Universidad de Deusto

Analíticas de evaluación para el diseño de una plataforma web de evaluación adaptativa de competencias digitales para la ciudadanía

Tesis doctoral presentada por Juan Bartolomé Boloix
dentro del Programa de Doctorado de
Ingeniería para la Sociedad de la Información y Desarrollo Sostenible

Dirigida por Dr. Pablo Garaizar Sagarmínaga
y Dr. Xabier Larrucea Uriarte

Universidad de Deusto

Analíticas de evaluación para el diseño de una plataforma web de evaluación adaptativa de competencias digitales para la ciudadanía

Tesis doctoral presentada por Juan Bartolomé Boloix
dentro del Programa de Doctorado de
Ingeniería para la Sociedad de la Información y Desarrollo Sostenible

Dirigida por Dr. Pablo Garaizar Sagarmínaga
y Dr. Xabier Larrucea Uriarte

El doctorando

Los directores

Bilbao, marzo de 2023

Analíticas de evaluación para el diseño de una plataforma web de evaluación adaptativa de competencias digitales para la ciudadanía

Autor: Juan Bartolomé Boloix
Director: Pablo Garaizar Sagarmínaga
Director: Xabier Larrucea Uriarte

Impreso en Bilbao
Primera edición, marzo de 2023

A mi familia, a quienes os debo todo.
Os quiero mucho.
Gracias por todo.

Resumen

En las últimas décadas, las tecnologías de la información y la comunicación han ido transformando nuestra vida cotidiana radicalmente. Si en la década de los 80 fue el ordenador personal, en la de los 90 Internet, en los 2000 el teléfono móvil inteligente y en los 2010 el acceso móvil a Internet, la actual década propone desafíos importantes para la sociedad a partir de las nuevas inteligencias artificiales de propósito general. Estos cambios sociales con profundas relaciones con la sostenibilidad se han visto acrecentados en un contexto de pandemia global que ha puesto de manifiesto la necesidad de replantearse los métodos y enfoques de trabajo existentes. Este nuevo contexto ha puesto de manifiesto las vulnerabilidades de nuestras industrias y de las frágiles cadenas de valor estratégicas, y ha acentuado la necesidad de buscar soluciones para hacer frente a estas vulnerabilidades. En este contexto, existe un reconocimiento generalizado de que la alfabetización digital es una competencia crucial desafiada por las demandas tecnológicas, informativas, cognitivas y socioemocionales de la era digital (List et al., 2020; Van Laar et al., 2017).

La agenda 2030 para el Desarrollo Sostenible (ODS) de las Naciones Unidas[1], recoge los principales retos y sus respectivos objetivos divididos en una serie de metas, todas ellas conectadas a la tecnología digital. Garantizar el acceso a la tecnología no es suficiente para afrontar los retos identificados, sino que es esencial dotar a las personas de las capacidades adecuadas para utilizar la tecnología de forma significativa (O'Sullivan et al., 2021).

Según la revisión bibliográfica y la consulta a expertos y responsables políticos a nivel internacional llevada a cabo por Ala-Mutka (2011), la adquisición de la competencia digital (CD) se considera tan relevante como las demás competencias clave para lograr una sociedad sostenible. En consecuencia, muchos expertos y responsables políticos han tratado de definir qué CD debe tener cada ciudadano. Más aún, adecuar la CD de los ciudadanos a los requisitos de la demanda de empleo se ha identificado como un factor clave para la futura mano de obra (Abidoye et al., 2022). En consecuencia, es crucial reducir la brecha digital de la ciudadanía, la cual está estrechamente relacionada con sus condiciones económicas, sociales (Portillo et al., 2020; Sá et al., 2021).

A pesar de todas las iniciativas puestas en marcha por la Comisión Europea como, por ejemplo, la *Digital Skills and Jobs Coalition*[2] y los avances en este sentido,

[1] https://www.un.org/sustainabledevelopment/es/development-agenda/
[2] https://digital-strategy.ec.europa.eu/en/policies/digital-skills-coalition

Europa sigue luchando contra la escasez de profesionales cualificados. Las instituciones educativas y de formación son incapaces de responder a esta demanda. Recientemente, 2023 ha sido declarado "año europeo de las capacidades" por la Comisión Europea, y se han puesto en marcha varias iniciativas en la educación, la capacitación y las CDs para cada ciudadano con el fin de continuar con el esfuerzo de alcanzar los objetivos establecidos por la Unión Europea en la *Brújula Digital*[3]: El 80% de los europeos tendrá CDs básicas y habrá 20 millones de especialistas en TIC para 2030.

El reconocimiento de la CD como componente transversal que sirve de apoyo a otras competencias clave es uno de los objetivos marcados junto a la adopción de un marco de referencia único de CD, la creación de directrices y perfiles competenciales, el fomento del reconocimiento de las competencias adquiridas en entornos no formales e informales y el desarrollo de herramientas de evaluación, como destacó CEDEFOP (2016) o reflejó la Comisión Europea en su programa Erasmus+[4]. En este contexto, la acreditación de la CD se ha convertido en un tema de creciente interés en los últimos años. Varios autores han examinado los principales avances y limitaciones, destacando que la mayoría de los sistemas de evaluación se basan en autoevaluaciones, no cubren los tres componentes de la CD (conocimientos, habilidades y actitudes), apenas muestran evidencias de su calidad, y evalúan principalmente habilidades cognitivas de orden bajo (p. ej., recordar y comprender) (Kluzer y Priego, 2018; Law et al., 2018; Saltos-Rivas et al., 2021; Siddiq et al., 2016; Zhao et al., 2021).

La Evaluación Mediada por la Tecnología (EMT) también llamado en inglés como Technology Assessment (TEA) ofrece inmensas oportunidades para mejorar la experiencia de quienes se examinan, y desarrollar modos de evaluación más pertinentes y ajustados a los requerimientos (Cho et al., 2019; Debuse y Lawley, 2016; Drasgow, 2016; McArthur, 2022); Scherer et al., 2017; Shute y Rahimi, 2017; Stödberg, 2012; Zenisky y Luecht, 2016). En concreto, la TEA posibilita el uso de entornos simulados (Binkley et al., 2012), y oportunidades para aplicar el conocimiento en un entorno seguro (Scherer et al., 2014). La TEA tiene un enorme potencial para proporcionar formatos de ítems innovadores, así como la posibilidad de obtener información sobre el comportamiento y el rendimiento de quienes se examinan durante las pruebas, recogiendo diferentes tipos de datos como, p. ej., datos de resultados, tiempo de respuesta y flujos de clics (Greiff et al., 2015; Osborne, Dunne, y Farrand, 2013; Timmis et al., 2016). Además, es posible utilizar

[3] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_es
[4] https://erasmus-plus.ec.europa.eu/es

estos datos con fines de validación para soportar las inferencias hechas a partir de las puntuaciones obtenidas (Oranje et al., 2017).

Para lograr una adecuada evaluación de la CD, es crucial diseñar los ítems de evaluación apropiados para que desencadenen los conocimientos y habilidades esperados, especialmente cuando se evalúan constructos cognitivos complejos en los que se utilizan distintos formatos dinámicos como las simulaciones interactivas o los juegos serios (O'Leary et al., 2018; Van Voorhis y Paris, 2019).

El lanzamiento del Marco Europeo de Competencias Digitales para la Ciudadanía (DigComp) por Ferrari y Punie (2013) del Instituto de Prospectiva Tecnológica (JRC-IPTS) de la Comisión Europea y sus posteriores versiones (Stephanie et al., 2017; Vuorikari et al., 2016; Vuorikari et al., 2022), facilitaron el desarrollo de implementaciones a medida, proporcionando un marco de referencia para trabajar en CD. Sin embargo, la mayoría de las implementaciones realizadas hasta ahora relacionadas con la evaluación de CD son autoevaluaciones compuestas por ítems de opción múltiple y escalas *Likert*, que sólo miden habilidades cognitivas de bajo orden (Kluzer y Priego, 2018; Mattar et al., 2022).

Una de estas implementaciones basadas en DigComp, es la llevada a cabo en Euskadi por el Gobierno Vasco para la evaluación y certificación de la CD de la ciudadanía, BAIT[5], como parte de sus *Estrategia para la Transformación Digital de Euskadi*[6]. La investigación llevada a cabo en esta tesis doctoral se encuentra estrechamente relacionada con BAIT, ya que su principal autor es parte de su equipo, y los avances logrados en la tesis están siendo incorporados en BAIT de manera continuada. Cabe destacar que tanto BAIT como Pathways for Employ (P4E), herramienta desarrollada en este estudio y que describimos más adelante, han sido seleccionadas por Kluzer y Priego (2018) como casos de éxito de implementaciones a medida llevadas a cabo basadas en DigComp.

Por último, los test tradicionales son menos eficaces a la hora de evaluar las capacidades quienes se examinan, especialmente de quienes cuentan con bajas y altas capacidades (Ling et al., 2017), en parte debido a que muchos de los ítems administrados no sirven para distinguir entre ambos correctamente (Aybek y Demirtasli, 2017). Una posible solución consiste en aplicar pruebas adaptativas haciendo uso de la tecnología (TAI) (Troussas et al., 2020). A día de hoy no tenemos constancia de que se haya implementado una evaluación adaptativa de CD. Por este motivo, hemos decidido finalizar la investigación analizando este hecho y examinando los principales puntos a examinar para transformar una prueba lineal

[5] https://www.bait.eus
[6] https://bideoak2.euskadi.eus/2021/03/30/news_67948/ETDE2025_Estrategia_ES.pdf

en adaptativa, acompañando los resultados con el uso de simulaciones que faciliten la toma de decisiones.

Agradecimientos

Muchas gracias a todos los que habéis formado parte de este largo camino. A mis directores de tesis Pablo y Xabi, por su dedicación y paciencia conmigo, a mis compañeros de trabajo en Tecnalia y amigos, que sin su ayuda y ánimos no habría sido posible, a mi empresa Tecnalia Research & Innovation por su soporte y ayuda, y a todos los participantes en los estudios llevados a cabo, por su participación desinteresada y colaboración.

Tabla de contenido

Índice de figuras

Índice de tablas

No te preocupes, hay millones de olas ahí afuera.
Tómate tu tiempo y tu ola llegará.

Duke Kahanamoku

1. Introducción

En esta memoria de tesis doctoral se recoge la investigación que he llevado a cabo en los últimos años a partir del desarrollo de una herramienta de evaluación de competencias digitales, la incorporación de diferentes formatos de preguntas con el objetivo de evaluar habilidades de distinto orden cognitivo, el análisis de las propiedades psicométricas de las pruebas y el análisis de viabilidad para transformar las pruebas de evaluación lineales en pruebas adaptativas.

1.1. Motivación y contexto

En una sociedad como la actual inmersa en una rápida y constante evolución, en la que la información y su comunicación a través de las tecnologías digitales han transformado nuestra vida cotidiana. Existe un reconocimiento generalizado de que la alfabetización digital es una competencia crítica para la ciudadanía en general, desafiada por las demandas tecnológicas, informativas, cognitivas y socioemocionales de la era digital (List et al., 2020).

La agenda 2030 para el Desarrollo Sostenible (ODS) de las Naciones Unidas, recoge los principales retos y sus respectivos objetivos divididos en una serie de metas, todas ellas están conectadas al potencial y los efectos de la tecnología digital. Garantizar el acceso a la tecnología no es suficiente y para afrontar los retos identificados en la ODS, sino que es esencial dotar a las personas de las capacidades adecuadas para utilizar la tecnología de forma significativa (O'Sullivan et al., 2021).

Según la revisión bibliográfica y la consulta a expertos y responsables políticos a nivel europeo e internacional llevada a cabo por Ala-Mutka (2011), la adquisición

de la competencia digital (CD) se considera tan relevante como las demás competencias clave. En consecuencia, y viendo que la adquisición de CD puede proporcionar importantes beneficios en la sociedad actual, muchos expertos y responsables políticos han tratado de definir qué CD debe tener cada ciudadano. Más aún, adecuar la CD de los ciudadanos a los requisitos de la demanda de empleo se ha identificado como un factor clave del desarrollo sostenible para la futura mano de obra (Abidoye et al., 2022). En consecuencia, es crucial reducir la brecha digital de la ciudadanía, la cual está estrechamente relacionada con las condiciones económicas, sociales y culturales de los ciudadanos e impide el desarrollo sostenible (Portillo et al., 2020; Sá et al., 2021).

A pesar de todas las iniciativas puestas en marcha por la Comisión Europea como, por ejemplo, la *Digital Skills and Jobs Coalition*[1] y los avances en este sentido, Europa sigue careciendo de la mano de obra digitalmente cualificada necesaria. Recientemente, 2023 acaba de ser declarado "*año europeo de las capacidades*" por la Comisión Europea, y se pondrán en marcha varias iniciativas con un enfoque y una inversión mucho más fuertes en la educación, la capacitación y las CD para cada ciudadano, para continuar con el esfuerzo de alcanzar los objetivos establecidos por la Unión Europea en la *Brújula Digital*[2] para la década digital de Europa: El 80% de los europeos tendrá CDs básicas y habrá 20 millones de especialistas en TIC para 2030.

Además, el reconocimiento de la CD como componente transversal que sirve de apoyo a otras competencias clave es uno de los puntos clave en relación con la adopción de un marco de referencia único de CD, la creación de directrices y perfiles competenciales, el fomento del reconocimiento de las competencias adquiridas en entornos no formales e informales y el desarrollo de herramientas de evaluación, como destacó CEDEFOP (2016) o reflejó la Comisión Europea (2019) en su programa Erasmus+[3].

En este contexto, la acreditación de la CD se ha convertido en un tema de creciente interés en los últimos años. Varios autores han examinado los principales avances y limitaciones, destacando que la mayoría de los sistemas de evaluación consisten en autoevaluaciones, no cubren los tres componentes de la CD y principalmente evalúan habilidades cognitivas de orden bajo (por ejemplo, recordar y comprender) (Kluzer y Priego, 2018; Law et al., 2018; Saltos-Rivas et al., 2021; Siddiq et al., 2016; Zhao et al., 2021).

La TEA ofrece inmensas oportunidades para mejorar la experiencia de las personas que se examinan y desarrollar modos de evaluación más pertinentes y ajustados a las necesidades actuales (Cho et al., 2019; Debuse y Lawley, 2016; Drasgow, 2016; Scherer et al., 2017; Shute y Rahimi, 2017; Stödberg, 2012; Zenisky y Luecht, 2016). En concreto, en la evaluación de la CD, posibilita el uso de entornos

[1] https://digital-strategy.ec.europa.eu/en/policies/digital-skills-coalition

[2] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_es

[3] https://erasmus-plus.ec.europa.eu/es

simulados (Binkley et al., 2012), y oportunidades para aplicar el conocimiento en un entorno seguro (Scherer et al., 2014). Más aún, la TEA tiene un enorme potencial para proporcionar formatos de ítems innovadores, así como la posibilidad de obtener información sobre el comportamiento y el rendimiento de quienes se examinan durante las pruebas, recogiendo diferentes tipos de datos como, por ejemplo, datos de resultados, tiempo de respuesta y flujos de clics (Bartolomé y Garaizar, 2022; Greiff et al., 2015; Osborne, Dunne, y Farrand, 2013; Timmis et al., 2016). Además, es posible utilizar estos datos con fines de validación para soportar las inferencias hechas a partir de las puntuaciones obtenidas (Oranje et al., 2017). En concreto, el diseño de los ítems de evaluación es fundamental para que desencadenen los conocimientos y habilidades esperados, especialmente cuando se evalúan constructos cognitivos complejos en los que se utilizan distintos formatos dinámicos como las simulaciones interactivas o los juegos serios (O'Leary et al., 2018; Van Voorhis y Paris, 2019).

El lanzamiento del Marco Europeo de Competencias Digitales para la Ciudadanía (DigComp) por Ferrari y Punie (2013) del Instituto de Prospectiva Tecnológica (JRC-IPTS) de la Comisión Europea y sus posteriores versiones (Stephanie et al., 2017; Vuorikari et al., 2016; Vuorikari et al., 2022), facilitaron el desarrollo de implementaciones a medida, proporcionando un marco de referencia para trabajar en CD. Sin embargo, la mayoría de las implementaciones realizadas hasta ahora relacionadas con la evaluación de CD son autoevaluaciones compuestas por ítems de opción múltiple y escalas Likert, que solo miden habilidades cognitivas de bajo orden (Kluzer y Priego, 2018).

Una de estas implementaciones basadas en DigComp, es la llevada a cabo en Euskadi por el Gobierno Vasco para la evaluación y certificación de la CD de la ciudadanía, BAIT[4], como parte de sus *Estrategia para la Transformación Digital de Euskadi*[5]. La investigación llevada a cabo en esta tesis doctoral está estrechamente relacionada con BAIT, ya que su principal autor es parte de su equipo, y los avances logrados en la tesis están siendo incorporados en BAIT de manera continuada.

Por último, los test tradicionales son menos eficaces a la hora de evaluar las capacidades quienes se examinan, especialmente de quienes cuentan con bajas y altas capacidades (Ling et al., 2017), en parte debido a que muchos de los ítems administrados no sirven para distinguir entre ambos correctamente (Aybek y Demirtasli, 2017). Una posible solución consiste en aplicar pruebas adaptativas haciendo uso de la tecnología (TAI) (Troussas et al., 2020). Cabe destacar que hoy en día no tenemos constancia de que se haya implementado una evaluación adaptativa de CD. Por este motivo, hemos decidido finalizar la investigación analizando este hecho y examinando los principales puntos a examinar para transformar una prueba lineal en adaptativa, acompañando los resultados con el uso de simulaciones que faciliten la toma de decisiones.

[4] https://www.bait.eus

[5] https://bideoak2.euskadi.eus/2021/03/30/news_67948/ETDE2025_Estrategia_ES.pdf

1.2. Hipótesis y preguntas de investigación

El primer objetivo de nuestra investigación es diseñar una herramienta de evaluación de CD, suficientemente flexible como para permitir la incorporación de distintos formatos de ítems que permitan evaluar habilidades de distinto orden cognitivo. Como parte de la investigación, en el proyecto europeo PathwaysforEmploy (P4E) investigamos la definición y evaluación de perfiles digitales llevando a cabo una implementación a medida basada en DigComp, que nos sirvió como punto de partida de la investigación.

Para ello, decidimos realizar nuestra propia implementación basada en el marco de referencia de DigComp, diseñando una plataforma web de evaluación de CD orientada a la ciudadanía. Además, incorporamos dos casos de estudio: una prueba para el área competencial (AC) de *Información y Alfabetización Digital* (IAD) y otra prueba para la CD de *Netiqueta*. Ambos casos representan dos enfoques diferentes adoptados por algunas de las iniciativas relevantes identificadas como casos de éxito por Kluzer y Priego (2018). Para la selección del AC, elegimos una de las tres principales AC de DigComp. Para la selección de la CD, elegimos una CD que no suele evaluarse en profundidad y que, al incluir formatos de ítems dinámicos, los resultados serían más notables.

Para abordar este objetivo planteamos las siguientes preguntas de investigación, que codificamos de la siguiente manera:

- PI_K1. ¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?
- PI_K2. ¿Qué propiedades psicométricas tienen las pruebas? ¿Qué evidencias se pueden presentar que soporten las inferencias realizadas de las puntuaciones obtenidas?

El segundo objetivo de nuestra investigación es utilizar la información registrada durante la realización de las pruebas por parte de los participantes con el objetivo de validar las inferencias realizadas entre las afirmaciones y el comportamiento observado. De las 5 formas posibles identificadas por Oranje et al., (2017), nos centramos en utilizar los datos registrados con dos fines: 1) generar y probar inferencias sobre el constructo de interés; 2) mejorar los diseños de los ítems analizando los patrones de comportamiento mostrados por los participantes en las pruebas.

Para ello, dotamos tanto a la plataforma de evaluación como a las simulaciones interactivas que integramos en la prueba, de capacidad para registrar una variedad de eventos por parte de los participantes durante la realización de las pruebas: resultados de las respuestas a los ítems, registros de respuestas, clics realizados, tiempo empleado por ítem y prueba, número de intentos erróneos en las simulaciones, así como el camino escogido para resolver la simulación en caso de tener varios caminos válidos. Además, realizamos un estudio exploratorio con el

objetivo de examinar los procesos de respuesta (RPs) de una selección de formatos de ítems mediante técnicas de seguimiento ocular.

Para abordar este objetivo planteamos las siguientes preguntas de investigación, que codificamos de la siguiente manera:

- PI_K3. ¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos generar y probar inferencias sobre el constructo de interés?
- PI_K4. ¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos utilizar esta información para mejorar los diseños de los ítems?

Tras examinar el creciente número y variedad de herramientas existentes para evaluar la CD y sus distintas implementaciones, el tercer y último objetivo de nuestra investigación consiste en realizar un análisis detallado de los distintos aspectos a considerar para llevar a cabo una evaluación adaptativa.

Para abordar este objetivo planteamos la siguiente pregunta de investigación, que codificamos de la siguiente manera:

- PI-K5. ¿Qué consideraciones deben tenerse en cuenta para pasar de un diseño lineal a un diseño adaptativo en este tipo de sistemas?

La hipótesis que intentaremos validar es la siguiente:

Es posible evaluar adecuadamente constructos cognitivos complejos dentro de la Competencia Digital mediante herramientas que empleen diferentes formatos dinámicos, el registro detallado de la interacción de los participantes y test adaptativos informatizados.

Tanto los objetivos como las preguntas de investigación e hipótesis han guiado las actividades que hemos llevado a cabo en esta investigación para su validación, tal y como describimos en las siguientes secciones.

1.3. Metodología de investigación

En la figura 1.1 hemos representado de manera conceptual la metodología que hemos seguido en la investigación. Como puede apreciarse, una serie de fases se han realizado de forma iterativa, con el objetivo de ir refinando la adecuación de la herramienta a los objetivos deseados.

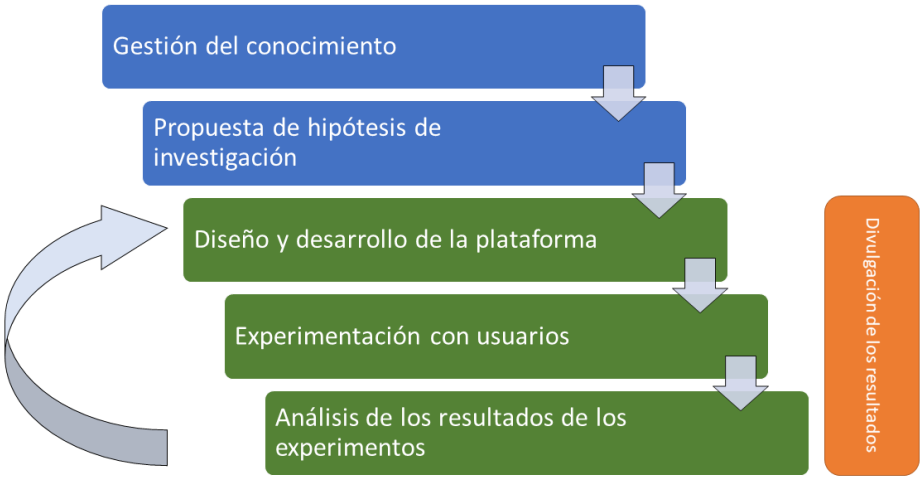


Figura 1.1. Representación conceptual de la metodología de investigación seguida en la tesis

A continuación, mostramos el propósito de las distintas fases que hemos seguido en la investigación:

1. Gestión del conocimiento

- Análisis del estado del arte existente a nivel científico en los campos de nuestra investigación.
- Análisis de las herramientas de evaluación de CD disponibles, así como los distintos enfoques seguidos y las tecnologías utilizadas en sus implementaciones, que sirvan de guía para el desarrollo de nuestra propia herramienta[6].
- Elección del enfoque a aplicar, así como las tecnologías a utilizar.

2. Propuesta de hipótesis de investigación

- Formulación de los objetivos, preguntas de investigación e hipótesis que dirigirán el proceso de nuestra investigación.
- Validación de que la hipótesis refleja el objetivo de nuestro estudio y de que previamente no haya sido resuelto en la literatura científica.

3. Diseño y desarrollo de la plataforma

- Diseño y desarrollo de la plataforma tecnológica de evaluación de CD que será utilizada para experimentar con usuarios.

[6] http://www.evaluatucompetenciadigital.com

- Diseño e implementación de los ítems dinámicos como las simulaciones interactivas, así como su integración en la plataforma de evaluación de CD.
- Diseño e implementación de los diferentes mecanismos de captura de las interacciones de los participantes durante la realización de las pruebas de evaluación.

4. Experimentación con usuarios

- Contacto con Kzgunea[7], la red pública de centros de capacitación en TIC del Gobierno Vasco, definición de agenda y organización de las sesiones con los dinamizadores de los centros en calidad de expertos.
- Contacto con el Laboratorio de Factores Humanos y Experiencia de Usuario (HF&UX) de Tecnalia[8] para llevar a cabo el estudio basado en técnicas de seguimiento ocular, definición de agenda, captación de voluntarios y organización de las sesiones con los participantes.
- Diseño de la prueba experimental con participantes pertenecientes al colectivo destinatario de nuestro estudio.
- Contacto con AllDigital[9], KZgunea y demás entidades en contacto con el colectivo destinatario para la captación de colaboradores en el estudio publicado dentro del marco de la *AllDigital week*[10].

5. Análisis de los resultados de los experimentos

- Tratamiento y filtrado de los datos registrados.
- Observación de los datos registrados en los distintos estudios con fines estadísticos.
- Observación de los datos registrados en los distintos estudios llevados a cabo con fines de validación.

6. Divulgación de los resultados

- Divulgación de los resultados de la investigación, tanto parciales como finales, en congresos internacionales y publicaciones científicas, así como con la posterior publicación de la tesis doctoral.

[7] https://www.kzgunea.eus/es/inicio

[8] https://www.tecnalia.com/infraestructuras/laboratorio-de-factores-humanos-y-experiencia-de-usuario-hfux

[9] https://all-digital.org/

[10] https://www.alldigitalweek.eu/

1.4. Estructura del documento

En este capítulo describimos la motivación y contexto que dan origen a esta investigación, resumiendo las hipótesis del estudio y describiendo la metodología seguida. En el capítulo 2 describimos los estudios previos encontrados en las áreas científicas relacionadas con esta tesis: la CD, la evaluación y acreditación de la CD, la evaluación mediante el uso de la tecnología y las analíticas de evaluación, y los test adaptativos informatizados (TAIs). En el capítulo 3 describimos el proceso iterativo llevado a cabo para el diseño y desarrollo de la herramienta de evaluación www.evaluatucompetenciadigital.com (ETCD), describiendo los distintos estudios que hemos llevado a cabo con expertos en la materia y con ciudadanos anónimos a través de Internet. En el capítulo 4 analizamos los puntos claves necesarios para transformar las pruebas lineales en adaptativas, y examinamos si es justificable la elección de un TAI para ambas pruebas. Por último, en el capítulo 5 hacemos un resumen de las principales aportaciones de nuestra investigación y las futuras líneas de investigación que identificamos.

Si luchamos, podemos perder; si no lo hacemos, estamos perdidos.

Ramón Calderón

Por esto amo la tecnología; si la usas bien, puede darte poder y privacidad.

Cory Doctorow

2. Estado del arte

En este capítulo revisamos los conceptos y evidencias encontradas en la literatura científica en el contexto en el que se desarrolla nuestra investigación. Centraremos el propio término de CD y su variedad de definiciones, para abordar después la problemática de su evaluación, especialmente en un contexto de políticas gubernamentales donde la CD es fundamental para capacitar a los ciudadanos para que vivan en una sociedad en la que sean consumidores y creadores de tecnología digital de forma crítica, creativa, autónoma y ética, lo que también es esencial para el desarrollo sostenible (Santos y Serpa, 2017). Posteriormente, revisaremos las herramientas que se han utilizado para evaluar la CD, prestando especial atención a la definición de perfiles digitales (PD) y su evaluación. Asimismo, examinaremos las posibilidades que ofrecen las pruebas mediadas por la tecnología (TEA) y el potencial que ofrece el uso de la información de los participantes durante la realización de las pruebas. En concreto, prestaremos especial atención al uso de tecnología de seguimiento ocular (ET). Por último, evaluaremos la aplicabilidad de pruebas adaptativas (TAI) en el contexto de evaluación de la CD.

2.1. La competencia digital

Hoy en día, las tecnologías digitales están presentes en casi todos los aspectos de nuestra vida diaria (Cascio y Montealegre, 2016). Este cambio digital está generando un impacto enorme en las competencias necesarias para la mayoría de los trabajos debido a la automatización de las tareas habituales, la creación de nuevos y diferentes tipos de trabajos, o la necesidad de profesionales más capacitados en las TIC en todos los sectores.

Las tecnologías digitales son las principales impulsoras de la innovación, el crecimiento y la creación de empleo en la economía global. Sin embargo,

incorporar tales tecnologías no es suficiente para asegurar el éxito (Schallenmueller, 2016). Previamente, el foco estuvo puesto en facilitar la disponibilidad y el acceso a la tecnología, pero durante los últimos años los esfuerzos se han centrado en evitar la exclusión de la ciudadanía de los servicios básicos en la sociedad y asegurar al acceso universal a la información (Van Deursen y Helsper, 2015). La alfabetización digital (también conocida como fluidez digital) refleja la competencia de un empleado para lograr los resultados deseados utilizando la tecnología y es clave en los entornos laborables de hoy en día (Colbert et al., 2016). Es un factor clave para evitar la exclusión social y laboral, donde el nivel educativo, la educación no formal, y el uso de las diferentes habilidades (lectura, numéricas y relacionadas con las TIC) están decisivamente relacionadas (Iñiguez-Berrozpe y Boeren, 2019).

La CD de la mano de obra juega un papel relevante y determinante en la adopción de la tecnología (Marsh, 2021), y es considerada como una habilidad que evoluciona (Mohammadyari y Singh, 2015), y que requiere ser constantemente actualizada para que tanto empleados como organizaciones puedan aprovechar el potencial del entorno de trabajo digitalizado. De hecho, se previó que para 2022, el 54% de la mano de obra existente necesitaría mejorar sus CD o volver a capacitarse (Schwab, 2018; Voogt et al., 2013). Más aun, el uso de teléfonos inteligentes que se ha extendido rápidamente debe abordarse urgentemente como parte del desarrollo de la alfabetización digital (Forkosh Baruch y Erstad, 2018).

El sector digital y, en especial, las habilidades digitales, ocupan un lugar destacado en la agenda europea desde 2014 con el *Digital Single Market (DSM)* como una de sus prioridades (EC, 2015). La *New Skills Agenda for Europe* es una de las herramientas más importantes para poner a disposición de los ciudadanos de la UE la formación, las capacidades y el apoyo adecuados (European Commission, 2016). La *Digital Skills and Jobs Coalition* es una de estas iniciativas con el objetivo de mejorar las habilidades digitales de la población en general (European Commission, 2017b). De hecho, el 44% de la población europea no tenía CD básicas, a pesar de que 9 de cada 10 puestos de trabajo requerirán pronto CD (The Digital Skills Gap in Europe, 2017), y que para el 2020 iban a faltar 500.000 expertos en el sector de las TIC (EC, 2016b).

A pesar de los avances, Europa sigue careciendo de la mano de obra digitalmente cualificada necesaria. De hecho, 2023 acaba de ser declarado *Año Europeo de las Capacidades* por la Comisión Europea, año en el que se va a poner en marcha varias iniciativas en la educación, la capacitación y las CD para cada ciudadano, de acuerdo con los objetivos establecidos por la Unión Europea en *la Brújula Digital para la década digital de Europa*[11], es decir, el 80% de los europeos con CD básicas y 20 millones de especialistas en TIC para 2030.

Según el análisis de la bibliografía llevado a cabo por Oberländer et al. (2020), existe una falta de investigación científica sobre la CD de adultos y un abandono en el contexto laboral. Respecto a la definición de perfiles digitales (PD), es decir,

[11] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_es

perfiles en los cuales las CD son requeridas en mayor o menor medida, la base de datos *Occupational Information Network* (O*NET)[12] proporciona descripciones detalladas de las ocupaciones, mostrando para cada perfil la importancia de las habilidades requeridas para desempeñar ese trabajo (Handel, 2016; Peterson et al., 2001). Esta herramienta también indica el nivel de competencia requerido para cada habilidad, pero no utiliza un marco de referencia para las CD identificadas. A nivel europeo, la *European multilingual classification of Skills, Competences, Qualifications and Occupations* (ESCO)[13] proporciona el vocabulario común necesario para facilitar el intercambio de información relacionada con las competencias y las cualificaciones, y ha aplicado directamente DigComp en la definición de los conocimientos y aptitudes esenciales para la lista de ocupaciones proporcionada, pero no indica el nivel de competencias requerido.

Desde una perspectiva práctica, la definición de un sistema para medir el nivel de CD de los trabajadores en función de sus PD podría ser una herramienta útil para mejorar la CD de diferentes profesiones. Sin embargo, no se ha abordado directamente la definición de qué CD son esenciales para PD específicos, p. ej. teletrabajadores o emprendedores.

2.1.1. Conceptos y definiciones

Hasta hace poco, no existía un entendimiento común de lo que son las CD y cuáles de ellas deberían ser necesarias para cada ciudadano (Ala-Mutka, 2011), y en los últimos años se han dado varias definiciones sobre lo que es la CD. Esta variedad puede deberse a que la CD es una definición que depende del contexto (Ferrari, 2012; Law et al., 2018). Existen varios conceptos estrechamente relacionados con la CD que se abordan desde diferentes perspectivas (p. ej. el Parlamento Europeo, el Consejo Europeo, la Comisión Europea, etc.) (Ala-Mutka, 2011). El término "*alfabetización digital*" se define como "*las habilidades necesarias para lograr la competencia digital [...] se sustenta en el uso técnico básico de los ordenadores e Internet*" (European Parliament and the Council, 2008b). La alfabetización digital es un término amplio que abarca no solo las habilidades, sino también los conocimientos y las actitudes hacia la tecnología. Del mismo modo, el término "*e-Skills*" abarca los diferentes niveles de competencias en TIC: competencias del usuario de las TIC, competencias del profesional de las TIC y competencias de negocio electrónico. Es el concepto utilizado por la Dirección General de Empresa e Industria y el sector de las TIC. Otro término cercano es el de "*alfabetización mediática*", que se define como la capacidad de acceder a los medios de comunicación, comprender y evaluar críticamente diferentes aspectos de los contenidos mediáticos y crear comunicaciones en diversos contextos. Este concepto es utilizado por la Comisión Europea para subrayar las habilidades relacionadas con las herramientas, la actitud crítica y la comprensión del uso seguro. En este contexto, la "*competencia digital*" implica "*el uso seguro y crítico de las tecnologías*

[12] https://www.onetonline.org/

[13] https://esco.ec.europa.eu/es

de la sociedad de la información (TSI) para el trabajo, el ocio y la comunicación. Se sustenta en las habilidades básicas de las TIC: el uso de ordenadores para recuperar, evaluar, almacenar, producir, presentar e intercambiar información, y para comunicarse y participar en redes de colaboración a través de Internet" (European Parliament & the Council, 2006).

En un contexto de políticas gubernamentales, Ferrari (2012) definió la CD como "*el conjunto de conocimientos, habilidades, actitudes, estrategias y concienciación que se requieren cuando se utilizan las TIC y los medios digitales para realizar tareas, resolver problemas, comunicarse, gestionar información, colaborar, crear y compartir contenidos, y construir conocimiento de una manera eficaz, eficiente y adecuada de forma crítica, creativa, autónoma, flexible, ética y sensible para el trabajo, el entretenimiento, la participación, el aprendizaje, la socialización, el consumo y el empoderamiento*". En consecuencia, la CD es fundamental para capacitar a los ciudadanos para que vivan en una sociedad en la que sean consumidores y creadores de tecnología digital de forma crítica, creativa, autónoma y ética, lo que también es esencial para el desarrollo sostenible (Santos y Serpa, 2017).

Por último, también existen otros conceptos relacionados con la CD como "*alfabetización informática*", "*alfabetización en Internet*", "*alfabetización informacional*", "*alfabetización en TIC*" y "*fluidez digital*". Teniendo en cuenta el complejo escenario de definiciones y conceptos en torno a la CD, Ala-Mutka (2011) definió la CD para DigComp como una alfabetización emergente a partir de otras alfabetizaciones, que incluyen los conceptos enumerados anteriormente.

2.1.2. DigComp, el Marco Europeo de Competencias Digitales para la Ciudadanía

DigComp (Ferrari y Punie, 2013) y sus actualizaciones (Stephanie et al., 2017; Vuorikari et al., 2016; Vuorikari et al., 2022) fueron lanzados por el Instituto de Prospectiva Tecnológica (JRC-IPTS) de la Comisión Europea con el objetivo de contribuir a una mejor comprensión y desarrollo de la CD en Europa (Janssen et al., 2013) y cumplir los siguientes objetivos: (1) identificar los componentes clave de la CD en términos de conocimientos, habilidades y actitudes necesarias para ser digitalmente competente; (2) desarrollar los descriptores de la CD con el fin de alimentar un marco conceptual y / o directrices que pueden ser validados a nivel europeo, teniendo en cuenta los marcos pertinentes actualmente disponibles (Ferrari, 2012); y (3) proponer una hoja de ruta para el posible uso y revisión del marco de la CD y los descriptores de las CD para todos los niveles de aprendizaje. Posteriormente, la versión 2.0 del marco definió 8 niveles de competencia para cada CD en lugar de 3 (Vuorikari et al., 2016), proporcionando una descripción más detallada de las características de cada nivel (en términos de conocimientos, habilidades y actitudes) y de los requisitos para pasar al siguiente nivel. De hecho, esta contribución es de especial interés para la definición de itinerarios de acreditación de PD que requieren competencias y niveles específicos. En definitiva,

DigComp se propuso crear un consenso europeo sobre los componentes de la CD mediante el desarrollo de un marco conceptual que pudiera servir de referencia para las iniciativas, los planes de estudio y las certificaciones actuales.

DigComp define el alcance y los componentes de la CD para los ciudadanos proporcionando una comprensión global, completa y compartida de lo que es la CD, estructurada en 5 dimensiones: (1) AC que implican las diferentes CD, (2) descriptores para cada CD, (3) niveles de competencia a nivel de CD, (4) conocimientos, habilidades y actitudes esperados en cada CD y (5) diferentes propósitos de aplicabilidad. En la figura 2.1 se muestran las 21 CD y su distribución en las 5 AC.

Figura 2.1. Modelo de referencia conceptual de DigComp. Adaptado de "Marco de Competencias Digitales parala Ciudadanía" (p. 9), traducción del marco DigComp 2.2 por Asociación Somos Digital, 2022.

En DigComp 2.0, Vuorikari et al. (2016) actualizaron la terminología, los conceptos y los descriptores de las CD. En DigComp 2.1, Stephanie et al. (2017) aplicaron cambios significativos aumentando a ocho los niveles de CD, haciendo uso de la taxonomía de Bloom en la definición de los descriptores de las CD. En DigComp Vuorikari Rina et al. (2022) incorporaron una actualización de los ejemplos de conocimientos, habilidades y actitudes para cada CD.

Las entidades interesadas en aplicar el marco de referencia deben llevar a cabo sus propias implementaciones en función de sus necesidades, identificando qué dispositivos digitales o aplicaciones de software específicas son relevantes para ellas. DigComp proporciona un vocabulario común y es lo suficientemente flexible como para ser adaptado a diferentes PD y contextos (Bartolomé et al., 2022).

2.1.3. Consideraciones para esta disertación en términos de CD

Considerando el contexto de BAIT, el cual está orientado a la mejora y acreditación de la CD de la ciudadanía, decidimos utilizar DigComp como marco de referencia debido a sus notables puntos fuertes: (1) fue diseñado tras un profundo análisis de los marcos de CD disponibles; (2) siguió un meticuloso proceso de consulta y desarrollo por parte de expertos en el área de la CD; y (3) como resultado, proporciona una visión global basada en CD y AC.

Por razones similares, la UNESCO también seleccionó DigComp para el desarrollo del Marco Global de Alfabetización Digital (cuyas siglas en ingles son DLGF) (Laanpere, 2019; Law et al., 2018). Más aún, el Banco Mundial también identificó DigComp como uno de los marcos más completos y utilizados para la CD general (Bashir y Miyamoto, 2020).

Por otra parte, DigComp promueve la independencia de las tecnologías y los dispositivos empleados. No obstante, de acuerdo con Fraillon (2018), las herramientas de software comunes tienden a proporcionar funciones similares a pesar de que el diseño de la interfaz puede variar. Además, los resultados de estudios recientes también cuestionan que la CD sea independiente del contexto de la tarea y de la tecnología utilizada ya que, en algunos campos específicos, las tecnologías digitales concretas o el manejo de tecnologías digitales específicas pueden ser una CD en si misma (Law et al., 2018).

Por último, cabe mencionar cómo los rápidos avances tecnológicos están causando cambios notables en el ámbito digital. En concreto, tecnologías emergentes como, por ejemplo, la Inteligencia Artificial o la Realidad Virtual / Aumentada, y otros fenómenos como, por ejemplo, la veracidad de la información y la desinformación o nuevos contextos como el teletrabajo, están causando que se eleve el grado de exigencia de CD por parte de la ciudadanía. Además, la necesidad de considerar los aspectos ecológicos y de sostenibilidad de la interacción con las tecnologías digitales se está acentuando. De ahí, que Vuorikari et al. (2022) publicaran una nueva actualización del marco DigComp considerando estos aspectos. Lo cual, reafirma la importancia de elegir un marco de referencia flexible y que constantemente se actualice de acuerdo con los últimos avances tecnológicos que se vayan produciendo.

2.2. Evaluación y acreditación de la CD

El reconocimiento de la CD como componente transversal que sirve de apoyo a otras competencias clave es uno de los puntos clave en relación con la adopción de un marco de referencia único de CD, la creación de directrices y perfiles competenciales, el fomento del reconocimiento de las competencias adquiridas en entornos no formales e informales y el desarrollo de herramientas de evaluación, como destacó CEDEFOP (2016) o se reflejó en el Programa Erasmus + (Comisión Europea, 2019).

En este contexto, la evaluación y acreditación de la CD se ha convertido en un tema de creciente interés en los últimos años y varios autores han examinado los principales avances y limitaciones (Kluzer y Priego, 2018; Law et al., 2018; Saltos-Rivas et al., 2021; Siddiq et al., 2016; Zhao et al., 2021). A pesar de emplear diferentes enfoques, algunas cuestiones siguen requiriendo un mayor estudio, destacando que la mayoría de los sistemas de evaluación son autoevaluaciones, no cubren los tres componentes de la CD (conocimientos, habilidades y actitudes) y evalúan principalmente habilidades cognitivas de orden bajo (según la taxonomía de Bloom, las habilidades cognitivas de bajo orden incluyen recordar, comprender y aplicar, mientras que las habilidades cognitivas de alto orden incluyen analizar, evaluar y crear).

Asimismo, es necesario dedicar recursos suficientes y reunir a los diferentes agentes involucrados para contribuir en la construcción de un ecosistema para la mejora de las CD de la ciudadanía, incluyendo aspectos clave como ofrecer la posibilidad de hacer un diagnóstico inicial reconociendo las habilidades existentes y las carencias, así como especificar la demanda de habilidades actuales y futuras (World Economic Forum, 2017).

2.2.1. Herramientas de evaluación y marcos de referencia

Además de las deficiencias previamente mencionadas, hasta hace poco, la mayoría de las evaluaciones de CD se basaban en marcos de evaluación de empresas comerciales (Law et al., 2018). En consecuencia, la selección de las CD enseñadas y evaluadas estaban influenciadas por el marco de referencia elegido, basado principalmente en aplicaciones comerciales como el paquete ofimático de Office y el sistema operativo de Microsoft. El lanzamiento de DigComp facilitó el desarrollo de implementaciones a medida, proporcionando un marco de referencia para trabajar en CD (Ferrari y Punie, 2013). Sin embargo, la mayoría de las implementaciones relacionadas con la evaluación de CD son autoevaluaciones compuestas por ítems de opción múltiple y escalas Likert, que solo miden habilidades cognitivas de bajo orden (Kluzer y Priego, 2018). Además, el componente de habilidad de la CD apenas es evaluado, probablemente porque el diseño de ítems con este fin es complicado y requiere mucho esfuerzo.

Las principales implementaciones identificadas por Kluzer y Priego (2018) y Law et al. (2018) con fines de acreditación, pueden clasificarse en 3 categorías diferentes: basadas en marcos de referencias comerciales, basadas en marcos de referencias propios, y personalizadas basadas en DigComp (como es el caso de esta tesis).

Como veremos más adelante, para este estudio nos basamos en DigComp como marco de referencia debido a sus notables puntos fuertes: (1) fue diseñado tras un profundo análisis de los marcos de CD disponibles; (2) siguió un meticuloso proceso de consulta y desarrollo por parte de expertos en el área de la CD; y (3) como resultado, proporciona una visión global basada en la CD y las AC. Por razones similares, la UNESCO también seleccionó DigComp como marco de referencia de CD para el desarrollo del DLGF (Laanpere, 2019; Law et al., 2018). Es

más, el Banco Mundial[14] también identificó en un informe reciente el marco DigComp como uno de los marcos más completos y utilizados para la CD general (Bashir y Miyamoto, 2020). A su vez, también fue el seleccionado por el servicio de evaluación y certificación de CD BAIT, estrechamente relacionado con este estudio.

Implementaciones basadas en marcos de referencias comerciales

La mayoría de los marcos de referencia de alfabetización digital adoptados se basan en cursos de formación y marcos de evaluación basados en marcos de referencia comerciales. De acuerdo con Law et al. (2018), 36 de 47 países adoptaron marcos comerciales: El ECDL[15], también conocido como ICDL, (n = 31), IC3[16] de Certiport (n = 13) y Microsoft Digital Literacy Standard Curriculum[17] (n = 11). Esto hace que las herramientas de formación y evaluación de CD estén fuertemente influenciadas por el marco de referencia elegido. En la tabla 2.1 puede verse una lista de los marcos de referencias identificados y analizadas para este estudio, con su dirección web (activa en enero de 2023).

En cuanto a los instrumentos de evaluación identificados por Law et al. (2018), a pesar de la inclusión de diferentes formatos de ítems como simulaciones interactivas, las evaluaciones dan prioridad a la tecnología en sí misma, en lugar del posible uso de diferentes aplicaciones para resolver determinadas tareas de manera efectiva y eficaz. ECDL es el marco comercial más relevante e implantado, así como la solución más completa. Proporciona una solución técnica basada en un software comercial que interactúa con el sistema operativo y las aplicaciones instaladas en el ordenador. Las personas examinadas tienen que llevar a cabo una serie de tareas reales usando aplicaciones reales integradas en el entorno de evaluación. Por otra parte, este entorno de evaluación tiene serias limitaciones ya que las pruebas han sido diseñadas en base a software con licencia privativa, no representando a muchas organizaciones que utilizan otro software, p. ej. Google Workspace[18] (Gilbert, 2019). Más aún, las pruebas principalmente ponen en práctica descriptores de los niveles básico e intermedio.

Las situaciones mostradas en sus evaluaciones están basadas en su mayoría en tareas que tienen que ser resueltas desde un ordenador. Según el Eurostat (European Statistical Office), este hecho no se corresponde con la realidad, donde en 2019 más del 90% de los jóvenes utilizaban dispositivos móviles para acceder a Internet y el 52% utilizaba un ordenador portátil. Además, cada vez son más los empleados equipados con dispositivos móviles, un 28% de todas las personas

[14] https://www.bancomundial.org/es/home

[15] https://icdleurope.org/

[16] https://certiport.pearsonvue.com/Certifications/IC3/Digital-Literacy-Certification/Overview.aspx

[17] https://www.microsoft.com/en-us/digital-literacy

[18] https://workspace.google.com/intl/es/

empleadas en empresas de la Unión Europea según el Eurostat de 2019, con una clara tendencia al alza. Esto refleja la importancia de que el personal pueda mantener el contacto con la empresa, así como con los proveedores y clientes, mientras se desplaza, permitiéndole consultar el correo electrónico, trabajar con documentos alojados en la nube o utilizar aplicaciones informáticas de la empresa. De acuerdo con el informe de GSMA (2022), en 2021 existían 474 millones de personas en Europa (el 86% de la población) abonadas a servicios móviles, y se espera que esta cifra aumente a 480 millones en 2025. Por lo tanto, los dispositivos móviles deberían incluirse en el marco de evaluación o podrían suponer una limitación relevante.

Marco de referencia	Web
ECDL (ICDL), The Digital Skills Standard	https://icdleurope.org/
IC3 Digital Literacy Certification	https://certiport.pearsonvue.com/Certifications/IC3/Digital-Literacy-Certification/Overview.aspx
+Microsoft Digital Literacy Standard Curriculum	https://www.microsoft.com/en-us/digital-literacy

Tabla 2.1. Marcos de referencia comerciales.

Implementaciones basadas en marcos de referencias propios

En el estudio llevado a cabo por Law et al. (2018) con el objetivo de poder desarrollar una metodología que pueda servir de base para el indicador temático 4.4.2 de los ODS: "*Porcentaje de jóvenes/adultos que han alcanzado al menos un nivel mínimo de competencia en CD*", los autores identificaron sólo 11 de los 47 países seleccionados que habían desarrollado sus propios marcos de evaluación nacionales como, p. ej. el *Marco de alfabetización digital de Columbia Británica*[19] o el *India Pradhan Mantri Gramin Digital Saksharta Abhiyan* (PMGDISHA)[20] . 7 de los cuales, además habían adoptado de manera complementaria marcos empresariales (ver tabla 2.2).

No obstante, como hay muchos otros factores contextuales y culturales que influyen en la aplicación de los marcos curriculares, no es posible concluir que estos marcos propios de evaluación nacionales sean comparables en cuanto a su ejecución o impacto. Además, no hay pruebas de que las competencias incluidas en los marcos nacionales difieran en función del desarrollo económico del país. Por lo tanto, no se podría obtener mucha información sobre las posibles diferencias en cuanto a la importancia relativa de las distintas CD (Law et al., 2018).

[19] https://www2.gov.bc.ca/assets/gov/education/kindergarten-to-grade-12/teach/teaching-tools/digital-literacy-framework.pdf

[20] https://www.pmgdisha.in/

País	Marco de referencia
Canadá	Marco de alfabetización digital de Columbia Británica
Canadá	UTILIZAR, COMPRENDER Y CREAR: Un marco de alfabetización digital para las escuelas canadienses
Chile	Competencias TIC SIMCE
Costa Rica	Estándares de rendimiento de los alumnos en el aprendizaje con tecnologías digitales
Hungría	Estrategia de educación digital
India	Misión Nacional de Alfabetización Digital
India	The Pradhan Mantri Gramin Digital Saksharta Abhiyan (PMGDISHA)
Indonesia	SiBerkreasi
Jordania	Fundación Reina Rania para la Educación y el Desarrollo (QRF) (antes Iniciativa Educativa Jordana)
Kenia	Digischool: el programa de alfabetización digital
Kenia	Plan Estratégico 2013-2018 de la Autoridad de las TIC
Kenia	Plan de estudios del programa Presidential Digitalent
Kenia	Plan Maestro Nacional de TIC de Kenia
Rep. De Corea	Baeumnara
Malasia	Programas de habilidades directivas, técnicas e informáticas
Filipinas	Normas Nacionales de Competencia en TIC (NICS) de Filipinas

Tabla 2.2. Marcos de referencia propios identificados por Law et al. (2018).

Implementaciones personalizadas basadas en DigComp

Afortunadamente, la llegada de DigComp facilitó el desarrollo de implementaciones personalizadas por parte de diferentes tipos de organizaciones (Kluzer y Priego, 2018). Teniendo en cuenta el carácter descriptivo y no prescriptivo del marco de referencia, las organizaciones interesadas en llevar a cabo su propia implementación han de adaptarlo a sus requisitos. DigComp se ha utilizado para una variedad de propósitos pragmáticos que se adaptan bien al propósito de la acreditación de CD. Sin embargo, es bastante difícil encontrar una prueba científicamente fiable y válida que evalué las 21 CD de DigComp en los 3 niveles

posibles. Como alternativa a un enfoque pragmático, existen los enfoques psicométricos como las recientes implementaciones guiadas por la MIRT que entiende la CD como un único rasgo latente que debe ser inferido indirectamente a través del análisis estadístico de los resultados de la prueba. Cabe destacar que el supuesto de independencia local puede ser bastante difícil de asegurar en el caso de una evaluación basada en el rendimiento, donde los participantes tengan que resolver una serie de tareas aplicando sus conocimientos.

Respecto a las propiedades psicométricas de las pruebas, hay dos aspectos que son claves: la fiabilidad y la validez.

Por fiabilidad se entiende la precisión de las puntuaciones obtenidas, es decir, la calidad de los datos obtenidos. Se puede decir que una prueba será fiable si cuando se administre a los mismos participantes, obtendremos los mismos resultados. Es un indicador de la estabilidad de las medidas. Los métodos más utilizados para examinar la fiabilidad de las puntuaciones obtenidas son: el coeficiente test-retest, las formas paralelas (o coeficiente de equivalencia), y la consistencia interna. Tradicionalmente, el más utilizado ha sido el coeficiente alfa de Cronbach (1951), el cual expresa la consistencia interna a partir de la covariación entre los ítems de la prueba mediante un número decimal positivo que oscila entre 0,00 y 1,00. Cuanto mayor es la covariación, mayor será el coeficiente alfa y según Barrios y Cosculluela (2013) una fiabilidad adecuada debería fluctuar entre 0,70 y 0,95. Cuando se utiliza el coeficiente alfa de Cronbach, se recomienda complementarlo con otras alternativas, debido a las limitaciones asociadas a este índice como, p. ej. la cantidad de ítems y el número de opciones de respuesta (Ventura-León y Caycho-Rodríguez 2017).

Por validez se entiende la calidad de las inferencias hechas a partir de las puntuaciones obtenidas. En concreto, el proceso de validación consiste en aportar evidencias que permitan soportar las inferencias realizadas, es decir mostrar que están fundadas y que son válidas. Los Estándares para Pruebas Educativas y Psicológicas (AERA, APA, y NCME, 2014), considerados como la posición consensuada y actual de la teoría de la validez (Zumbo, 2014), proporcionaron un marco conceptual para la validación de las pruebas, estableciendo 5 fuentes de evidencia de validez:

- Evidencia de validez basada en el contenido de la prueba. Se utiliza para evaluar cómo el contenido de la prueba está alineado con el propósito de la prueba, y si representa correctamente los conocimientos, habilidades y aptitudes que se desean medir. Habitualmente, este tipo de evidencia suele ser obtenida a través de expertos en el campo de evaluación (Sireci y Faulkner-Bond, 2014).

- Evidencia de validez basada en los RP. Se utiliza para evaluar la forma en que los participantes interactúan con la prueba. Concretamente, se persigue obtener evidencias que permitan confirmar que los participantes hacen uso de los procesos cognitivos previstos cuando responden a los ítems (Ercikan y Pellegrino, 2017; Zumbo y Hubley, 2017). El proceso de obtención de evidencias de este tipo requiere examinar los procesos cognitivos desplegados durante la resolución de la prueba.
- Evidencia de validez basada en la estructura interna. Se evalúa la calidad de las puntuaciones de las pruebas, abarcando desde la precisión y confiabilidad hasta su dimensionalidad. Además, existen más tipos de evidencias de estructura interna como los que incluyen estudios de funcionamiento diferencial de los ítems, en los cuales se examina la invariabilidad de la herramienta de evaluación entre grupos específicos.
- Evidencias de validez basada en relaciones con otras variables. En estos estudios se examinan las puntuaciones obtenidas en las pruebas y otras variables relevantes para el constructo que se está evaluando. Los estudios más habituales utilizan técnicas de regresión o correlación para investigar las relaciones de prueba-criterio.
- Evidencia de validez y consecuencias de las pruebas. En estos estudios se examinan las repercusiones de las pruebas en las personas y en la sociedad. En concreto, determinan si los resultados de una prueba satisfacen su propósito inicial sin causar ningún efecto negativo (Cronbach, 1989; Messick, 1989). Este tipo de evidencias de validez también requieren soportar que el diseño y desarrollo se hayan llevado a cabo de forma que las pruebas sean justas para todos los participantessin que se vean afectados por sus características demográficas o culturales, o por las necesidades especiales que puedan tener (Sireci y Randall, 2021).

Cabe destacar que estas cinco fuentes de evidencia de validez no son distintos tipos de validez, sino que contribuyen al conjunto de evidencias posibles de utilizar con el objetivo de respaldar el uso que se hace de una prueba determinada para un propósito concreto.

A nuestro entender, un enfoque pragmático tiende a dar lugar a una validez interna más débil, pero a una mejor validez externa del instrumento de evaluación, ya que es mejor comprendido, aceptado y adoptado por las distintas organizaciones que puedan estar interesadas. Por consiguiente, en un enfoque pragmático es necesario equilibrar la validez interna y externa, tanto a través de consideraciones metodológicas como del diseño del instrumento de evaluación. De esta manera, en vez de buscar una alta fiabilidad y validez interna, lo que hacemos es priorizar la validez externa y la utilidad percibida de la evaluación por los participantes. Teniendo en cuenta el primero de los objetivos de nuestro estudio, diseñar una herramienta de evaluación de CD suficientemente flexible como para permitir la

incorporación de distintos formatos de ítems, en ECTD optamos por un enfoque similar al sugerido por Laanpere (2019) para el DLGF, es decir, siguiendo un enfoque pragmático en el que evaluamos cada CD como un constructo independiente ignorando las relaciones existentes entre las distintas CD, tal y como se refleja en DigComp.

Respecto a las implementaciones basadas en DigComp identificadas y analizadas por Kluzer y Priego (2018), solo unas pocas abordan la evaluación, el reconocimiento y la certificación de CD mediante un ordenador. De estas, la mayoría de las herramientas de evaluación son pruebas en línea basadas en autoevaluaciones, compuestas por preguntas de opción múltiple y escalas Likert, en las que sólo se considera los componentes de conocimiento y actitud. Sólo unas pocas implementaciones evalúan el componente de habilidad incluyendo, p. ej. simulaciones interactivas o evaluaciones basadas en tareas en las que los participantes deben interactuar con sus estaciones de trabajo (BAIT[21]; IKANOSTEST[22]; PIX[23]; TUCERTICYL[24]; TOSA DigComp Certification[25]). El objetivo principal de estas herramientas no es evaluar las habilidades relacionadas con herramientas específicas, sino proporcionar los elementos para comprender cualquier entorno digital de manera eficaz. Su enfoque proporciona la visión más precisa de la CD de los participantes, ya que éstos deben poner en práctica sus conocimientos. En la figura 2.2 se muestra la interfaz de una prueba de evaluación en BAIT, mostrando una pregunta basada en una simulación interactiva.

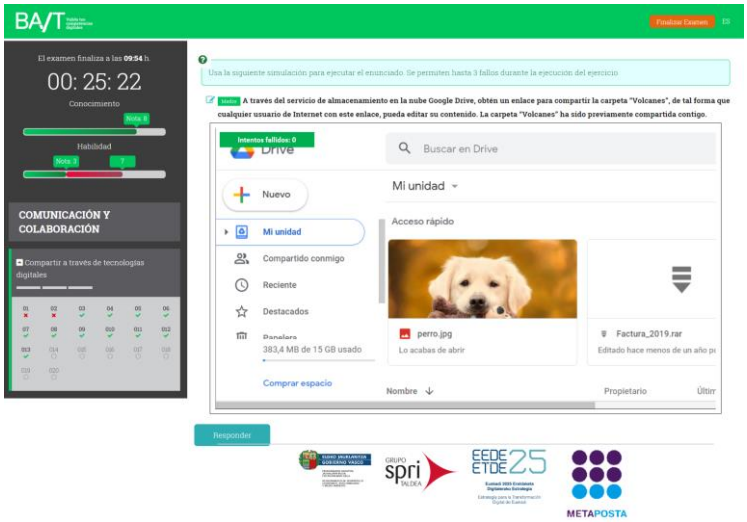


Figura 2.2. Ejemplo de una pregunta basada en una simulación interactiva en BAIT

[21] https://www.bait.eus

[22] https://test.ikanos.eus/

[23] https://pix.org/fr/

[24] https://tucerticyl.es/

[25] https://www.tosa.org/EN/digcomp-certification?sbj_id=217

IKANOSTEST, una de las herramientas pioneras de evaluación basadas en DigComp, proporciona un test de autodiagnóstico, estructurado en 3 bloques temáticos que incluyen una serie de aspectos relacionados con las TIC: (1) desarrollo del potencial; (2) estudios y certificaciones; y (c) CD. El objetivo es identificar los puntos fuertes y las áreas de mejora de los usuarios y crear una experiencia formativa mediante la interpretación de los resultados.

BAIT y TUCERTICYL, están orientadas a la certificación. Ambas incluyen una prueba de demostración con preguntas similares a las incluidas en la prueba de certificación, que sirven de ayuda a los participantes para familiarizarse con el entorno de evaluación. En cambio, PIX ofrece una prueba adaptativa de autodiagnóstico por AC, cuyo objetivo es identificar los niveles de los participantes antes de realizar una prueba de certificación. Éstos pueden seleccionar las CD a incluir en la prueba y el algoritmo adapta el nivel de las pruebas en función de sus respuestas. Estas 3 herramientas, BAIT, TUCERTICYL y PIX, tuvieron en cuenta diferentes dispositivos a la hora de diseñar los ítems de evaluación. Además, implementaron los ítems utilizando principalmente dos alternativas compatibles: simulaciones interactivas y tareas reales a realizar en los puestos de trabajo. Las simulaciones en BAIT y TUCERTICYL se implementaron utilizando el software Articulate Storyline[26], una aplicación para crear medios de aprendizaje basados en la tecnología (Siegel y Hadi, 2017). En ellas, los participantes tienen que examinar una situación dada e interactuar con las simulaciones para llegar a la solución final. En el caso de las tareas reales, se crearon desarrollos a medida donde los participantes tienen que interactuar con sus estaciones de trabajo para llevar a cabo las instrucciones dadas en un enunciado, p. ej. descargar una plantilla, implementar una lista de instrucciones y subir el documento final a la plataforma para ser evaluado automáticamente. La mayoría de estos elementos se integraron en la plataforma mediante distintas APIs.

A nivel de alcance, las distintas implementaciones llevadas a cabo han seguido diferentes enfoques. Por ejemplo, TUCERTICYL ofrece dos certificaciones: ciudadano digital nivel básico y ciudadano digital nivel intermedio. En cada certificación se incluyen todas las AC (unas 63 preguntas por prueba). PIX, en cambio, sólo evalúa 16 competencias. BAIT es el que ha implementado la evaluación en mayor profundidad, basando sus pruebas de certificación en CD y niveles. Actualmente, estas certificaciones sólo están disponibles para los trabajadores públicos del Gobierno Vasco y se espera que a corto plazo se abra a la ciudadanía. Además, BAIT cuenta con una configuración muy flexible, lo que también le permite ofrecer certificaciones basadas en AC. Las pruebas de BAIT, basadas en CD y niveles, pueden aumentar el nivel de exigencia de las certificaciones, incluyendo más ítems y cubriendo más objetivos de aprendizaje o sub-competencias. TOSA DigComp Certification utiliza la misma tecnología de ICDL para la implementación de las preguntas de habilidad y, como ya comentamos previamente, este entorno de evaluación tiene serias limitaciones ya que las pruebas han sido diseñadas en base a

[26] https://articulate.com/360/storyline

software con licencia privativa (p.ej. Microsoft Office). A pesar de este inconveniente, en TOSA DigComp Certification también han incorporado preguntas interactivas haciendo uso de otras aplicaciones y escenarios, pero estas simulaciones son muy sencillas y prácticamente evalúan habilidades cognitivas de bajo orden.

En cuanto a la definición de PD, sólo IKANOS, BAIT y PATHWAYSFOREMPLOY[27] (P4E) (Bartolomé et al., 2022) prestan atención a esta cuestión definiendo PD, para que posteriormente los usuarios puedan evaluar sus CD y comparar sus resultados con los niveles exigidos en el PD. Otras implementaciones menos relevantes se centraron en perfiles específicos como, p. ej. estudiantes universitarios del primer año de enfermería (Evangelinos et al., 2014); ciudadanía y empleados de organizaciones tanto privadas como públicas (Digital Competence[28]); empleados y empresas (SMARTIVEMAP[29]); o empleados y directivos de la región holandesa (Digitale interaktive Didaktik[30]).

Para llevar a cabo una evaluación eficaz de las CD de los participantes, estos deben poder practicarlas en entornos reales y, posteriormente, ser evaluados a través de evaluaciones pertinentes y más cercanas a la práctica real (Litchfield y Dempsey, 2015). La mayoría de las evaluaciones encontradas en la revisión de la literatura, están compuestas de autoevaluaciones con ítems de opción múltiple y respuesta construida (Sparks et al., 2016). Proporcionan una solución tecnológicamente sencilla, pero la mayoría de ellas tienden a evaluar mal las habilidades de los participantes debido al sesgo positivo intrínseco causado en parte por su exceso de confianza (Kruger y Dunning, 1999; Pajares y Graham, 1999). Aunque la confianza de los participantes con las TIC y su motivación para usarlas son importantes para obtener un buen rendimiento (Guillén-Gámez et al., 2018), el sesgo y la precisión de su autoeficacia con las TIC son igualmente relevantes. Por lo tanto, estos sesgos deben ser tenidos en cuenta para mejorar la validez de estas soluciones y su capacidad predictiva (Aesaert et al., 2017). Sin embargo, existe una ambigüedad conceptual en las definiciones y el funcionamiento de los métodos de evaluación, y una tendencia a medir solo habilidades de orden cognitivo bajo, o no cubrir las 5 AC de DigComp al completo (Siddiq et al., 2016). Además, la CD es una expresión observable de conocimientos, habilidades y actitudes desplegadas por los participantes de forma integrada. Con lo cual, para evaluar con precisión su nivel de competencia, se debe observarlos mientras realizan tareas con diferentes niveles de dificultad.

Respecto al tipo de evaluaciones disponibles, según Stephanie et al. (2017) los instrumentos de evaluación comprenden: (a) la autoevaluación, que consiste en una evaluación subjetiva que puede no reflejar realmente la competencia de un examinando; (b) la evaluación basada en el conocimiento, que mide los

[27] http://pathwaysforemploy.com

[28] https://digitalekompetencer.dk/

[29] https://www.smartivemap.com/eng

[30] http://www.forum-did.de/themen/digcomp/

conocimientos de un examinando en un determinado campo; (c) la evaluación basada en el desempeño, en la que un examinando tiene que demostrar su capacidad realizando determinadas tareas; y (d) la recopilación y análisis de datos secundarios, que proporciona información relacionada con un grupo pero no a nivel individual. La mayoría de los sistemas de acreditación y certificación identificados han seguido un enfoque basado en el rendimiento (Law et al., 2018). El objetivo principal es evaluar si los participantes son capaces de poner en práctica sus conocimientos resolviendo determinadas tareas.

Por otra parte, aunque las competencias básicas formuladas en DigComp pueden permanecer estables en el tiempo, en el contexto actual de evolución tecnológica, la construcción de la CD no puede representarse como una entidad fija (Aesaert et al., 2014; Siddiq et al., 2016). Por lo tanto, los descriptores de las CD requieren una revisión constante y lo mismo ocurre con los sistemas de evaluación utilizados. Además, Litchfield y Dempsey (2015) sugirieron que los conceptos y constructos para la evaluación de las CD deberían realizarse desde una perspectiva de dominio (TIC, siglo XXI, etc.) o de conocimiento (alfabetización digital, CD, etc.).

En la tabla 2.3 puede verse una lista de los marcos de referencia de implementaciones basados en DigComp identificados y analizadas para este estudio, con su dirección web (activa en enero de 2023).

Implementación	Destinatario	Finalidad	Web
BAIT	Funcionarios del País Vasco (Próximamente ciudadanía)	Certificación	https://www.bait.eus
ComDIX – Certificación galega de competencias dixitais	Ciudadanía (*)	Certificación	https://competenciasdixitais.xunta.gal/index.html
CRISS	Alumnos de primaria y secundaria	Evaluación y Certificación	https://www.crissh2020.eu/
Digital Competence	Ciudadanía y empleados de organizaciones tanto privadas como públicas	Evaluación	https://digitalekompetencer.dk/
Digitale interaktive Didaktik	Empleados y directivos de la región holandesa	Evaluación	http://www.forum-did.de/themen/digcomp/

Digital Knowledge Certificate (Dig-START, Dig-CERT and DigComp-CERT)	Ciudadanía (*)	Certificación	https://www.etc.at/f4i-dig-cert/ https://www.fit4internet.at/view/FAQ-digcomp-cert
ECCC – European Computer Competence Certificate	Ciudadanía	Certificación	http://www.ecccf.eu/
EDSC DigComp User	Ciudadanía (*)	Certificación	http://bit.ly/3QWbfZB
IDCERT - Digital Competence Specialised Level Certification	Ciudadanía (*)	Certificación	https://it.idcert.io/courses/idcert-digital-competence/details
IKANOSTEST	Ciudadanía y profesionales con distintos PD	Autoevaluación	https://test.ikanos.eus/
PATHWAYSFOREMPLOY (P4E)	Ciudadanía (con especial interés en PD como el del emprendedor o el teletrabajador)	Evaluación y acreditación	http://pathwaysforemploy.com
PIX	Ciudadanía	Certificación	https://pix.org/en-gb/
SMARTIVEMAP	empleados y empresas	Evaluación	https://www.smartivemap.com/eng
TOSA DigComp Certification	Ciudadanía	Certificación	https://www.tosa.org/EN/digcomp-certification?sbj_id=217
TUCERTICYL	Ciudadanía	Certificación	https://tucerticyl.es/

Tabla 2.3. Implementaciones basadas en DigComp: (*) Aún no disponible.

2.2.2. Evaluación de perfiles digitales

El sector digital y, en especial, las CD, ocupan un lugar destacado en la agenda europea desde 2014, con el DSM como una de sus prioridades (EC, 2015). La *Agenda de Competencias para Europa* es una de las herramientas más importantes para poner una serie de iniciativas a disposición de los ciudadanos de la Unión Europea la formación, las competencias y el apoyo adecuados (EC, 2016). La *Digital Skills and Jobs Coalition* es una de estas iniciativas con el objetivo de mejorar las CD de la población en general, no solo de los profesionales de las TI (EC, 2017b). Sin embargo, el 44% de la población europea no tiene ninguna habilidad digital básica, aunque 9 de cada 10 puestos de trabajo requerirán CD pronto (The Digital Skills

Gap in Europe, 2017). A pesar de los avances en este problema, Europa sigue careciendo de la mano de obra digitalmente cualificada necesaria.

La base de datos de la O*NET[31] proporciona descripciones detalladas de las ocupaciones y perfiles competenciales, mostrando las habilidades requeridas para ese trabajo y su importancia (Handel, 2016; Peterson et al., 2001). También se indica el nivel de competencia requerido para cada habilidad, pero sin utilizar un marco de referencia. A nivel europeo, la ESCO[32] proporciona el vocabulario común necesario para facilitar el intercambio de información en lo que se refiere en materia de capacidades y cualificaciones, y más aún, ha aplicado directamente DigComp en la definición de las CD necesarias, pero sin indicar el nivel requerido.

Por último, perfiles como el emprendedor o el teletrabajador no están disponibles en O*NET ni en ESCO, probablemente porque estos perfiles transversales pueden encontrarse como parte de ocupaciones más específicas.

Desde un punto de vista práctico, la definición de un sistema para medir el nivel de CD de los trabajadores en función de sus roles y funciones (es decir, diferentes PDs) podría ser una herramienta útil para aumentar la CD de las diferentes profesiones. Sin embargo, no se ha abordado directamente la definición de qué CD son esenciales para una gran mayoría de PD de manera que puedan satisfacer las necesidades futuras, y menos aún, basándose en un marco de referencia común de CD, que facilite su definición y adopción por las distintas entidades que puedan estar interesadas.

En los siguientes puntos, presentamos los PD del emprendedor y del teletrabajador, elegidos como caso de estudio de Bartolomé et al. (2022), trazando el desarrollo de la CD para ambos perfiles.

Emprendedor

Las TIC han transformado la naturaleza de una gran variedad de procesos y resultados empresariales, p. ej. facilitando las operaciones de creación de nuevas empresas e iniciativas de emprendimiento, así como permitiendo nuevos modelos de negocio digitales (Nambisan et al., 2018; Steininger, 2019). Como resultado, han surgido nuevas competencias para facilitar la transición digital del emprendimiento, las cuales pueden variar en función del contexto en el que se lleve a cabo el emprendimiento (David, 2015). Además, según los datos aportados por Global Entrepreneurship Monitor (2018), el 43% de la población mundial ve buenas oportunidades para iniciar un negocio en los próximos seis meses.

El emprendedor digital se define como un emprendedor en el que parte o la totalidad de las actividades empresariales se producen de forma digital en lugar de en formatos más tradicionales. Los emprendedores digitales deben enfrentarse a retos singulares, como las respuestas instantáneas de los contendientes y la

[31] https://www.onetcenter.org/

[32] https://esco.ec.europa.eu/es

capacidad de teletrabajar y perseguir oportunidades únicas (Hull et al., 2007). Hair et al. (2012) afirmaron que la orientación al mercado es más difícil de conseguir para las nuevas empresas digitales, pero, por otro lado, también disponen de más información sobre sus competidores y clientes, lo que puede utilizarse para entrar en el mercado de forma eficiente proporcionando una clara ventaja competitiva.

Sahut et al. (2019) identificaron cuatro grandes corrientes de investigación en materia de emprendimiento digital: (1) contribuciones del modelo de negocio digital a la literatura de estrategia; (2) digitalización de los procesos de emprendimiento; (3) contribuciones de la plataforma digital a la literatura de estrategia; y (4) literatura del ecosistema emprendedor digital.

De hecho, la capacidad de utilizar eficientemente la última tecnología es relevante en el desarrollo de la actitud emprendedora e intensifica la importancia de desarrollar las CD en lugar de sólo formarse en habilidades técnicas (Scuotto y Morellato, 2013). Los factores y aspectos motivacionales de los emprendedores con éxito han sido identificados y categorizados como: características personales, conocimientos, habilidades técnicas, roles y actitudes (Bi et al., 2017; Dhaliwal y Sahay, 2020; Millán et al., 2019; Sussan y Acs, 2017; Yunis et al., 2018). Oggero et al. (2019) encontraron que los individuos con buen nivel de CD tienen más probabilidades de ser empresarios. A pesar de ello, no se ha realizado un esfuerzo suficiente en el análisis de la importancia de las tecnologías digitales en la configuración de las oportunidades, las decisiones, las acciones y los resultados empresariales (Nambisan, 2017). Otras contribuciones a la investigación sobre el emprendedor digital se han centrado en investigar la adopción, la utilización y las implicaciones de los grandes avances de la tecnología digital, como las aplicaciones móviles o las plataformas tecnológicas digitales (Beliaeva et al., 2019; Durkin et al., 2013; Evans y Schmalensee, 2016; Hayter et al., 2017; McIntyre y Srinivasan, 2017; Nambisan et al., 2018; Olsson y Bernhard, 2020; Srinivasan y Venkatraman, 2018).

Además, la literatura sobre emprendimiento digital carece de estudios centrados en el desarrollo de la capacidad emprendedora, en concreto, poniendo el foco en qué CD son imprescindibles. Sussan y Acs (2017) contextualizaron el emprendimiento en la era digital integrando dos conceptos bien establecidos: El ecosistema digital y el ecosistema emprendedor. Proporcionaron un marco teórico compuesto por cuatro conceptos: (1) Gobernanza de la infraestructura digital, (2) ciudadanía del usuario digital, (3) emprendimiento digital y (4) mercado digital.

El desarrollo de la capacidad empresarial de los ciudadanos y las organizaciones europeas es uno de los objetivos políticos clave de la UE. La Comisión Europea identificó el sentido de la iniciativa y el espíritu empresarial como una de las 8 competencias clave necesarias para una sociedad basada en el conocimiento, y desarrolló *EntreComp: El marco de competencias del espíritu empresarial* (Bacigalupo et al., 2016), identificando los elementos que son necesarios para convertirse en emprendedor y estableciendo una referencia común.

De acuerdo con los estudios analizados en la revisión de la literatura llevada a cabo y los ejemplos identificados por McCallum et al. (2018), no existen ejemplos de casos que traten directamente la definición y evaluación de las CD necesarias para

los emprendedores independientemente del ámbito de actividad. Bartolomé et al. (2022) sólo identificaron 3 casos en los cuales existía un cierto vínculo entre EntreComp y DigComp. ENTRECOM4ALL[33] vinculó las competencias empresariales con las CD pertinentes, aunque la herramienta de autoevaluación sólo identifica y evalúa el nivel de sus competencias empresariales. REACT[34] desarrolló una herramienta de mapeo de prácticas que permite explorar cómo cada práctica aborda los marcos EntreComp y DigComp. Sin embargo, no se tiene en cuenta la definición y evaluación de las CD necesarias para convertirse en un empresario. WOW[35] integró las competencias empresariales con DigComp para demostrar cómo los marcos se complementan entre sí y cómo se podría implementar un modelo conceptual digital de referencia en el ámbito del espíritu empresarial. En particular, se centró en un tipo de negocio específico (un negocio financiado por micromecenazgo) y en un colectivo (mujeres desfavorecidas). Curiosamente, este estudio es el único que se centró en la definición de la CD de las personas interesadas en convertirse en emprendedores y desarrolló una herramienta de evaluación de la CD, la cual sólo evalúa el componente de conocimiento.

En la tabla 2.4 puede verse las principales características identificadas y analizadas en materia de CD para el PD del emprendedor.

Característica	Referencia
Las TIC facilitan la creación de nuevas empresas, iniciativas de emprendimiento, y permiten nuevos modelos de negocio.	(Nambisan et al., 2018; Steininger, 2019)
Nuevas competencias para facilitar la transición digital del emprendimiento.	(David, 2015)
Deben enfrentarse a retos singulares, como las respuestas instantáneas de los contendientes y la capacidad de teletrabajar y perseguir oportunidades únicas.	(Hull et al., 2007)
Disponen de más información sobre sus competidores y clientes, y puede utilizarse para entrar en el mercado de forma eficiente.	(Hair et al., 2012)
Utilizar eficientemente la última tecnología es relevante en el desarrollo de la actitud emprendedora e intensifica la importancia de desarrollar la CD en lugar de sólo formarse en habilidades técnicas	(Scuotto y Morellato, 2013)
Personas con buen nivel de CD tienen más probabilidades de ser empresarios	(Oggero et al., 2019)
Es crucial la adopción, utilización y las implicaciones de los grandes	(Beliaeva et al., 2019; Durkin et al., 2013;

[33] http://entrecom4all.eu/

[34] https://www.reactproject.online/

[35] https://uwawme.eu/

avances de la tecnología digital, como las aplicaciones móviles o las plataformas tecnológicas digitales.	Evans y Schmalensee, 2016; Hayter et al., 2017; McIntyre y Srinivasan, 2017; Nambisan et al., 2018; Olsson y Bernhard, 2020; Srinivasan y Venkatraman, 2018)
El sentido de la iniciativa y el espíritu empresarial como una de las 8 competencias clave necesarias para una sociedad basada en el conocimiento, EntreComp: El marco de competencias del espíritu empresarial.	(Bacigalupo et al., 2016)

Tabla 2.4. Principales características identificadas y analizadas en materia de CD para el PD del emprendedor.

En la tabla 2.5 puede verse las 3 iniciativas identificados y analizadas para este estudio, con su dirección web (activa en enero de 2023).

Iniciativa	Web
ENTRECOM4ALL	http://entrecom4all.eu/
REACT	https://www.reactproject.online/
WOW	https://uwawme.eu/

Tabla 2.5. Iniciativas en las cuales existía un cierto vínculo entre EntreComp y DigComp.

Teletrabajador

Hoy en día, la gran mayoría de los trabajos actuales se apoyan en Internet y/o pueden realizarse desde casi cualquier lugar y en cualquier momento, presentando nuevas oportunidades y retos. Debido al desplazamiento del mercado laboral de la manufactura y la construcción, a los negocios basados en los servicios, el teletrabajo ha sido globalmente aceptado por las empresas. Estos "*teletrabajadores*" (también conocidos como "*trabajadores a distancia*") pueden trabajar de forma independiente o como parte de un equipo en una variedad de acuerdos virtuales. El "*trabajo a distancia*" es un concepto amplio que describe las situaciones en las que el trabajo se realiza total o parcialmente en un lugar alternativo al lugar de trabajo por defecto. El "*teletrabajo*" se considera una subcategoría del trabajo a distancia, en la que el trabajo realizado a distancia implica el uso de dispositivos electrónicos (dispositivos móviles y ordenadores de sobremesa) y de herramientas (paquetes ofimáticos, aplicaciones en la nube, herramientas de reunión, mensajería, herramientas de colaboración, etc.) (Eurofound and the International Labour Office, 2017).

Además, los perfiles competenciales de los teletrabajadores requieren nuevos conjuntos de habilidades, como las habilidades técnicas para funcionar en un lugar de trabajo totalmente remoto y digital, habilidades de resolución de problemas en

un entorno de trabajo habilitado por las TIC, y habilidades sociales necesarias para la interacción no presencial.

La pandemia ha provocado un cambio sustancial en la adopción del teletrabajo oscilando la media de la Unión Europea entre el 4,8 y el 5,5% (Eurostat, 2021). Como consecuencia de la pandemia, esta situación cambió radicalmente y se impulsó definitivamente el teletrabajo en toda la Unión Europea (Sostero et al., 2020). A partir del segundo trimestre de 2020 la media teletrabajadores de la UE subió hasta el 12,3%. Además, tal y como Newman (2016) muestra, la movilidad de los empleados puede mejorar su productividad, lo que supone un 30% de mejora de los procesos, un 23% más de productividad y un 100% más de satisfacción de los empleados. El teletrabajo es una opción cada vez más extendida entre los trabajadores de todo el mundo, con efectos positivos como la reducción del tiempo de desplazamiento, una mayor autonomía del tiempo de trabajo que conduce a una mayor flexibilidad en cuanto a la organización del tiempo de trabajo, un mejor equilibrio general entre la vida laboral y personal, y una mayor productividad. Por otra parte, las empresas ahorran dinero al tiempo que permiten a los trabajadores la libertad de crear sus propios horarios y trabajar desde donde quieran (de Macêdo et al., 2020; Eurofound and the International Labour Office, 2017). El teletrabajo no es simplemente una nueva forma de trabajar, establece una nueva forma de organización con nuevos tipos de tareas, problemas más complicados y diferentes responsabilidades de gestión (Mahler, 2012). No obstante, el trabajo a distancia es, en general, ventajoso para los empleadores y los empleados, y la desconexión del trabajo del lugar forma parte, sin duda, de la naturaleza cambiante del trabajo en el siglo XXI (Felstead y Henseke, 2017). Sostero et al. (2020) desarrollaron un índice de teletrabajo basado en un marco conceptual y una taxonomía de tareas para el análisis ocupacional creados por Fernández-Macías et al. (2016) y utilizando datos de encuestas ocupacionales europeas. Dingel y Neiman (2020) siguieron un enfoque similar para desarrollar el perfil del trabajo desde casa utilizando encuestas de la O*NET y descubrieron que el 37% de los trabajos en Estados Unidos pueden realizarse completamente en casa.

La digitalización del trabajo impuesta por el teletrabajo ha subrayado la importancia de que los empleados mejoren sus CD y cómo las lagunas existentes pueden afectar a su rendimiento (Zamfir y Aldea, 2020). Estudios anteriores sugirieron que la formación sobre la configuración de equipos, conectividad, el uso en ubicaciones remotas y la resolución de problemas son esenciales, a pesar de que la formación se identificó de dificultad baja en comparación con otras cuestiones de tecnologías de la información (Dingel y Neiman, 2020; Garrett y Danziger, 2007; Haddon y Brynin, 2005; Harmer y Pauleen, 2012; Sostero et al., 2020; Vargas-Llave et al., 2020). Vargas-Llave et al. (2020) afirmaron que las CD no sólo son cruciales para acceder al trabajo, sino también para la autopromoción y la construcción de una reputación en línea que garantice las oportunidades de empleo y amplíe las perspectivas de carrera. En la comparación internacional de marcos de competencias del siglo XXI llevada a cabo por Voogt y Roblin (2012), identificaron la recuperación y el procesamiento de la información digital, y la comunicación a través de dispositivos digitales como dos componentes esenciales de la

competencia. Además, Aesaert et al. (2014) identificaron la recuperación y el procesamiento de la información digital, así como la comunicación de forma segura, sensata y apropiada como dos temas presentes en los planes de estudio nacionales sobre las TIC.

Hasta donde sabemos, ninguna investigación anterior ha abordado o discutido las CD en profundidad que son esenciales para cualquier persona interesada en trabajar a distancia, teniendo en cuenta el amplio espectro de tipos de teletrabajo y la naturaleza de sus condiciones de trabajo en las que estas competencias son fundamentales.

En la tabla 2.6 puede verse las principales características identificadas y analizadas en materia de CD para el PD del teletrabajador.

Característica	Referencia
El trabajo realizado a distancia implica el uso de dispositivos electrónicos (dispositivos móviles y ordenadores de sobremesa) y de herramientas (paquetes ofimáticos, aplicaciones en la nube, herramientas de reunión, mensajería, herramientas de colaboración, etc.)	(Eurofound and the International Labour Office, 2017)
Nueva forma de organización con nuevos tipos de tareas, problemas más complicados y diferentes responsabilidades de gestión	(Mahler, 2012)
Desconexión del lugar de trabajo de forma física.	(Felstead y Henseke, 2017)
Índice de teletrabajo basado en un marco conceptual y una taxonomía de tareas para el análisis ocupacional.	(Fernández-Macías et al., 2016; Sostero et al., 2020)
Libertad de crear sus propios horarios y trabajar desde donde quieran.	(de Macêdo et al., 2020; Eurofound and the International Labour Office, 2017)
Perfil del trabajo desde casa utilizando encuestas de la O*NET.	(Dingel y Neiman, 2020)
Importancia de la mejora de la CD y cómo las lagunas existentes puede afectar al rendimiento del trabajador.	(Zamfir y Aldea, 2020)
Es esencial la formación sobre la configuración de equipos, conectividad, el uso en ubicaciones remotas y la resolución de problemas.	(Dingel y Neiman, 2020; Garrett y Danziger, 2007; Haddon y Brynin, 2005; Harmer y Pauleen, 2012; Sostero et al., 2020; Vargas-Llave et al., 2020)
Las CD son cruciales para autopromoción y la construcción de una reputación en línea que garantice las oportunidades de empleo y amplíe las perspectivas de carrera.	Vargas-Llave et al. (2020)
La recuperación y el procesamiento de la información digital, y la	(Aesaert et al., 2014; Voogt y

comunicación a través de dispositivos digitales como dos componentes esenciales de la competencia.	Roblin, 2012)

Tabla 2.6. Principales características identificadas y analizadas en materia de CD para el PD del teletrabajador.

2.2.3. Evaluación de la CD de la ciudadanía

Partiendo de DigComp, seleccionamos dos casos de estudio, Información y alfabetización digital (IAD) y Netiqueta, ya que representan dos enfoques diferentes que están siendo adoptados por algunas de las iniciativas más relevantes identificadas como casos de éxito por Kluzer y Priego (2018), que consisten en pruebas basadas en AC o en CD respectivamente.

Para la selección del AC, elegimos una de las tres principales AC de DigComp, IAD, pero se podría haber elegido cualquiera de las 3 AC principales de DigComp priorizando sobre las otras dos son transversales (Seguridad y Resolución de problemas). Para la selección de la CD elegimos una CD que habitualmente no suele evaluarse en profundidad. Más aún, los resultados de incluir ítems con formatos dinámicos serán más notables que en otras CD.

Información y alfabetización digital

De acuerdo con DigComp, IAD es una AC compuesta por tres CD (ver tabla 2.7).

CD	Descripción
Navegación, búsqueda y filtrado de datos, información y contenidos digitales	Articular las necesidades de información, buscar datos, información y contenidos en entornos digitales, acceder a ellos y navegar entre ellos. Crear y actualizar información personal
Evaluación de datos, información y contenidos digitales	Analizar, comparar y evaluar críticamente la credibilidad y fiabilidad de las fuentes de información y contenidos digitales. Analizar, interpretar y evaluar críticamente los datos información y contenidos digitales.
Gestión de datos, información y contenidos digitales	Organizar, almacenar y recuperar datos, información y contenidos en entornos digitales. Organizarlos y procesarlos en un entorno estructurado.

Tabla 2.7. IAD y CD comprendidas según Stephanie et al. (2017).

También conocida como *alfabetización informacional* o *alfabetización informacional digital*, está en constante evolución debido a los cambios recurrentes en la forma en que los ciudadanos acceden y gestionan la información a través de diferentes tipos de dispositivos. Los ciudadanos, y especialmente los jóvenes, están sustituyendo los medios de comunicación tradicionales por las redes sociales, pero que al mismo

tiempo suponen una fuente de información no controlada que tiende a crear confusión, generar controversias y desconfianza (Dessart, 2017; Pérez-Escoda y Esteban, 2021; Pérez-Escoda, A., Pedrero-Esteban et al., 2021), permiten a los usuarios ser creadores activos de contenidos (Larrondo Ureta et al. 2020), e influyen en la elección de modelos de conducta de los jóvenes (Castillo-Abdul et al., 2020). Además, la facilidad y rapidez de propagación que proporcionan las redes sociales para la desinformación se ha convertido en una de las amenazas más peligrosas (Masip et al., 2020; Orso et al., 2020; Viner, 2016; Vraga y Bode, 2020), junto con la aparición de discursos basados en la apelación emocional para influir en la elección haciendo uso de diferentes vías como el "*click baiting*" ("enlaces cebo"), los algoritmos basados en inteligencia artificial, la creación de burbujas filtradas, la personalización de la información, etc. (Kopecky et al., 2020; Orso et al., 2020; Viner, 2016). El Eurobarómetro (2020) ya mostró un aumento de la preocupación por cuestiones como el rápido crecimiento de las noticias falsas (74%) y hacia los medios de comunicación social (65%). IAD ha sido identificada como una alfabetización clave para identificar las noticias falsas (Jones-Jang et al., 2021). Así que, en este contexto, es necesario examinar y evaluar cómo los ciudadanos perciben y evalúan los medios de comunicación en términos de noticias falsas.

Respecto a su evaluación, existen muchas herramientas de autoevaluación en las cuales los participantes deben autoevaluar su nivel, y la mayoría de estas herramientas están compuestas por preguntas de opción múltiple que evalúan habilidades cognitivas de orden bajo (Catalano, 2015; Foo et al., 2017; Kluzer y Priego, 2018; Siddiq et al., 2016; Walsh, 2009). Por una parte, las autoevaluaciones ofrecen una solución fácil de aplicar, pero por otra parte tienden a obtener resultados poco realistas de los participantes, en parte debido a su exceso de confianza, especialmente los participantes con una capacidad muy baja (Kruger y Dunning, 1999; Mahmood, 2016). También hay algunas excepciones en lo que a instrumentos se refiere, p. ej. el uso de tareas abiertas con rúbricas de puntuación (Leichner et al., 2013; Markowski et al., 2018), pero estas alternativas serían muy complicadas de integrar en un contexto de certificación que requiere entornos acotados, seguros y que ofrezcan los resultados de manera inmediata.

Desde el punto de vista de la operacionalización del constructo de IAD con fines de evaluación, Sparks et al. (2016) indicaron que los diseñadores de pruebas se decantan por dos enfoques posibles: (1) seleccionar un marco concreto alineado con el constructo definido en su aplicación y, a continuación, diseñar los ítems de acuerdo con los descriptores del marco (esta opción es adecuada para evaluar un conjunto específico de habilidades) o (2) operacionalizar el constructo a nivel conceptual, desarrollando así tareas más cercanas a la práctica real que evalúen IAD de forma más amplia. Esta opción es adecuada para definir el constructo de forma más holística y examinar si los examinados pueden poner en práctica sus conocimientos en un contexto real. Por lo tanto, es crucial definir al comienzo los objetivos de aprendizaje previstos y el tipo de evaluación prevista. De hecho, más allá de la definición del constructo, en el desarrollo de la evaluación deben tenerse en cuenta otras cuestiones como, p. ej. los contextos en los que se va a acceder,

evaluar y utilizar la información o si una tecnología específica es un objetivo de evaluación en sí mismo o constituye una forma de alcanzar un objetivo.

Respecto al tipo herramientas de evaluación y los tipos de preguntas utilizados, Sparks et al. (2016) las clasificaron como: (1) compuestas por preguntas de respuesta construida centradas en el IAD, como el International Computer and Information Literacy Study (ICILS)[36]; (2) compuestos por preguntas de respuesta construida centradas en la alfabetización tecnológica, como el ICDL; y (3) consistentes en tareas basadas en el rendimiento centradas en IAD.

La evaluación de IAD en la educación superior también es un tema clave (ACRL, 2016; Sparks et al., 2016), y el interés por desarrollar instrumentos para evaluar IAD ha crecido en los últimos años. Sin embargo, la mayoría de las pruebas se desarrollan desde dos perspectivas, la bibliotecaria y la académica, y suelen ser específicas de un dominio (Catalano, 2016; Hollis, 2018).

En cuanto a la calidad de los instrumentos de evaluación, en la mayoría de las pruebas identificadas en la revisión sistemática llevada a cabo por Mahmood (2017), se aplicó la teoría clásica de los test, y los análisis más comúnmente realizados fueron la validez de contenido y discriminante, y la fiabilidad de consistencia interna. En consecuencia, los expertos han manifestado la necesidad de disponer de instrumentos de evaluación libres para medir la IAD, realizando una evaluación más eficaz, validada e independiente del dominio y del contexto (Hollis, 2018).

En la tabla 2.8 puede verse las principales referencias identificadas y analizadas en materia de alfabetización informacional o alfabetización informacional digital.

Característica	Referencia
Sustitución de los medios de comunicación tradicionales.	(Dessart, 2017; Pérez-Escoda y Esteban, 2021; Pérez-Escoda, A., Pedrero-Esteban et al., 2021)
Las redes sociales permiten a los usuarios ser creadores activos de contenidos.	(Larrondo Ureta et al. 2020)
Redes sociales que influyen en la elección de modelos de conducta de los jóvenes.	(Castillo-Abdul et al., 2020)
Amenaza debida a la facilidad y rapidez de propagación que proporcionan las redes sociales para la desinformación.	(Masip et al., 2020; Orso et al., 2020; Viner, 2016; Vraga y Bode, 2020)
Amenaza por la aparición de discursos basados en la apelación emocional para influir en la elección, algoritmos basados en inteligencia artificial, la creación de burbujas filtradas, la	(Kopecky et al., 2020; Orso et al., 2020; Viner, 2016)

[36] https://www.iea.nl/studies/iea/icils

personalización de la información, etc.	
Aumento de la preocupación por el crecimiento de las noticias falsas y hacia los medios de comunicación social.	(Eurobarómetro, 2020)
Identificada como una alfabetización clave para identificar las noticias falsas.	(Jones-Jang et al., 2021)

Tabla 2.8. Principales referencias identificadas y analizadas en materia de alfabetización informacional o alfabetización informacional digital.

En la tabla 2.9 puede verse las principales referencias identificadas y analizadas en materia evaluación de la alfabetización informacional o alfabetización informacional digital.

Característica	Referencia
La mayoría son autoevaluaciones compuestas por preguntas de opción múltiple que evalúan habilidades cognitivas de orden bajo, y que tienden a obtener resultados poco realistas, en parte debido al exceso de confianza de quienes se evaluan.	(Catalano, 2015; Foo et al., 2017; Kluzer y Priego, 2018; Kruger y Dunning, 1999; Mahmood, 2016; Siddiq et al., 2016; Walsh, 2009)
A nivel de operacionalización del constructo de IAD con fines de evaluación, los diseñadores de pruebas se decantan por dos enfoques: (1) seleccionar un marco concreto alineado con el constructo definido en su aplicación, y diseñar los ítems de acuerdo con los descriptores del marco ; (2) operacionalizar el constructo a nivel conceptual, desarrollando así tareas más cercanas a la práctica real.	(Sparks et al., 2016)
Los distintos tipos de herramientas de evaluación y tipos de preguntas utilizados: (1) compuestas por preguntas de respuesta construida centradas en el IAD; (2) compuestos por preguntas de respuesta construida centradas en la alfabetización tecnológica; y (3) consistentes en tareas basadas en el rendimiento centradas en IAD.	(Sparks et al., 2016)
La mayoría de las pruebas se desarrollan desde dos perspectivas, la bibliotecaria y la académica, y suelen ser específicas de un dominio.	(Catalano, 2016; Hollis, 2018; Sparks et al., 2016)
En cuanto a la calidad de los instrumentos de evaluación, en la mayoría se aplicó la TCT, y los análisis más realizados fueron la validez de contenido y discriminante, y la fiabilidad de consistencia interna.	(Mahmood, 2017)
Necesidad de disponer de instrumentos de evaluación libres, realizando una evaluación más eficaz, validada e independiente del dominio y del contexto.	(Hollis, 2018)

Tabla 2.9. Principales referencias identificadas y analizadas en materia de evaluación de la alfabetización informacional o alfabetización informacional digital.

Netiqueta

Según DigComp, la netiqueta es una de las 6 CD del AC de *Comunicación y Colaboración*, definida como "*Ser consciente de las normas de comportamiento y de los conocimientos técnicos al utilizar las tecnologías digitales e interactuar en entornos digitales. Adaptar las estrategias de comunicación a la audiencia específica y ser consciente de la diversidad cultural y generacional en los entornos digitales*".

En la sociedad actual, en la que las TIC impregnan la mayoría de los ámbitos, las redes sociales y el uso extensivo de los dispositivos móviles, han modificado radicalmente la forma de interactuar entre las personas, y en consecuencia la netiqueta se está mostrando como una CD crucial (Vaterlaus et al., 2021). Surge un nuevo escenario para entender las relaciones humanas, desde cómo se ejercen las habilidades interpersonales en línea hasta cómo se exhiben los comportamientos sociales en grupos y comunidades online (del Carmen García Galera et al., 2017). Cabezas-González et al. (2021) descubrieron que los individuos que se comunican en línea con frecuencia y hacen uso de las redes sociales con mucha frecuencia tienden a mostrar niveles más bajos de CD, en contra de lo esperado. Por lo tanto, es de gran importancia investigar la formación actual en *Comunicación y Colaboración* haciendo uso de la tecnología, pero también teniendo a la netiqueta presente (Kozík y Slivová, 2014). Sin embargo, la netiqueta apenas se ha definido y todavía no parece haber atraído la atención necesaria (Soler-Costa et al., 2021). Sólo unos pocos estudios han analizado las directrices relacionadas con el uso correcto del correo electrónico (p. ej. Brusco (2011) o Hammond y Moseley (2018)), o han presentado directrices generales para Internet (p. ej. McMurdo (1995)). Ningún estudio ha intentado definir qué CD debe tener un ciudadano para comunicarse de forma eficiente a través de herramientas cotidianas como las aplicaciones de mensajería instantánea, las redes sociales o el correo electrónico, teniendo a la netiqueta presente. Por ello, es necesario revisar los antecedentes teóricos y analizar las propuestas experimentales.

En cuanto a las herramientas de evaluación que incluyen la netiqueta, los artículos empíricos identificados incluían la elaboración de pruebas a medida, como las de Arouri y Hamaidi (2017) o Linek y Ostermaier-Grabow (2018), cuyas pruebas de validez y fiabilidad eran insuficientes. En términos generales, la netiqueta aún no se ha evaluado en profundidad y la mayoría de ellas sólo incluyen algunas preguntas generales (Kluzer y Priego, 2018; Law et al., 2018). Desde un punto de vista más amplio, los expertos como Siddiq et al. (2016), ya identificaron la falta de instrumentos para evaluar la CD de los individuos en el AC de *Comunicación y Colaboración en línea*. Sólo la plataforma de evaluación y certificación BAIT, estrechamente relacionada con este estudio, ofrece actualmente una prueba dedicada exclusivamente a la evaluación de la CD de netiqueta.

En la tabla 2.10 puede verse las principales referencias identificadas y analizadas en materia de netiqueta.

Característica	Referencia
Las redes sociales y el uso extensivo de los dispositivos móviles, han modificado radicalmente la forma de interactuar entre las personas, surgiendo un nuevo escenario para entender las relaciones humanas.	(del Carmen García Galera et al., 2017; Vaterlaus et al., 2021)
Las personas que con frecuencia se comunican en línea con frecuencia y hacen uso de las redes sociales tienden a mostrar niveles más bajos de CD.	(Cabezas-González et al., 2021)
Es crucial investigar la formación actual en Comunicación y Colaboración en línea, teniendo a la netiqueta presente.	(Kozík y Slivová, 2014)
La netiqueta apenas se ha definido y todavía no ha captado la atención necesaria.	(Soler-Costa et al., 2021)
Pocos estudios han analizado directrices relacionadas con el uso correcto del correo electrónico, o generales para Internet.	(Brusco, 2011; Hammond y Moseley, 2018; McMurdo, 1995)

Tabla 2.10. Principales referencias identificadas y analizadas en materia de netiqueta.

En la tabla 2.11 puede verse las principales referencias identificadas y analizadas en materia evaluación de netiqueta.

Característica	Referencia
Existencia de apenas unas pocos pruebas a medida, cuyas pruebas de validez y fiabilidad son insuficientes.	(Arouri y Hamaidi, 2017; Linek y Ostermaier-Grabow, 2018)
La netiqueta aún no se ha evaluado en profundidad.	(Kluzer y Priego, 2018; Law et al., 2018)
En general, falta de instrumentos para evaluar la CD de los individuos en el AC de "Comunicación y Colaboración en línea".	(Siddiq et al., 2016)

Tabla 2.11. Principales referencias identificadas y analizadas en materia de evaluación de netiqueta.

2.2.4. Consideraciones para este estudio en términos de evaluación y acreditación de CD

Para llevar a cabo este estudio nos basamos en DigComp como marco de referencia por sus fortalezas tal y como hemos desarrollado a lo largo de este capítulo, mismas razones por las cuales también fue el marco de referencia de CD seleccionado por BAIT, el cual está estrechamente relacionado con este estudio.

Cabe destacar que a pesar de que los componentes de la CD son el conocimiento, la habilidad y la actitud, decidimos dejar la evaluación del componente actitudinal fuera del ámbito de este estudio debido a su complejidad y

a la falta de consenso a la hora de evaluarlo. De hecho, en BAIT también se ha optado por dejar el componente actitudinal fuera de las pruebas de evaluación.

Adoptamos un enfoque de evaluación basada en el desempeño a través de tareas reales o simuladas, para evaluar los conocimientos, las habilidades y los hábitos de trabajo a través de la realización de tareas que son significativas y atractivas para los participantes. Los participantes tienen que acceder a materiales, interactuar con programas y servicios, crear nuevos contenidos, buscar y evaluar la información encontrada, comunicar y compartir información.

Por último, la validez representa la forma en que los resultados de la prueba pueden interpretarse y utilizarse de acuerdo con los objetivos perseguidos en la evaluación (AERA, APA y NCME, 1999). En el contexto de la prueba, la validez interna representa cómo una prueba apoya una afirmación sobre causa y efecto, mientras que la validez externa puede entenderse como su potencial para poder extender las conclusiones obtenidas a otros contextos ajenos al estudio. El enfoque pragmático basado en DigComp utilizado puede afectar a la validez interna, pero por otro lado debería favorecer la validez externa. Nuestro objetivo fue lograr un equilibrio con la validez de la prueba a través del diseño del instrumento de evaluación, para que sea entendido y adoptado entidades que puedan estar interesadas, entendiendo las decisiones tomadas.

Tras todo este análisis y considerando que no podemos abordar todos los puntos identificados en una sola investigación, hemos priorizado y seleccionado un conjunto de características en las cuales centrar nuestra experimentación. Al no encontrar ninguna herramienta que las cumpla todas, decidimos desarrollar nuestras propias herramientas, que presentamos en los siguientes capítulos. Presentamos estas características en la tabla 2.12, en la que a título informativo hemos incluido BAIT, herramienta origen y clave de esta investigación, así como nuestras dos herramientas P4E y ETCD.

Herramienta	Des	Fin	Enf	Com	For	Dis	Psi	Ada	Cog	Prf
BAIT	CI, PD	CE	CN	C, H	OM, IR, SI, TR	TO	NI	NO	A	A
ComDIX	CI	CE	PD + MC	NI	NI	NI	NI	NO	NI	NI
Digital Competence	CI, PD	AU	MC	A	LK	TO	NI	NO	B	B
Digital Knowledge Certificate	CI	CE	MC	C	OM	TO	NI	NO	B	B
ECCC	CI	CE	AC	C, H, A	NI	TO	NI	NO	A	A
ECTD	CI	EV, CE	CD, AC	C, H	OM, IR, SI, TR	TO	SI	EI	A	A
EDSC DigComp	CI	CE	MC	NI	NI	TO	NI	NO	NI	NI

User										
IDCERT	CI	CE	MC	NI	NI	TO	NI	NO	NI	NI
IKANOSTEST	CI, PD	AU	MC	C, A	OM, IR, LK	TO	NI	NO	B	B
P4E	CI, PD	EV, AC	PD + AC	C, H, A	OM, IR, SI, TR	TO	SI	NO	A	M
PIX	CI	CE	MN + AC	C, A	OM, IR, SI, TR	TO	NI	NI	A	A
SMARTIVEMAP	PD	EV	MC	C, A	OM, LK	TO	NI	NO	B	B
TOSA DigComp Certification	CI	CE	MC	C, H	OM, IR, SI	TR	NI	NI	A	M
TUCERTICYL	CI	CE	PD + MC	C, H	OM, IR, SI	TO	NI	NO	A	M

Tabla 2.12. Principales características de las implementaciones basadas en DigComp identificadas y analizadas: Des= Destinatario (CI=ciudadanía/PD/AL=alumnos), Fin=Finalidad (AU=autoevaluación, EV=evaluación, AC=acreditación, CE=certificación), Enf=Enfoque (CD/CN=competencia y nivel/AC/MC=marco completo/PD/MN=marco no completo del todo), Com= Componentes de CD evaluados (C=conocimiento, H=habilidad, A=actitud, NI=sin información), For= Formatos de preguntas incluidas (OM=opción múltiple/IR=ver imagen y responder/SI=simulaciones interactivas/TR=tareas reales interactuando con el puesto, LK=escalas likert), Dis= Dispositivos y sistemas operativos abarcados (PC=PC/laptop, MO=móvil, TA=tableta, OT=otros, MW=Microsoft Windows, MO=Microsoft Office, TO=Todos, TR=Todos pero muy dependiente de soluciones de Microsoft), Psi= Propiedades psicométricas (SI, NO, ES=escasas, NI=sin información), Ada= Motor adaptativo (SI, NO, NI=si pero sin información, EI=estudios iniciales), Cog = Orden cognitivo evaluado (B=bajo, M=medio, A=avanzado, NI=sin información), Prf=Profundidad de la evaluación (A=alta, M=media, B=baja, NI=sin información).

2.3. Evaluación mediante el uso de la tecnología y analíticas de evaluación

La llegada de DigComp ha facilitado el desarrollo de implementaciones personalizadas de marcos de evaluación y certificación de CD proporcionando un lenguaje común para entender y hablar de la CD (Kluzer y Priego, 2018). Sin embargo, la mayoría de las implementaciones relacionadas con la evaluación de la CD o el reconocimiento y la certificación, proporcionan herramientas en línea basadas en la autoevaluación, compuestas por preguntas de opción múltiple y escalas Likert, donde apenas se evalúa el componente de habilidad de la CD.

La evaluación digital ofrece inmensas oportunidades para mejorar la experiencia de los participantes y desarrollar modos de evaluación más pertinentes y ajustados a las necesidades actuales, tal y como una gran variedad de autores han destacado en sus estudios como, p. ej. Cho et al. (2019), Debuse y Lawley (2016), Drasgow (2016), Scherer et al. (2017), Shute y Rahimi (2017), Stödberg (2012), o Zenisky y Luecht (2016). Esto no ha pasado desapercibido para los gobiernos, donde el potencial de la tecnología digital para promover y medir las habilidades del siglo XXI necesarias para la prosperidad económica, como el pensamiento

crítico y la resolución de problemas en colaboración, tal y como ha quedado reflejado en una gran variedad de informes de todo el mundo (OCDE, 2016).

El uso de la TEA está muy extendido en la evaluación de la CD, posibilitando el uso de entornos simulados (Binkley et al., 2012), y oportunidades para aplicar el conocimiento en un entorno seguro (Scherer et al., 2014). La TEA tiene un enorme potencial para proporcionar formatos de ítems innovadores y más cercanos a la práctica real, así como la posibilidad de obtener información sobre el comportamiento y el rendimiento de los examinados durante las pruebas de evaluación, recogiendo diferentes tipos de datos como datos de resultados, tiempo de respuesta o flujos de clics (Bartolomé y Garaizar, 2022; Greiff et al., 2015; Osborne et al., 2013; Timmis et al., 2016). Además, numerosos autores han utilizado estos datos para mejorar la interpretación de las puntuaciones o para contribuir a la evaluación de la validez (Oranje et al., (2017)).

Desde el punto de vista del diseño de las pruebas de evaluación, en los últimos años un creciente cuerpo de literatura ha examinado y encontrado una fuerte conexión entre el rendimiento y el compromiso de un examinando y la forma en que las tareas y evaluaciones estaban diseñadas en una prueba (Nguyen, Rienties et al., 2017; Papamitsiou y Economides, 2016; Rienties y Toetenel, 2016). Esto supone un reto, especialmente cuando se evalúan constructos cognitivos complejos en los que los diseñadores de pruebas utilizan formatos dinámicos como las simulaciones interactivas o los juegos serios (Van Voorhis y Paris, 2019) o simulaciones de laboratorio en realidad virtual (O'Leary et al., 2018). Para la interpretación de las respuestas obtenidas es esencial saber hasta qué punto los formatos de las preguntas de evaluación podrían utilizarse para obtener evidencias de las capacidades de los participantes para los niveles de CD a la que se dirige la evaluación. Habitualmente, las pruebas de evaluación suelen incluir diferentes tipos de preguntas para permitir tales inferencias. Sin embargo, el gran esfuerzo requerido para desarrollar tipos específicos de preguntas, como los formatos dinámicos, limitan el alcance de las inferencias. Además, las respuestas de los examinados a las tareas de evaluación no suelen proporcionar suficiente información sobre el proceso de respuesta (RP) desplegado por los examinados para llegar a sus respuestas. Por lo tanto, es muy complicado confirmar si los examinados se involucraron en las tareas de las formas esperadas o mostraron diferentes estrategias para llegar a las mismas respuestas finales.

Según el metaanálisis reciente llevado a cabo por Papamitsiou y Economides (2016), es necesario llevar a cabo un análisis más detallado de las estrategias de diseño de la evaluación y del aprendizaje para comprender el contexto en el que tienen lugar el aprendizaje y las estrategias de TEA. El análisis de los datos generados a partir de los RP durante las evaluaciones puede ayudar a validar el diseño de los ítems de evaluación, presentando evidencias de que los ítems desencadenaron los conocimientos y habilidades esperados. Sin embargo, examinar si los examinados responden a las tareas realizando RP relevantes para el constructo, no suele tenerse en cuenta a la hora de examinar la validez de una prueba. Recientes revisiones sobre la fiabilidad y la validez de las herramientas para la evaluación de la CD han llegado a la conclusión de que se dispone de muy poca

información sobre las formas de garantizar la validez y la fiabilidad del instrumento utilizado. Y aún más, de los diferentes tipos posibles de evaluación de la calidad, las evidencias basadas en RP son escasas o inexistentes (Saltos-Rivas et al., 2021; Siddiq et al., 2016; Zhao et al., 2021).

Los métodos de ET, amplían aún más la información basada en los RP, proporcionando observaciones del examinando a un nivel más detallado (Oranje et al., 2017). Sin embargo, hay relativamente pocos estudios que abarquen ET en el contexto de la evaluación educativa y menos todavía en el área de la evaluación de la CD (Bartolomé et al., 2020).

2.3.1. Evaluación mediada por la tecnología (TEA)

Durante las últimas décadas, las tecnologías existentes y emergentes están empezando a desempeñar un papel en el cambio de la evaluación, contribuyendo para que la evaluación sea más inteligente, rápida, justa y eficaz (Pauli y Ferrell, 2020). Si se lleva a cabo correctamente, impulsa la mejora, moldea el comportamiento de los alumnos y proporciona responsabilidad a los empleadores y a otras personas (Pauli y Ferrell, 2020). Las principales aportaciones de la TEA son:

- La evaluación y el modelado psicométrico de constructos complejos como la resolución de problemas colaborativos y el pensamiento computacional, haciendo uso de formatos de ítems innovadores más cercanos a la práctica real (Graesser et al., 2017; Greiff et al., 2013; Grover y Pea, 2013; Kuo y Wu, 2013; Mayrath et al., 2012; Scherer, 2015; Shute y Rahimi, 2017; Van Voorhis y Paris, 2019; von Davier et al., 2017). La TEA permite diseñar ítems más cercanos a la práctica real, es decir, poniendo a prueba los conocimientos y habilidades de una manera más realista, contextualizada y motivadora. La evaluación puede transformarse en un proceso interactivo que va más allá de la medición del recuerdo de conocimientos y capta el rendimiento en tareas complejas (Shute et al., 2016). En este caso, es crucial entender qué funciona y qué no, para que las evaluaciones puedan proporcionar medidas fiables y válidas del constructo en cuestión (Duckworth y Yeager, 2015). De especial interés es el estudio realizado por los autores Engelhardt et al. (2017), que se centraron en la evaluación de las competencias TIC y examinaron los efectos de las características de la tarea en la validez de constructo.
- La automatización de la puntuación (Gierl et al., 2014; Kuo y Wu, 2013; Paiva et al., 2022; Taub et al.,2017; Vista et al., 2017; Zechner et al., 2017), el diseño y desarrollo de pruebas (Engelhardt et al., 2017; Nguyen et al., 2017), y del ensamblado de pruebas (Veldkamp, 2015; Veldkamp et al., 2017).
- La disponibilidad de datos de proceso para describir no solo el rendimiento en base al resultado final (correcto e incorrecto), sino también el comportamiento estratégico mostrado, las secuencias llevadas a cabo y los patrones de respuesta (Greiff et al., 2016; Sweeney et al., 2017; Timmis et al., 2016; Veldkamp et al., 2017).

- La provisión de comentarios constructivos, adecuados y fáciles de entender (Carless et al., 2011; Debuse y Lawley, 2016; Eggen et al., 2011; Helfaya y O'Neill, 2018; Helfaya, 2019; Makransky et al., 2020; Van der Kleij et al., 2012).
- El diseño de pruebas accesibles para que todos puedan utilizarlas en la mayor medida posible, incluidas las personas con algún tipo de discapacidad.
- La provisión de entornos seguros para la realización de las pruebas, asegurándose de que cada participante realiza la prueba cumpliendo las normas establecidas.

La TEA puede permitir que las evaluaciones se adapten mejor a cada participante. Con un banco de ítems suficientemente grande, bien diseñados y haciendo uso de un algoritmo adaptativo, es posible administrar diferentes versiones de una prueba a diferentes grupos de participantes de acuerdo con las respuestas que se vayan dando. De este modo, la evaluación puede adaptarse al nivel de conocimientos de cada persona, logrando mayor rapidez al evitar las preguntas demasiado difíciles y fáciles, y además pudiendo llegar a ser más precisa, ya que permite determinar con detalle las capacidades de cada participante.

Alejándose de las pruebas tradicionales, los TEA están permitiendo ítems más cercanos a la práctica real, p. ej. pidiendo a los participantes que geolocalicen un sitio desde su dispositivo móvil, que creen documentos en línea y los compartan con otras personas, que graben y editen vídeos, o que utilicen las redes sociales para compartir contenidos (OECD, 2013), además de posibilitar la obtención de información sobre el comportamiento de los participantes durante la realización de las pruebas (Drasgow, 2016; Goldhammer et al., 2016; Greiff et al., 2015; Osborne et al.,2013; Timmis et al., 2016).

En concreto, el uso de tecnologías inmersivas como la Realidad Virtual o Aumentada, permiten evaluar a los alumnos en entornos educativos más generales (Pauli y Ferrell, 2020) y, además facilitan nuevas formas de presentar estímulos y recoger respuestas (O'Leary et al., 2018).

En el caso de las pruebas basadas en ordenadores, según Binkley et al. (2012), los ordenadores son necesarios para evaluar a las personas en entornos reales, como los que ofrecen los entornos simulados, necesarios para medir habilidades como la resolución de problemas, la alfabetización informativa y la colaboración. Bennett y Bejar (1998) propusieron un modelo en el que las evaluaciones por ordenador pueden considerarse como un sistema integrado compuesto por componentes interrelacionados. Kuo y Wu (2013) aplicaron y elaboraron ese modelo, explorando los patrones de las aplicaciones tecnológicas, centrándose en las categorías en las que las aplicaciones tecnológicas aportan ventajas y nuevas oportunidades. De hecho, utilizamos este marco para guiar nuestro análisis con el fin de identificar las posibles aplicaciones de la tecnología, en particular en el desarrollo de nuevos tipos de ítems que evalúen habilidades de orden superior.

Desde el punto de vista del registro de información en las pruebas de evaluación, Azevedo (2015) destaca como la TEA ha aumentado significativamente

las capacidades de recopilación de datos como, p. ej. registrando respuestas de los participantes, registros, capturas de pantalla, patrones de movimiento de los ojos, datos de sensores fisiológicos, etc. Este hecho ha favorecido el seguimiento y registro de RP (Duchowski y Duchowski, 2017; Van Gog y Scheiter, 2010), pudiendo ser utilizada para apoyar la práctica de validación y desarrollo de pruebas de evaluación (Ercikan y Pellegrino, 2017; Zumbo y Hubley, 2017).

La tendencia hacia nuevos paradigmas de evaluación se ve facilitada por el uso de ordenadores. Por ejemplo, las evaluaciones internacionales a gran escala tienen una influencia cada vez mayor en un mundo educativo globalizado (Shute et al., 2016), como son el Programa para la Evaluación Internacional de Alumnos (PISA[37]), el Programa para la Evaluación Internacional de las Competencias de los Adultos (PIAAC[38]), el Estudio Internacional sobre Tendencias en Matemáticas y Ciencias (TIMSS[39]), el Estudio Internacional sobre el Progreso de la Lectura (PIRLS[40]) y el Estudio Internacional sobre Conocimientos de Informática e Información (ICILS[41]). Han adoptado un enfoque TEA con pruebas realizadas por ordenador, utilizando nuevos formatos de ítems como las tareas interactivas. Siiman et al. (2016) revisaron las tareas interactivas de varias evaluaciones educativas a gran escala para comprender mejor las ventajas de estos ítems de evaluación y cómo pueden guiar el desarrollo de ítems por ordenador para evaluar la CD de los estudiantes. Los autores encontraron que sólo 10 de las 21 CD de DigComp (entre ellas la netiqueta) pueden estar representadas por tareas interactivas y, por lo tanto, requieren desarrollar nuevos ítems interactivos para poder ser evaluadas.

Sin embargo, O'Leary et al. (2018) indican que hay que tener presente que a pesar de que el uso de simulaciones interactivas amplía el abanico de posibles comportamientos a evaluar, el desarrollo de procedimientos de puntuación fiables y válidos que tengan en cuenta todas estas posibilidades es complejo.

En la tabla 2.13 puede verse las principales referencias identificadas y analizadas en TEA.

Característica	Referencia
Ofrece inmensas oportunidades para mejorar la experiencia de los participantes y desarrollar modos de evaluación más pertinentes y ajustados a las necesidades actuales, de especial interés en la evaluación y el modelado psicométrico de constructos complejos	(Binkley et al., 2012; Cho et al., 2019; Debuse y Lawley, 2016; Drasgow, 2016; Graesser et al., 2017; Greiff et al., 2013; Greiff et al., 2015; Grover y Pea, 2013; Kuo y Wu, 2013; Mayrath et al., 2012;

[37] https://www.oecd.org/pisa/

[38] https://www.oecd.org/skills/evaluaciones-de-competencias/evaluaciondecompetenciasdeadultospiaac.htm

[39] https://nces.ed.gov/timss/

[40] https://nces.ed.gov/surveys/pirls/

[41] https://www.iea.nl/studies/iea/icils

como la CD. Además, proporciona formatos de ítems innovadores y más cercanos a la práctica real, así como la posibilidad de obtener información sobre el comportamiento y el rendimiento de los participantes durante las pruebas de evaluación.	OCDE, 2016; ; Osborne et al., 2013; O'Leary et al., 2018; Pauli y Ferrell, 2020; Scherer, 2015; Scherer et al., 2017; Shute y Rahimi, 2017; Stödberg, 2012; Timmis et al., 2016; Van Voorhis y Paris, 2019; von Davier et al., 2017; Zenisky y Luecht, 2016)
Posibilita la evaluación en entornos seguros	(Scherer et al., 2014)
Uso de los datos registrados para mejorar la interpretación de las puntuaciones o para contribuir a la evaluación de la validez	(Duckworth y Yeager, 2015; Oranje et al., 2017; Shute et al., 2016)
Fuerte conexión entre el rendimiento y el compromiso de quienes se examinan y el diseño de las tareas y evaluaciones, fundamental para que desencadenen los conocimientos y habilidades esperados, especialmente cuando se evalúan constructos cognitivos complejos.	(Nguyen, Rienties et al., 2017; O'Leary et al., 2018; Papamitsiou y Economides, 2016; Rienties y Toetenel, 2016; Van Voorhis y Paris, 2019)
Es necesario profundizar en las estrategias de diseño de la evaluación para comprender el contexto en el que tienen lugar.	(Engelhardt et al., 2017; Papamitsiou y Economides, 2016)
En cuanto a la calidad de las herramientas de evaluación, las evidencias basadas en RP son escasas o inexistentes	(Saltos-Rivas et al., 2021; Siddiq et al., 2016; Zhao et al., 2021)
ET amplía la información de los RP proporcionando observaciones de quienes se evaluan como el comportamiento estratégico mostrado, las secuencias llevadas a cabo y los patrones de respuesta, pero sigue sin aplicarse en el área de la evaluación de la CD.	(Azevedo, 2015; Bartolomé et al., 2020; Drasgow, 2016; Goldhammer et al., 2016; Greiff et al., 2015; Greiff et al., 2016; Oranje et al., 2017; Osborne et al.,2013; Sweeney et al., 2017; Timmis et al., 2016; Veldkamp et al., 2017)
La TEA favorecido el seguimiento y registro de RP, pudiendo ser utilizada para apoyar la práctica de validación y desarrollo de pruebas de evaluación.	(Duchowski y Duchowski, 2017; Ercikan y Pellegrino, 2017; Van Gog y Scheiter, 2010; Zumbo y Hubley, 2017)
Posibilitan la automatización de la puntuación, el diseño y desarrollo de las pruebas, así como su ensamblado.	(Engelhardt et al., 2017; Gierl et al., 2014; Kuo y Wu, 2013; Nguyen et al., 2017; Paiva et al., 2022; Taub et al.,2017; Veldkamp, 2015; Veldkamp et al., 2017; Vista et al., 2017; Zechner et al., 2017)
Facilita la provisión de comentarios constructivos, adecuados y fáciles de entender.	(Carless et al., 2011; Debuse y Lawley, 2016; Eggen et al., 2011; Helfaya y O'Neill, 2018; Helfaya, 2019; Makransky et al., 2020; Van der Kleij et al., 2012)
TEA puede considerarse como un sistema integrado compuesto por componentes interrelacionados, en el que hay que identificar las posibles aplicaciones de la tecnología con un fin en particular.	(Bennett y Bejar, 1998; Kuo y Wu, 2013)

Tabla 2.13. Principales referencias identificadas y analizadas en TEA.

Captura de información adicional mediante rastreadores oculares

ET es un procedimiento en el que un dispositivo recoge el punto donde los ojos fijan la mirada o los movimientos de los ojos en relación con la cabeza del usuario y los objetos necesarios para realizar las tareas (estímulo en adelante) (Duchowski y Duchowski, 2017). El interés por el seguimiento de la mirada ha aumentado especialmente en las últimas décadas debido a los rápidos avances en la tecnología ET, p. ej. proporcionando dispositivos no intrusivos con mayores tasas de registro por segundo.

La fuerte relación entre el comportamiento cognitivo y los patrones de movimiento ocular ha impulsado el uso de la tecnología ET en diferentes campos en los últimos años como, p. ej. en el estudio de la lectura, la búsqueda visual, la percepción de escenas naturales y los estudios de usabilidad (Blascheck et al., 2017; Rayner, 1998).

En diferentes revisiones sistemáticas sobre los procesos cognitivos en entornos de aprendizaje multimedia, se mostró cómo la tecnología ET puede ayudar a comprender el rendimiento cognitivo de los alumnos en un entorno de aprendizaje multimedia (Alemdag y Cagiltay, 2018; Mutlu-Bayraktar et al., 2019), en entornos de aprendizaje multimedia animados/simulados (Coskun y Cagiltay, 2022), y en entornos de realidad virtual o mixta (Rappa et al., 2019).

Posteriormente, la tecnología ET se incorporó a la evaluación, demostrando la viabilidad y la validez en el uso de las métricas de ET y el seguimiento de la mirada para distinguir entre individuos de diferentes niveles de habilidad (Just y Carpenter, 2018; Halszka et al., 2017), para analizar la relación entre las métricas de ET en los entornos de aprendizaje multimedia animados/simulados (Coskun y Cagiltay, 2022), pero también para el diseño y la validación de pruebas (Nisiforou y Laghos, 2013; Oranje et al., 2017; Tai et al., 2006; Yaneva et al., 2021).

Según la revisión sistemática de Tien et al. (2014), la superposición de las ubicaciones de la mirada y las fijaciones se estudiaron ampliamente como indicadores indirectos de las áreas de importancia percibida por el participante. Además, se indicó que el uso de ejemplos de modelado de movimientos oculares (MO), es decir, la mirada de otra persona, es útil en diferentes situaciones como, p. ej. guiar la atención y promover el rendimiento de los radiólogos novatos en el diagnóstico médico (Litchfield et al., 2010), mejorar la interpretación de los estudiantes de imágenes médicas basadas en el modelado de MO de un profesor (Seppänen y Gegenfurtner, 2012), guiar la atención visual de los estudiantes de medicina en el razonamiento clínico mientras realizan una búsqueda visual de síntomas (Jarodzka et al., 2012), mejorar el seguimiento del rendimiento de los estudiantes (Kok et al., 2022), apoyar el procesamiento integrador de los estudiantes de la información verbal y gráfica durante la lectura de un texto ilustrado (Mason et al., 2017), o para fomentar la comprensión de los comportamientos de resolución de los estudiantes en las evaluaciones de ciencias (Tien et al., 2014). Hasta donde sabemos, ninguna investigación anterior ha aplicado el uso de ejemplos de modelado de MO en el campo de la evaluación de la CD.

Sin embargo, también hay estudios contradictorios que sugieren la necesidad de una mayor investigación, ya que la búsqueda visual depende del dominio, la tarea y el nivel de experiencia (van der Gijp et al., 2017). Por ejemplo, Eder et al. (2022) descubrieron que los estudiantes que interpretaban radiografías panorámicas dentales no mejoraban su rendimiento tras utilizar ejemplos de modelado de MO como método de apoyo para la interpretación de imágenes médicas.

Además, el uso de la tecnología ET mientras los participantes realizan las pruebas de evaluación, puede proporcionar información adicional sobre su rendimiento. Los MO son esenciales para los procesos cognitivos, ya que dirigen la atención visual a las partes específicas de algún estímulo, que son procesadas por el cerebro (Just y Carpenter, 1976) y reflejan los procesos cognitivos actuales en tareas de lectura y procesamiento de información (Rayner, 1998).

También se han realizado estudios para investigar cómo las personas interactúan con las herramientas digitales, crucial para evaluar la CD de los participantes (Ala-Mutka, 2011; Ferrari y Punie, 2013). Varios estudios señalaron que los patrones de fijación son relevantes en el análisis y reconocimiento de la información, y cómo interactuamos con la misma (Ashraf et al., 2018; Lund, 2016; Tsai et al., 2012).

Respecto a las estrategias cognitivas, se han estudiado en profundidad analizando los lugares donde miran los participantes y el orden en que lo hacen (Brunyé et al., 2019). Hu et al. (2017) hicieron un seguimiento de MO de los estudiantes que resolvían problemas interactivos y analíticos de la evaluación PISA y concluyeron que los RP de los estudiantes diferían entre los distintos tipos de problemas y entre las distintas etapas del RP. En un estudio reciente, Lewandowski y Kammerer (2021) comprobaron que el uso del ET es útil para comprender el comportamiento de los usuarios en los motores de búsqueda. Además, numerosos estudios han investigado qué elementos visuales de las páginas web son los más fijados y qué caminos se siguen, p. ej. para crear una ruta de exploración común en términos de elementos visuales (Eraslan et al., 2014).

En cuanto a las métricas de ET, dado que la mayor parte de la adquisición de información y el procesamiento cognitivo se produce durante las fijaciones, la mayoría de las métricas cuantitativas utilizadas en los estudios se basan en las fijaciones (Duchowski y Duchowski, 2017; Privitera y Stark, 2000). Un procedimiento común cuando se quiere medir la dificultad de procesamiento de la web es analizar los lugares donde se producen las fijaciones más largas y las duraciones de las fijaciones (Goldberg et al., 2002), utilizando las duraciones de las fijaciones como un proxy para medir la carga cognitiva (Just y Carpenter, 1980). Los investigadores suelen definir Áreas de interés (AOI) para estudiar los MO dentro de estas áreas específicas, y Tiempo de interés (TOI) para seleccionar intervalos de tiempo específicos de los datos de registro.

Además, el uso de métodos de análisis cualitativos y exploratorios basados en técnicas de visualización ha aumentado en los últimos años (Blascheck et al., 2017), lo que permite a los investigadores examinar otros aspectos de los datos. Más

concretamente, se pueden utilizar diferentes formas de representaciones de datos para describir la estrategia de los participantes: métodos de edición de cadenas y secuencias (Blascheck et al., 2017; Glady et al., 2013; Kucharský et al., 2020; Kübler et al., 2017), clasificación de estadísticas en bruto (Boisvert y Bruce, 2016; Kanan et al., 2014), y los modelos de Markov (Coutrot et al., 2018; Groner et al., 1984).

La identificación de grupos latentes, como participantes exitosos y no exitosos, puede llevarse a cabo basándose en patrones de ET que pueden complementar los análisis más convencionales del comportamiento de respuesta (Li et al., 2017; Steingroever et al., 2019). Además, se han desarrollado diferentes algoritmos para analizar las trayectorias de exploración individuales para descubrir las trayectorias más seguidas en términos de elementos visuales (Tai et al., 2006).

En la tabla 2.14 puede verse las principales aplicaciones y características de ET identificadas y analizadas.

Aplicación / Característica	Referencia
Estudios de la lectura, búsqueda visual, percepción de escenas naturales y usabilidad.	(Blascheck et al., 2017; Just y Carpenter, 1976; Rayner, 1998)
Análisis del rendimiento cognitivo en entornos de aprendizaje multimedia, y de realidad virtual o mixta.	(Alemdag y Cagiltay, 2018; Coskun y Cagiltay, 2022; Mutlu-Bayraktar et al., 2019; Rappa et al., 2019)
Distinguir entre individuos de diferentes niveles de habilidad.	(Just y Carpenter, 2018; Halszka et al., 2017; Li et al., 2017; Steingroever et al., 2019)
Análisis de estrategias cognitivas examinando los lugares observados y el orden seguido.	(Brunyé et al., 2019)
Diseño y la validación de pruebas.	(Nisiforou y Laghos, 2013; Oranje et al., 2017; Tai et al., 2006; Yaneva et al., 2021)
Uso de ejemplos de modelado de MO útil, p. ej. para guiar la atención y promover el rendimiento de novatos.	(Jarodzka et al., 2012; Kok et al., 2022; Litchfield et al., 2010; Mason et al., 2017; Seppänen y Gegenfurtner, 2012; Tien et al., 2014)
La búsqueda visual depende del dominio, la tarea y el nivel de experiencia.	(Eder et al., 2022; van der Gijp et al., 2017)
Análisis de interacciones con herramientas digitales, crucial para evaluar la CD de los participantes.	(Ala-Mutka, 2011; Ashraf et al., 2018; Ferrari y Punie, 2013; Lund, 2016; Tsai et al., 2012)
Los RP de los participantes difieren entre los distintos tipos de	(Hu et al., 2017)

problemas y entre las distintas etapas del RP.	
Uso del ET para modelar el comportamiento de los usuarios en los motores de búsqueda.	(Lewandowski y Kammerer, 2021)
Análisis de los elementos visuales de las páginas web más fijados y caminos seguidos para crear una ruta de exploración común.	(Eraslan et al., 2014; Tai et al., 2006)
La mayoría de las métricas cuantitativas se basan en fijaciones, ya que es en las fijaciones donde se produce la mayor parte de la adquisición de información y el procesamiento cognitivo.	(Duchowski y Duchowski, 2017; Privitera y Stark, 2000)
Duraciones de las fijaciones como un proxy para medir la carga cognitiva y la dificultad.	(Goldberg et al., 2002; Just y Carpenter, 1980)
Uso de métodos de análisis cualitativos y exploratorios basados en técnicas de visualización para describir la estrategia de los participantes: métodos de edición de cadenas y secuencias, clasificación de estadísticas en bruto, y los modelos de Markov.	(Blascheck et al., 2017; Boisvert y Bruce, 2016; Coutrot et al., 2018; Glady et al., 2013; Groner et al., 1984; Kanan et al., 2014; Kucharský et al., 2020; Kübler et al., 2017)

Tabla 2.14. Principales aplicaciones y características de ET identificadas y analizadas.

2.3.2. Analíticas de evaluación (AA)

Las analíticas del aprendizaje (LA) utilizan grandes cantidades de información sobre las interacciones realizadas por los alumnos en los entornos de aprendizaje digitales, para comprender y mejorar el aprendizaje. Aunque la evaluación es una dimensión central de las LA, hasta ahora pocos autores han examinado los vínculos entre las LA y la evaluación. Primero realizamos una breve introducción al campo de las AA, y a continuación describimos como hacemos uso de la información proveniente de las pruebas de evaluación, en concreto de los RP de los participantes al responder los ítems, para ser usada como evidencias de validación en el proceso de desarrollo de las pruebas de evaluación.

Orígenes, conceptos y definiciones

Desde su concepción, las LA se han preocupado por resolver los problemas asociados al crecimiento de la disponibilidad, la cantidad, la velocidad y el tipo de datos en los entornos de aprendizaje. Las LA se desarrollaron rápidamente en el campo de la mejora del aprendizaje a través de la tecnología con el objetivo de "*medir, recoger, analizar y reportar datos sobre los estudiantes y sus contextos, con el fin de comprender y optimizar el aprendizaje y el entorno en el que se produce*" (Siemens, 2013).

Hoy en día, las LA se encuentran en un punto en el que, gracias a los esfuerzos conjuntos entre la investigación y la industria, han conseguido hacer manejables los problemas técnicos, el uso masivo de cursos en línea masivos y abiertos (MOOC) y

la necesidad del cambio a la instrucción a distancia causado por la pandemia, han demostrado que la adquisición y el análisis a gran escala de los datos de las trazas de las interacciones de los alumnos con el contenido de aprendizaje y entre los propios alumnos, son al menos posibles (Reich, 2022). Sin embargo, existe un desafío constante en los sistemas educativos que tienen que afrontar las continuas necesidades de adaptación causadas por los cambios de comportamiento, actitudes y procesos que surgen en respuesta a los rápidos cambios tecnológicos que se producen (Heifetz y Laurie, 1997).

Durante la última década, el campo de las LA ha crecido tanto en número de autores y estudios, así como en la amplitud de sus áreas de interés. Desde 2011, la comunidad de LA cuenta con el apoyo de la Sociedad para la Investigación de LA (SOLAR)[42], con más de 1000 miembros. Entre sus funciones, cabe mencionar la organización de la conferencia (LAK)[43] así como la publicación del Journal de LA[44].

Las LA ofrecen muchas oportunidades para entender mejor y mejorar el aprendizaje de los alumnos, así como los entornos en los que éste tiene lugar, gracias al análisis de las interacciones de los usuarios en los entornos de aprendizaje (Lang et al., 2022). Respecto al abanico de oportunidades que ofrece las LA, abarcan desde las bondades que aporta el uso de la tecnología en general, pasando por la eficiencia y la reducción del trabajo (Foster y Siddle, 2020; Littlejohn, 2021), hasta el potencial de mejora de la desigualdad y el acceso (Nguyen et al., 2020; Ochoa, 2019; Uttamchandani y Quick, 2022).

Las LA se han aplicado en una gran variedad de contextos y situaciones, pero las principales han sido: la predicción y descripción de los resultados y procesos de aprendizaje (p. ej. Baker et al., 2015; Gardner y Brooks, 2018), la predicción del éxito de los estudiantes y la identificación de estudiantes en riesgo (p. ej. Jovanović et al., 2021), el perfilado de los estudiantes identificando sus estrategias (p. ej. Barthakur et al., 2021; Jovanović et al., 2017; Matcha et al., 2020), los marcos de ética, protección de la privacidad y adopción (p. ej. Tsai et al., 2018), la comprensión de los estados afectivos (p. ej. D'Mello, 2017), la determinación del papel de las redes sociales en el aprendizaje (p. ej. Poquet y Jovanovic, 2020), las recomendaciones de aprendizaje adaptativo y la retroalimentación personalizada (p. ej. McNamara y Nulsen, 2012; Pardo, 2018), y el análisis de la práctica docente (p. ej. Martínez-Maldonado et al., 2022).

Por otra parte, las LA también tienen una serie de puntos a vigilar, por citar algunos, la privacidad (Ifenthaler y Schumacher, 2016; Pardo y Siemens, 2014), la ética (Ferguson et al., 2016; Prinsloo y Slade, 2017), o la propiedad de los datos (Kitto et al., 2015; Siemens, 2013). Además, ha incrementado la consciencia de que los datos no pueden separarse de como la tecnología facilita su recopilación, y que la interacción entre los humanos y la tecnología necesita ser explorada en profundidad (Siemens, 2012). A su vez, tal y como Wise et al. (2021) mencionan,

[42] https://solaresearch.org/

[43] https://www.solaresearch.org/events/lak/

[44] https://learning-analytics.info/index.php/JLA

también se está empezando a prestar atención a los medios para capturar tipos de datos mejores y más útiles, considerando medios no utilizados tradicionalmente.

Además, existen una serie de campos muy relacionados con las LA, como la minería de datos educativos, la inteligencia artificial en la educación, las ciencias del aprendizaje, el aprendizaje colaborativo mediado por ordenador o la ingeniería de datos educativos y aprendizaje. Las delimitaciones de cada campo son difusas, en ocasiones llegando a enmarcarse el trabajo de los autores en varios campos a la vez. A diferencia de la minería de datos educativos, que se centra en el desarrollo de métodos para explorar los tipos de datos generados en los entornos educativos, las LA se focalizan más en la creación de significado y la acción (Baker et al., 2016; Siemens y Baker, 2012). Incluso, Dormetzil et al. (2019) sostuvieron que la minería de datos educativos es un subcampo de las LA. Más aún, varios autores como Siemens (2013) o Baek y Dolek (2021) manifestaron que un rasgo definitorio de LA es su enfoque expansivo a la metodología.

A diferencia de los métodos de evaluación tradicionales, que dejan poco rastro, la TEA pueden generar grandes cantidades de datos. Además, los incrementos en la potencia de cálculo, una rápida mejora de la ciencia de los datos y los métodos de investigación sobre grandes conjuntos de datos, contribuyen a la evaluación y a la analítica del aprendizaje.

Los datos generados en las evaluaciones rara vez se incluyen como parte de los conjuntos de datos disponibles para las LA, a pesar de que autores como Knight et al. (2013) y Milligan (2020) manifestaron que las LA son intrínsecamente una forma de evaluación. Tal y como Gašević et al. (2022) indican, las débiles conexiones de hoy en día entre las LA y la evaluación educativa, puede ser debido a que los métodos de LA existentes no cumplen los criterios utilizados en psicometría en lo que se refiere a formas de validez en la evaluación (Kane, 2013; Messick, 1995). Aun así, las LA tienen un gran potencial para ofrecer beneficios al campo de la evaluación (Milligan, 2020). La mejora de los servicios de evaluación forma parte de los objetivos principales de la investigación en LA (Chatti et al., 2012; Papamitsiou y Economides, 2014). Las AA pueden ser utilizadas con los datos obtenidos en los procesos de evaluación, tales como, las clasificaciones de las notas, los resultados de la progresión, los resultados de las evaluaciones o los logros obtenidos.

Sin embargo, es necesario abordar algunos desafíos críticos como, p. ej. es necesario saber cómo las LA pueden utilizarse para obtener información sobre la evaluación durante el proceso formativo, o cómo se puede mejorar la validez y la fiabilidad en la recogida y el análisis de los datos.

Las LA proveen una serie enfoques que pueden mejorar las prácticas de evaluación haciendo uso de datos multimodales recogidos en entornos físicos de aprendizaje como, p. ej. comportamiento de los clics realizados y respuestas a preguntas. Las analíticas del aprendizaje multimodal (MLA) son un subcampo de las LA que reconoce que el aprendizaje es un fenómeno multimodal que se produce a través de múltiples espacios físicos y digitales, y requiere considerar múltiples fuentes de datos para analizar el aprendizaje como un proceso complejo como, p. ej. movimientos del ratón, ET, o biomarcadores fisiológicos (Azevedo y Gašević,

2019; Sharma y Giannakos, 2020). Por lo tanto, si queremos reforzar los vínculos entre la LA y la evaluación, será necesario centrarse en enfoques que puedan hacer uso de datos multimodales para abordar cuestiones de validez y fiabilidad en la evaluación (Fan et al., 2022).

Conviene señalar que la parte correspondiente a la mejora constante del proceso de evaluación a partir de la información registrada durante la realización de las pruebas suele estar poco detallada y trabajada (Siddiq et al., 2016). Esta información puede ser útil en pruebas de evaluación que requieran tipos de ítems más complejos, para poder examinar si los ítems fueron bien diseñados y realmente activaron los comportamientos previstos (Hunt y Jordan, 2016; Siddiq et al., 2016). Además, permite profundizar en cómo los estudiantes llegaron a sus respuestas, especialmente cuando se pretende que los examinados exhiban comportamientos similares a los del entorno laboral (Rayón et al., 2014).

Autores como Ifenthaler y Greiff (2021, 2022) ya han trabajado en este contexto de unión entre LA y AA, y estudiaron el uso de datos registrados durante la evaluación haciendo uso de técnicas de análisis de datos.

Por otro lado, las diferencias en el comportamiento de los usuarios durante la evaluación tienen un profundo impacto en su rendimiento y su nivel de logro. Para profundizar en este punto, Papamitsiou y Economides (2017) exploraron la dimensión temporal para modelar el comportamiento de los usuarios, y utilizaron los tiempos de respuesta para estudiar el comportamiento de la gestión del tiempo analizando: para ver cómo progresaban en cada ítem en comparación con el resto de los participantes, para identificar la dificultad real de los ítems y adaptarse mejor al nivel de la prueba, y para detectar patrones de comportamiento no deseados.

Vínculos entre las LA y AA

Gašević et al. (2022) realizaron una revisión de los artículos que abordan diferentes temas de evaluación que utilizan datos de rastreo para analizar las prácticas de evaluación existentes o proponer y validar nuevas formas de evaluación. Los autores agruparon los vínculos entre LA y AA en 3 categorías: (1) analíticas para la evaluación, (2) analíticas de la evaluación y (3) validez de la evaluación.

La primera categoría corresponde a las analíticas para la evaluación o, dicho de otra forma, enfoques de LA cómo formas de evaluación. Peters et al. (2021) propusieron crear y validar un nuevo enfoque para la evaluación de la inteligencia utilizando un videojuego popular. Los autores descubrieron que los datos de rastreo eran altamente predictivos del rendimiento en las pruebas de inteligencia en el juego y moderadamente predictivos del rendimiento en las pruebas convencionales. Rowe et al. (2021) usaron un videojuego para medir las prácticas de pensamiento computacional que siguen los estudiantes mientras juegan. Los autores desarrollaron un conjunto de clasificadores de aprendizaje automático entrenados con datos de rastreo del juego, los cuales detectaron automáticamente con buena precisión las prácticas de pensamiento computacional. Dowell y Poquet (2021)

combinaron análisis de redes sociales y análisis de la comunicación de grupo para la caracterización de los roles sociocognitivos que asumen los estudiantes durante las interacciones en línea en un MOOC. Barthakur et al. (2021) siguieron un enfoque basado en un análisis de clases latentes de los datos de rastreo de las interacciones de los estudiantes con los recursos disponibles en la plataforma MOOC, para evaluar las estrategias que siguen los estudiantes. Milligan y Griffin (2016) diseñaron un instrumento de evaluación de la capacidad de aprender en entornos, basado en las trazas registradas en los MOOC. Sun y Theussen (2022) propusieron un marco conceptual para evaluar las habilidades de negociación mediante la identificación de diferentes conjuntos de habilidades de negociación basada en el análisis de redes sociales. Los resultados mostraron que la mayoría de los grupos de estudiantes practicaron conjuntos de habilidades de negociación más complejos hacia el final del juego, y que la complejidad de los conjuntos de habilidades se relacionó positivamente con los resultados en el juego de simulación. Barthakur et al., (2022) desarrollaron una metodología y un sistema automatizado para la evaluación del dominio de las habilidades de liderazgo basado en la profundidad de la reflexión exhibida durante el proceso de aprendizaje, con el objetivo de identificar diferentes grupos de estudiantes. Akyar y Demirhan (2022) examinaron los estilos de negociación de candidatos a entrenadores deportivos utilizando una herramienta de evaluación basada en un juego, para mostrar las relaciones entre los distintos estilos. Rahimi et al. (2021) realizaron un análisis de los datos de comportamiento de los estudiantes, extraídos de los registros de un juego de física basado por ordenador, para confirmar que los estudiantes accedían a los apoyos de aprendizaje diseñados para facilitar la adquisición de conocimientos de contenido y no abusaban de los vídeos de soluciones.

La segunda categoría corresponde a las analíticas de la evaluación o, dicho de otra forma, aplicaciones de LA para responder a preguntas sobre las prácticas de evaluación. Stadler et al. (2020) utilizaron los recogidos durante la realización de las pruebas de resolución de problemas complejos para probar que el comportamiento mostrado era un indicador eficaz de la capacidad evaluada. Nicolay et al. (2021) utilizaron datos recogidos durante la realización de evaluaciones de resolución de problemas complejos para mostrar que muchos participantes no fueron capaces de transferir conocimientos, especialmente para los ítems más complejos de la evaluación. Zhang et al. (2021) propusieron un enfoque analítico para modelar la interacción entre la adaptación frente a la adversidad y la habilidad en evaluaciones que permiten múltiples intentos, descubriendo que la adaptación afectaba al rendimiento y, por lo tanto, cuestionaba la validez de las evaluaciones sumativas. Misiejuk et al. (2021) propusieron un enfoque analítico del aprendizaje basado en redes epistémicas para mostrar que los estudiantes valoran la especificidad, la justificación y la construcción en la evaluación entre pares, pero la amabilidad es menos prioritaria. Misiejuk y Wasson (2021) mostraron el potencial de los datos de la evaluación retrospectiva para mostrar nuevas ideas sobre las percepciones de los estudiantes de lo que es una retroalimentación útil, sus reacciones a la retroalimentación recibida y sus consecuencias para la aplicación de

la retroalimentación. Israel-Fishelson et al. (2021) analizaron los datos de una prueba de creatividad estandarizada y los cruzaron con los registros de las actividades de los estudiantes en un entorno de aprendizaje basado en juegos. Los autores entre otros puntos mostraron como la creatividad contribuye al pensamiento computacional. Zhang et al. (2021) analizaron los datos de registro recogidos en un juego diseñado para entrenar la sub-competencia de las funciones ejecutivas de los cambios. Los autores diseñaron características específicas a nivel de juego y para cada nivel a partir de los datos registrados en los registros. Posteriormente, utilizaron estas características para construir modelos de predicción con la precisión y el tiempo de reacción de los estudiantes durante el juego para predecir su cambio de habilidad, así como para predecir las ganancias de aprendizaje. Alonso-Fernandez et al. (2022) combinaron diferentes técnicas (visualizaciones y modelos de minería de datos) para obtener información significativa y comprender mejor las acciones y los resultados de los estudiantes en los juegos serios. Raković et al. (2022) utilizaron datos de un curso de biología, para examinar cómo los juicios evaluativos de los estudiantes realizados después de un primer examen de unidad predijeron los cambios en los comportamientos de aprendizaje y su posterior rendimiento en un examen posterior. Shaw (2022) realizó análisis de regresión basados en las puntuaciones obtenidas en una tarea de construcción de Minecraft mediante la técnica de evaluación consensuada, y revelaron que la creatividad de Minecraft se predecía mediante el pensamiento divergente, las puntuaciones del Test de Aptitud Escolar y la apertura a la experiencia. Liu y Israel (2022) aplicaron el modelo continuo oculto de Markov y técnicas de minería de secuencias, para descubrir los procesos de resolución de problemas de los estudiantes. El objetivo fue utilizar esta información como ayuda para activar futuros soportes e intervenciones de apoyo al aprendizaje personalizado de los estudiantes en entornos de aprendizaje basados en juegos. Emerson et al., (2022) llevaron a cabo un enfoque de modelado predictivo de estudiantes utilizando las preguntas de conocimiento del contenido posteriores al juego para la predicción temprana del rendimiento individual de los estudiantes en entornos de aprendizaje basados en juegos. Los resultados mostraron que, al incorporar el conocimiento sobre las preguntas de evaluación, los modelos de predicción temprana mejoraban los modelos de predicción que sólo usan los datos de las trazas de los estudiantes en el juego.

En materia de resolución de problemas complejos (CPS), Weise et al. (2022) examinaron el comportamiento estratégico para identificar tempranamente los efectos producidos al explorar el sistema en diferentes fases y su relación con el rendimiento en la CPS y hasta qué punto mediaba la relación entre la inteligencia y el rendimiento en la CPS. Molnár et al. (2022) analizaron los datos de los registros de uso e identificaron diferencias en el comportamiento de los estudiantes a la hora de realizar los exámenes en cuanto a la eficacia de su estrategia de exploración, el tiempo en la tarea y el número de intentos. Alrababah et al. (2022) investigaron el papel de la exploración estratégica y de los diferentes comportamientos de resolución de problemas en el éxito de CPS, utilizando datos de uso para visualizar y cuantificar el comportamiento de resolución de problemas de los estudiantes.

Detectaron notables diferencias en el comportamiento de los estudiantes en cuanto a la eficacia de su estrategia de exploración, identificando cuatro clases latentes basadas en el comportamiento de la estrategia de exploración de los estudiantes. Jordan (2014) analizó las respuestas de los estudiantes a las preguntas interactivas, para identificar conceptos erróneos e identificar los patrones característicos de compromiso con las tareas y la retroalimentación proporcionada.

La tercera categoría corresponde a la validez de la evaluación, esto es, la conceptualización y enfoques prácticos para garantizar la validez de la evaluación en LA. Winne (2020) analizó los factores que pueden confundir la validez del aprendizaje autorregulado, como la voluntad del estudiante mientras estudia y el contraste entre los eventos dinámicos en el aprendizaje frente a las medidas de evaluación estáticas. Shute y Rahimi (2021) analizaron la validez de una evaluación encubierta de la creatividad en un videojuego de física. Confirmaron que tiene una buena validez externa y que la creatividad estimada a través de la evaluación encubierta es un buen predictor del rendimiento en el juego. Liu et al. (2021) utilizaron un análisis factorial confirmatorio basado en características textuales extraídas de datos que utilizan marcos lingüísticos conocidos para validar un modelo de evaluación formativa de la reflexión escrita. Otros autores también han aplicado técnicas psicométricas o estadísticas para validar sus enfoques de evaluación basadas en el uso de datos de rastreo (Barthakur et al., 2021; Dowell y Poquet, 2021; Peters et al., 2021). Gane et al., (2021) utilizaron modelos cognitivos de pensamiento computacional para diseñar las evaluaciones y obtener evidencias para apoyar un argumento de validez y demostrar que se podían realizar inferencias válidas sobre las competencias en pensamiento computacional de los estudiantes. Para ello, utilizaron las trayectorias de aprendizaje y el diseño centrado en la evidencia para desarrollar las evaluaciones. El enfoque de diseño y el análisis de los datos, basados en múltiples enfoques psicométricos, sugirieron que las evaluaciones fueron adecuadas para evaluar el conocimiento, las habilidades y las capacidades de los estudiantes en materia de pensamiento computacional a través de múltiples trayectorias de aprendizaje. Dumas et al. (2022) formularon un coeficiente de relación relativamente sencillo de calcular que pretende captar hasta qué punto las puntuaciones de un determinado test pueden predecir un criterio libre de la influencia indebida de los datos demográficos de los estudiantes. Fan et al. (2022) examinaron la validez de los protocolos de evaluación basados en las trazas registradas durante el aprendizaje autorregulado, basándose en tres alineaciones entre los datos de las trazas y los datos del pensamiento en voz alta, para mejorar la validez de la evaluación.

De las 3 categorías posibles según Gašević et al. (2022), el objetivo de esta tesis queda enmarcado en la tercera categoría. En concreto buscamos validar la evaluación basándonos en el análisis de los RP de los participantes al responder los ítems de las pruebas de CD.

Otro aspecto a tener en cuenta es que los datos utilizados en las LA suelen ser un reflejo de los sesgos habituales presentes en la sociedad, y dado de que el uso de modelos basados en esos datos puede perpetuar los sesgos e incluso profundizar la desigualdad, se ha prestado poca atención de manera sistemática a la equidad de los subgrupos en las LA (Carter y Egliston, 2021; Gardner et al., 2019; O'Neil, 2016; Selwyn, 2020; Sha et al., 2021).

En la tabla 2.15 puede verse las 3 categorías de vínculos entre LA y AA de acuerdo con Gašević et al. (2022), y las investigaciones identificadas y analizadas para este estudio.

Categoría	Investigación
Analíticas para la evaluación (LA cómo formas de evaluación)	Caracterización de los roles y estilos (Akyar y Demirhan, 2022; Dowell y Poquet, 2021).
	Evaluación de estrategias y comportamientos (Barthakur et al., 2021; Peters et al., 2021; Rahimi et al., 2021; Rowe et al., 2021).
	Evaluación de habilidades (Barthakur et al., 2022; Milligan y Griffin, 2016; Sun y Theussen, 2022).
Analíticas de la evaluación (LA para responder a preguntas sobre las prácticas de evaluación)	Confirmación de prácticas evaluadas (Israel-Fishelson et al., 2021; Misiejuk y Wasson, 2021; Misiejuk et al., 2021; Stadler et al., 2020).
	Modelado de acciones (Emerson et al., 2022; Raković et al., 2022; Shaw, 2022; Zhang et al., 2021).
	Análisis de comportamientos (Alonso-Fernandez et al., 2022; Alrababah et al., 2022; Jordan, 2014; Liu y Israel, 2022; Molnár et al., 2022; Nicolay et al., 2021; Weise et al., 2022).
Validez de la evaluación (garantizar la validez de la evaluación)	Análisis y apoyo del argumento de validez (Barthakur et al., 2021; Dowell y Poquet, 2021; Dumas et al., 2022; Fan et al., 2022; Gane et al., 2021; Liu et al., 2021; Peters et al., 2021; Shute y Rahimi, 2021; Winne, 2020).

Tabla 2.15. Categorías de vínculos entre LA y AA (Gašević et al., 2022), e investigaciones identificadas y analizadas

Validación, evidencias basadas en los RP

La validación puede definirse como el proceso mediante el cual se acumulan evidencias para apoyar interpretaciones y usos específicos de las puntuaciones de las pruebas. En consecuencia, la validez ha sido objeto de estudio por una parte importante de autores, tanto en estudios centrados explícitamente en la validez, como en otros estudios que abordan cuestiones de validez en la evaluación para el

aprendizaje y la evaluación del aprendizaje. Sin embargo, el campo sigue necesitando un marco teórico claro que guíe la consideración de la validez en las LA, que reconozca las propiedades específicas de los datos in situ que utiliza la analítica del aprendizaje (Gašević et al., 2015; Wise y Shaffer, 2015). Al mismo tiempo, gracias al enorme potencial de la TEA para obtener información sobre el comportamiento y el rendimiento de los examinados durante las pruebas de evaluación, recogiendo diferentes tipos de datos, existe una gran oportunidad para aprovechar estas fuentes de información y utilizarlas para examinar la validez de las evaluaciones. Por ejemplo, Zhang et al. (2021) descubrieron que la resiliencia afectaba al rendimiento en evaluaciones que permiten múltiples intentos, cuestionando la validez de las evaluaciones sumativas.

En el marco proporcionado por la *Asociación Americana de Investigación Educativa* (AERA) (AERA, APA, y NCME, 2014), se identificaron cinco fuentes de evidencia: 1) contenido de la prueba; 2) RP; 3) estructura interna; 4) relaciones con otras variables; y 5) consecuencias del test. Además, las evidencias basadas en el contenido de las pruebas se complementan con las pruebas basadas en los RP.

Según Ercikan y Pellegrino (2017), "*los RP se refieren a los enfoques y comportamientos mostrados por los participantes cuando interpretan las situaciones de evaluación y formulan y generan soluciones, tal y como se revela a través de verbalizaciones, movimientos oculares, tiempos de respuesta o clics en el ordenador. Estos datos del RP pueden proporcionar información sobre el grado en que los ítems y las tareas involucran a los examinados de la manera prevista*". Hubley y Zumbo (2017) definieron los RP como "*los mecanismos que subyacen a lo que las personas hacen, piensan o sienten cuando interactúan con el ítem o la tarea y responden a ellos, y que son responsables de generar la variación de puntuación observada*".

Las pruebas de validez basadas en los RP incluyen aspectos relacionados con la forma en que los examinados interactúan con una prueba. En concreto, se busca confirmar que los examinados activan los procesos cognitivos previstos cuando responden a los ítems. Los RP pueden proporcionar información útil e inferencias de apoyo sobre el rendimiento de los examinados, que pueden ayudar a los diseñadores de pruebas a reducir la brecha entre las respuestas observadas y los constructos previstos. Sin embargo, esta contribución a la investigación de la validación no es nueva, p. ej. Cronbach (1949).

La recopilación de estas evidencias implica la investigación de los procesos cognitivos utilizados durante la prueba. En este ámbito se suelen presentar diferentes tipos de pruebas: realización de entrevistas de pensamiento en voz alta mientras los examinados responden a los ítems, realización de entrevistas cognitivas, análisis de las pulsaciones del teclado y del ratón, métodos de seguimiento basados en ET o análisis del tiempo de respuesta (AERA, APA y NCME, 2014).

Según Chronbach (1980), "*el trabajo de la validación no es apoyar una interpretación, sino averiguar qué puede ser erróneo en ella. Una proposición sólo merece cierto grado de confianza cuando ha sobrevivido a intentos serios de*

falsearla. (p. 103)". La sugerencia de Chronbach para las pruebas de validez es relevante en el contexto de los estudios de ET, ya que no aportan pruebas directas de que el proceso cognitivo implicado en los RP de los ítems sea el mismo utilizado en la tarea del mundo real examinada, pero pueden contribuir a demostrar que no es equivalente (Kane y Mislevy, 2017). Por ejemplo, si los datos del ET demuestran que los examinados resolvieron las tareas presentadas en las simulaciones sin examinar detenidamente el enunciado y las instrucciones, se desvirtuaría que las simulaciones requerían el mismo proceso cognitivo que se requiere en las tareas del mundo real. Además, los datos del ET podrían mostrar cómo examinados con diferentes niveles de competencia podrían mostrar diferentes RP. Es decir, los datos del ET no proporcionan una confirmación directa del proceso cognitivo, pero las pruebas proporcionadas pueden utilizarse para obtener inferencias útiles sobre el proceso (Fitts et al., 1950).

Las evidencias de los RP son particularmente importantes para las pruebas destinadas a evaluar constructos complejos como la CD, o para el diseño de nuevos formatos de evaluación (Ercikan y Pellegrino, 2017). Es posible que los participantes en las pruebas de evaluación sean capaces de resolver los ítems utilizando caminos o estrategias que no requieren el uso de las habilidades cognitivas previstas cuando se diseñaron los ítems. Por ejemplo, Yaneva et al. (2021) utilizaron datos de ET para evaluar cómo las distribuciones de las opciones en las preguntas de opción múltiple influían en cómo los participantes respondían a las preguntas, evaluando una interpretación alternativa de las puntuaciones de la prueba.

Las evidencias de los RP han sido ampliamente investigadas para analizar la adecuación entre la respuesta mostrada por los participantes y el constructo que debe evaluar la prueba (Baxter y Glaser, 1998; De Boeck y Jeon, 2019; Ercikan et al., 2015; Schnipke y Scrams, 2002), y para informar de las comparaciones de grupos, explicando las diferencias en el significado de las puntuaciones de las pruebas entre los diferentes grupos de participantes (Ercikan et al., 2020; Ercikan y Pellegrino, 2017).

Más aún, los datos de RP a pequeña escala han ganado cada vez más atención en las pruebas de capacidad cognitiva debido a la digitalización de la evaluación, aumentando las capacidades de recopilación de datos y permitiendo el seguimiento y registro de los RP (Duchowski y Duchowski, 2017; Van Gog y Scheiter, 2010). En las evaluaciones a gran escala, se ha utilizado la información proveniente de los RP para apoyar la práctica de validación y el desarrollo de pruebas (Ercikan y Pellegrino, 2017; Zumbo y Hubley, 2017), incluyendo el análisis de los distintos registros generados al llevar a cabo las pruebas mediante un ordenador, la captura de la pantalla, la expresión facial, los patrones de movimiento de los ojos, el comportamiento grabado en vídeo y los datos de los sensores fisiológicos (Azevedo, 2015).

Varios autores como Greiff et al. (2016), Kroehne y Goldhammer (2018) o Tempelaar et al. (2015), han puesto de manifiesto la necesidad de que los investigadores incluyan el análisis del tiempo de respuesta. El análisis del tiempo de respuesta puede ser muy valioso, y ya se ha utilizado para estudiar el compromiso y

la motivación de los participantes durante la realización de pruebas, específicamente en ciertos campos, p. ej. en la detección de la adivinación rápida (Van Der Linden, 2009). Las comparaciones entre grupos suelen centrarse en las respuestas de los participantes, ignorando las diferencias en el proceso de respuesta que puedan tener los grupos. Sin embargo, considerar el comportamiento basado principalmente en el tiempo de respuesta no explica las razones por las que emplearon una determinada cantidad de tiempo, es decir, sería recomendable complementar esta información con otros datos de proceso (Li et al., 2017).

De acuerdo con Oranje et al. (2017), los datos de RP, como los datos registrados, el ET y el tiempo de respuesta, pueden contribuir a validar las inferencias realizadas entre las afirmaciones y el comportamiento observado de al menos cinco formas posibles: 1) inferencia, para generar y probar inferencias sobre el constructo de interés; 2) diseño, para mejorar los diseños de tareas e ítems analizando los patrones de comportamientos mostrados por los participantes; 3) evidencia contextual, para ofrecer un contexto a las inferencias relacionadas con el constructo de interés; 4) indicadores de compromiso, para proporcionar un indicador del nivel de compromiso de los participantes con las pruebas; y 5) limpieza y filtrado de datos, para eliminar comportamientos y respuestas no válidos.

Lamentablemente, a pesar de la relevancia de examinar los RP de los participantes (Ercikan y Pellegrino, 2017; Zumbo y Hubley, 2017), los análisis basados en los RP apenas se han incorporado dentro de los estudios de investigación de validación (Newton, 2019; Papamitsiou y Economides, 2016) y más aún, no se comprenden adecuadamente (Zumbo y Chan, 2014). Hasta donde sabemos, ninguna investigación anterior ha investigado los RP en el campo de la validación de la evaluación de CD haciendo uso de datos de ET.

En la tabla 2.16 puede verse las 5 fuentes de evidencia de validez identificadas por AERA, APA, y NCME (2014), donde se puede apreciar la carencia de evidencias de RP en materia de evaluación de la CD e indicando para este caso, que alternativas pueden ser manejadas para solventar la carencia identificada.

Contenido de la prueba	✓	• Análisis de la adecuación del contenido de la prueba representa y su relevancia para la interpretación propuesta de los puntajes de la prueba. • Juicios expertos de la prueba y el constructo.
RP	✗	• Entrevistas de pensamiento en voz alta mientras los participantes realizan las pruebas. • Entrevistas cognitivas. • Análisis de las pulsaciones del teclado y del ratón. • Métodos de seguimiento basados en ET. • Análisis del tiempo de respuesta.

		• Análisis de registros. • Captura de la pantalla. • Captura de la expresión facial. • Comportamiento grabado en vídeo. • Datos de los sensores fisiológicos
Estructura interna	✓	• Depende de cómo se utilice la prueba, p.ej., para validar la unidimensionalidad los ítems deberán mostrar evidencias de homogeneidad. • Funcionamiento diferencial de los ítems.
Relaciones con otras variables	✓	• Evidencia convergente y discriminante (puede involucrar evidencia experimental como correlacional). • Relaciones prueba-criterio (diseños predictivos y concurrentes) • Generalización de validez (estudio de generalización de validez)
Consecuencias del test	✓	• Interpretación y usos de puntajes de la prueba previstos por los desarrolladores de la prueba. • Afirmaciones hechas sobre el uso de la prueba que no se basan directamente en interpretaciones de los puntajes de la prueba. • Consecuencias imprevistas.

Tabla 2.16. Fuentes de evidencia de validez según AERA, APA, y NCME (2014) y alternativas disponibles para abordar la carencia identificada.

2.3.3. Consideraciones para este estudio en términos de TEA y AA.

Según las 3 categorías en las que podrían agruparse los vínculos entre LA y AA según Gašević et al. (2022), el objetivo de esta tesis quedaría enmarcado en la tercera categoría, la cual busca proponer y validar nuevas formas de evaluación. En concreto nos hemos basado en el análisis de los RP de los participantes al responder los ítems de las pruebas de CD. Las pruebas fueron administradas por ordenador utilizando diferentes formatos de ítems, desde los tradicionales ítems de opción múltiple hasta ítems dinámicos como las simulaciones interactivas. En concreto, nos hemos centrado en formatos de preguntas interesantes para medir habilidades de orden superior de acuerdo con los niveles intermedios y avanzados de DigComp.

Además, cabe mencionar que, para el alcance de este estudio, de las cinco formas posibles para contribuir a la validación de las inferencias realizadas entre las afirmaciones y el comportamiento observado de los participantes según (Oranje et al., (2017), nos hemos centrado en: 1) para generar y probar inferencias sobre el constructo de interés; y 2) para mejorar los diseños de los ítems analizando los patrones de comportamiento mostrados por los participantes.

En la tabla 2.17 puede verse las consideraciones más destacables para este estudio en términos de TEA, ET y AA, y con el objetivo de analizar los RP de quienes se examinan.

Campo	Característica
TEA	Desarrollo de modos de evaluación más pertinentes utilizando formatos de ítems innovadores y más cercanos a la práctica real, así como obtener información sobre el comportamiento y el rendimiento de los participantes durante las pruebas de evaluación.
	Uso de los datos registrados para mejorar la interpretación de las puntuaciones o para contribuir a la evaluación de la validez
	Validación del diseño de las tareas de evaluación para que desencadenen los conocimientos y habilidades esperados, y comprender el contexto en el que tienen lugar.
	Mejorar la calidad de las herramientas de evaluación. Obtención de suficientes evidencias que la soporten. Especial interés en las basadas en RP por su actual escasez.
	Automatización de la puntuación, diseño y desarrollo de las pruebas, así como su ensamblado.
	Facilita la provisión de comentarios constructivos, adecuados y fáciles de entender.
AA	Validez de la evaluación (Análisis y apoyo del argumento de validez)
ET	Amplía la información de los RP proporcionando observaciones como el comportamiento estratégico mostrado, las secuencias llevadas a cabo y los patrones de respuesta. Esta información puede ser utilizada para el diseño y validación de las pruebas.
	Posibilita el modelado de individuos de diferentes niveles de habilidad. El modelado de MO puede ser útil para guiar la atención y promover el rendimiento de novatos.
	Análisis de estrategias cognitivas examinando los lugares observados y el orden seguido. De especial interés, el análisis de las interacciones con herramientas digitales y el análisis de los elementos visuales de las páginas web más fijados y caminos seguidos para crear una ruta de exploración común.
	Selección de métricas cuantitativas basadas en fijaciones, ya que es ahí donde se produce la mayor parte de la adquisición de información y el procesamiento cognitivo. Uso de sus duraciones como un proxy para medir su carga cognitiva y dificultad.
	Uso de métodos de análisis cualitativos y exploratorios basados en técnicas de visualización para ayudar a describir mejor la estrategia de los participantes.

Tabla 2.17. Consideraciones más destacables para este estudio en términos de TEA, ET y AA.

2.4. Test Adaptativo Informatizado (TAI)

Gracias a los últimos avances tecnológicos se ha favorecido la creación de entornos de evaluación más realistas usando formatos de elementos innovadores. Además, estos avances han permitido diseñar una experiencia de aprendizaje centrada en el

estudiante, proporcionando herramientas y mecanismos que pueden modelar los objetivos, necesidades y preferencias de aprendizaje de los estudiantes, los niveles cognitivos, el grado actual de conocimiento y los requisitos específicos para una evaluación electrónica eficaz (Benchoff et al., 2018). Una solución para ofrecer pruebas de evaluación más eficaces es aplicar pruebas adaptativas haciendo uso de la tecnología, como parte importante e integral del aprendizaje adaptativo (Troussas et al., 2020). Además, el uso de dispositivos como ordenadores permiten aumentar la precisión estadística de las puntuaciones mediante pruebas adaptativas informatizadas (TAI). En lugar de mostrar a cada participante la misma prueba, en las pruebas adaptativas después de cada nueva respuesta, haciendo uso de LA, se actualiza la estimación de la capacidad del participante y se selecciona el siguiente ítem para que sea el que tenga propiedades óptimas de cara a una nueva estimación (Van der Linden y Glas, 2010). Consecuentemente, cada participante puede acabar respondiendo a un conjunto diferente de ítems, aquellos que son más informativos con respecto a sus estimaciones de habilidad. Además, la administración individualizada de los ítems basada en TAI da lugar a evaluaciones más cortas sin perder la precisión de las mediciones (Gardner et al., 2004).

Con el desarrollo de las distintas tecnologías apoyadas en la web, los TAI han despegado exponencialmente en las últimas décadas (Zenisky y Luecht, 2016), y las pruebas adaptativas son, hoy en día, un procedimiento bien establecido a nivel mundial en la evaluación educativa y psicológica. En concreto, se han desarrollado varias plataformas que proveen pruebas adaptativas vía web tal y como aparecen listadas en la web de la International Association for Computerized Adaptive Testing[45], siendo probablemente la plataforma *Concerto* (Scalise y Allen, 2015) la más reconocida en este campo y cuyo desarrollo sigue en activo. *Concerto* está diseñada para una implantación autónoma y utiliza la biblioteca *catR* (Magis y Raîche, 2012), que ofrece diferentes estrategias de pruebas basadas en la *teoría de la respuesta al ítem* (TRI), así como varios métodos de selección del siguiente ítem y reglas de parada. Otra plataforma destacable es el *Open Source Computerized Adaptive Testing System* (OSCATS) (Wietsma, 2016), que hace uso de una biblioteca en C que implementa los algoritmos de TRI y de selección de ítems más utilizados en TAI.

Además, también se han aplicado evaluaciones adaptativas a las evaluaciones educativas internacionales a gran escala como, p. ej. el *Programa para la Evaluación Internacional de Estudiantes* (PISA) de la *Organización para la Cooperación y el Desarrollo Económico* (OCDE) y el *Programa para la Evaluación Internacional de las Competencias de los Adultos* (PIAAC). La mayoría de estas implementaciones realizan la adaptación a nivel de ítem, pero recientemente han surgido avances en la adaptación basada en *testlets*, también llamada prueba multietapa, la cual facilita el control del contenido (Yan et al., 2016).

Normalmente, cuando se diseña una prueba convencional se suelen incluir ítems que cubran el rango completo de habilidades que se pretende medir, es

[45] http://www.iacat.org/

decir, desde los niveles más bajos hasta los niveles más altos. El objetivo no es otro que poder cuantificar como de hábil es el participante que responde bien la mayoría de los ítems, o, al contrario, como de inhábil es el participante que responde mal a la mayoría de los ítems (Wainer et al., 2000). Normalmente, teniendo en cuenta que la mayoría de los participantes se suelen encontrar en la mitad del rango de habilidades, una práctica muy extendida consiste en adoptar una proporción donde el 50% de los ítems son de un nivel de dificultad medio, el 25% de nivel de dificultad más bajo y el restante 25% de nivel de dificultad más alto. Esto implica que los participantes muy hábiles puedan llegar a aburrirse y a desmotivarse, por tener que responder una gran cantidad de ítems muy fáciles para ellos antes de obtener el resultado, causándoles una sensación de pérdida de tiempo. En el caso de los participantes con bajo nivel de habilidad, paso algo similar, pero con los ítems de mayor dificultad, ya que de forma continuada fallan los ítems muy difíciles, sin que estos resultados aporten nueva información sobre su nivel de habilidad y causando al participante una sensación de malestar y frustración.

El desarrollo de la teoría de la respuesta al ítem (TRI) a mediados del siglo pasado ha proporcionado una base psicométrica sólida para el diseño de pruebas adaptativas. La característica clave de la TRI es su modelización de las probabilidades de respuesta para un ítem con parámetros distintos para la capacidad del examinando y las características de los ítems. En la TRI el foco está puesto a nivel de ítem, y no a nivel de test como en la teoría clásica de los test (TCT). Esta característica permite afrontar la cuestión estadística de los valores óptimos de los parámetros de los ítems para la estimación de la capacidad del participante. La principal respuesta a esta cuestión la dio Birnbaum (1968), quien, para la medida de información de Fisher, demostró que el ítem óptimo es el que tiene el valor más alto para su parámetro de discriminación y un valor para el parámetro de dificultad igual a la capacidad del participante. El objetivo final no es otro que poder ordenar los ítems en una escala por nivel de dificultad, de manera que nos permita usar esta escala para poder situar a los participantes según su nivel de habilidad. De esta forma, no sería necesario administrarle todos los ítems a un participante para lograr su nivel de habilidad, bastaría con administrarle los ítems de la escala que se encuentren cerca de su zona de habilidad.

Los TAI surgen en este contexto, gracias a la posibilidad de poder estimar el nivel de rasgo de los participantes en una misma escala, incluso sin haber mostrado todavía ningún ítem en común a todos los participantes. Un TAI comienza administrando ítems y estimando la capacidad inicial del participante. A continuación, se siguen proporcionando más ítems al participante de forma iterativa hasta que su capacidad se estima con una precisión determinada. La adaptación se basa en la forma en que el sistema de evaluación de la competencia selecciona los ítems necesarios para la estimación, ajustando la dificultad de las preguntas en función del rendimiento de la persona examinada. Si lo hace bien, se le administrarán ítems más difíciles. En caso contrario, se le administrarán ítems más fáciles. Luego, es de vital importancia mostrar los ítems que mejor se ajusten al nivel de los participantes para poder obtener la información más útil (Wainer et al., 2000).

Aun así, es necesario optimizarlas para que los participantes puedan recibir una evaluación precisa en el menor tiempo posible.

Durante los últimos años, los investigadores han ido desarrollando distintos enfoques sobre cómo puede funcionar el algoritmo adaptativo, a distintos niveles: adaptando la dificultad después de cada ítem respondido, adaptando la dificultad a nivel de bloques de ítems en pruebas multietapa (TME), y adaptando la prueba de formas totalmente diferentes, p. ej. haciendo uso de árboles de decisión basados en modelos de aprendizaje automático, o basados en modelos de diagnóstico cognitivo (DC).

Hasta hace poco, la mayoría de los esfuerzos de investigación se han centrado en la selección de los ítems de las pruebas teniendo en cuenta diferentes características de los participantes, como, p. ej. su nivel de conocimientos (Fang et al., 2017). También se han diseñado soluciones novedosas como, p. ej. el TAI propuesto por Chrysafiadi et al. (2020), combinando la lógica difusa y las teorías cognitivas para mejorar la personalización y la adaptabilidad. O el TAI diseñado por Badaracco y Martínez (2013), basado en un nuevo algoritmo de selección de ítems haciendo uso de un modelo de decisión que integra el conocimiento de los expertos modelado por lógica difusa. Otros estudios, como p. ej. los llevado a cabo por Bernardi y Di Mascio (2019), Oppl et al. (2017) o Chalmers (2016), se han centrado en el uso de la TRI y en cómo puede contribuir en los TAI. En cambio, otros autores han aplicado el aprendizaje profundo en las pruebas adaptativas, como Wang et al. (2017), para identificar los ítems a mostrar en las pruebas basándose en sus preferencias. Más aún, varios estudios como p. ej. los de Graf et al. (2017) y Tseng (2016), se centraron en modelar las habilidades de los estudiantes con el objetivo de diseñar pruebas adaptativas más eficaces. Otros enfoques aplicados a las pruebas adaptativas incluyen la búsqueda de la mejora de la herramienta de evaluación en sí misma. Por ejemplo, Vie et al. (2017) pusieron el foco en la retroalimentación entregada al participante, mientras que Pan y Lin (2018) se centraron en los aspectos técnicos del diseño y la implementación del sistema de tutoría propuesto.

Como resultado de la evolución de la teoría psicométrica y la aparición de nuevos modelos, se ha hecho posible la aplicación de TAI con diferentes formatos de respuesta (p. ej. politómicos o continuos), y pruebas que evalúan más de una dimensión utilizando modelos TRI multidimensionales y bifactoriales. Esto ha permitido el desarrollo de TAI basados en estos modelos, en los cuales se supone que los rasgos latentes subyacentes son continuos. No obstante, las metodologías adaptativas se han aplicado recientemente a un nuevo marco psicométrico, el marco de modelización del DC, el cual surgió con el propósito de clasificar diagnósticamente a los participantes en un conjunto predeterminado de rasgos latentes discretos, también denominados atributos. Los atributos son de naturaleza discreta, y normalmente sólo hay dos niveles para indicar si los participantes dominan o no cada atributo específico. De acuerdo con Ma (2019) Zhan et al. (2018), DC es un área de investigación muy activa sólo cuenta con un pequeño número de investigaciones en comparación con las que se han llevado a cabo en el

contexto de la TRI (p. ej. Cheng, 2009; Hsu et al., 2013; Kaplan et al, 2015; Xu et al., 2016; Yigit et al., 2019).

2.4.1. La Teoría de Respuesta al Ítem (TRI)

La TRI surgió en parte para abordar algunos de los inconvenientes de la tradicional TCT, que establece que la puntuación empírica de un examinando en una prueba es igual a la suma de su puntuación verdadera y el error. La TRI intenta dar un fundamento probabilístico al problema de la medición de rasgos y constructos no observables, o rasgos latentes (Baker y Kim, 2004; Muñiz Fernández, 1997).

La TRI se utiliza ampliamente para calibrar y evaluar los ítems de los instrumentos de evaluación y proporciona formas de evaluar las propiedades de los instrumentos en términos de fiabilidad y validez (Aesaert et al., 2014; Hambleton y Jones, 1993; Muñiz Fernández, 1997; Rasch, 1993; Thissen, 1982; Wilson, 2004). El punto fuerte de la TRI es que permite obtener evaluaciones invariantes respecto a las herramientas de evaluación utilizadas y de los participantes involucrados.

Desde el contexto de la TRI, los objetivos claves son: 1) establecer modelos estadísticos que permitan evaluar su grado de ajuste a los datos; 2) realizar estimaciones invariantes de los niveles de rasgo de los participantes y de las propiedades psicométricas de los ítems; y 3) obtener para cada participante datos de la precisión de la medida.

Ventajas de la TRI

Los principales rasgos y ventajas que caracterizan a la TRI son: 1) La existencia de rasgos latentes que pueden explicar el comportamiento de un participante en una prueba; 2) La relación entre el rendimiento y el conjunto de rasgos evaluados; 3) La especificación de la dimensionalidad; 4) La posición del ítem en el conjunto de valores del rasgo; 5) Instrumentos de evaluación con propiedades que no dependen del grupo específico de participantes ni de ítems mostrados. Tanto los ítems como los participantes reciben una puntuación en la misma escala al mismo tiempo; 5) Las unidades básicas de análisis se basan en los ítems y no en el instrumento de evaluación; y 6) La fiabilidad de un instrumento de evaluación depende de la acción entre el participantey el instrumento de evaluación.

Supuestos de la TRI

En los modelos TRI se asume que existe una relación entre los valores del rasgo que evalúan los ítems y la probabilidad de responderlos correctamente. A esta relación funcional se la conoce como curva característica del ítem (CCI). Es decir, la probabilidad de responder correctamente a un ítem depende de los valores del rasgo evaluado por el ítem, con lo cual, personas que obtengan distintos valores tendrán distintas probabilidades en responder correctamente al ítem. La CCI es específica para cada ítem y lo caracteriza. El tipo de curva a utilizar deberá

corresponderse con la modelización del comportamiento del rasgo a evaluar. De acuerdo con Hambleton et al. (1991), los parámetros para caracterizar las CCI y los ítems incluyen su dificultad (sitúa el ítem en la escala de habilidad que establece la probabilidad de ser contestado correctamente), la discriminación (representa el grado de variación en la tasa de éxito de los individuos en función de su habilidad), y un parámetro de adivinación. La CCI es definida cuando se dota de valor a estos tres parámetros y se establece una función matemática específica para la curva.

Otro supuesto de la TRI es el de la unidimensionalidad e independencia local. Como hemos visto previamente, la probabilidad de que un participante responda correctamente a un ítem sólo dependerá del rasgo que se pretende evaluar, asumiendo de manera implícita que los ítems que evalúan ese rasgo constituyen una única dimensión (Embretson y Reise, 2013). La TRI se basa en el principio de que es posible medir rasgos latentes, es decir, rasgos que no son directamente observables. Más aun, se dice que un modelo es unidimensional cuando el rendimiento en un ítem depende sólo de un rasgo latente, y que el modelo es multidimensional cuando el rendimiento en un ítem depende de más de un rasgo. Además, la dimensionalidad observada podría servir para confirmar si los datos obtenidos son consistentes con la teoría que respalda el constructo evaluado y las puntuaciones obtenidas en la prueba.

Consecuentemente, en un modelo unidimensional existe independencia local entro los distintos ítems, es decir, si un participante tiene un valor dado para el rasgo evaluado de carácter unidimensional, su respuesta a un ítem no es influida por las respuestas que haya dado el participante a otros ítems. Cabe destacar que podría darse el caso de existir independencia local con datos no unidimensionales, tal y como sucede en los modelos multidimensionales de la TRI.

Modelos de la TRI

Un modelo relaciona matemáticamente la probabilidad de que un participantede una determinada respuesta a un ítem, con otras características del participante como, p. ej. su nivel en un rasgo en concreto, y ciertas características del ítem como, p. ej. su grado de dificultad o de discriminación. Las funciones matemáticas que más se han utilizado para definir la CCI, el principal distintivo de la TRI, son la función logística y la curva normal acumulada. Ambas funciones pueden ser aplicadas a los 3 parámetros que definen la CCI con lo cual podremos llegar a tener hasta 6 modelos diferentes.

La aplicación del modelo de TRI más adecuado depende, en primer lugar, de las características de los ítems utilizados. Para los ítems dicotómicos, como los de ETCD, los modelos más utilizados son los modelos logísticos de 1, 2 o 3 parámetros. La función logística hasta ahora ha sido la más utilizada gracias a su mayor trazabilidad. Los modelos más utilizados de la TRI son unidimensionales y son:

- Modelo logístico de 1 parámetro o modelo de Rasch (Rasch, 1960), para modelos dicotómicos, asume que el rendimiento en un ítem sólo depende del nivel de rasgo del participante y de la dificultad del ítem.
- Modelo logístico de 2 parámetros, contempla a parte de la dificultad un parámetro de discriminación del ítem indica en qué proporción discrimina el ítem entre los niveles inferiores y superiores a la dificultad del ítem.
- Modelo logístico de 3 parámetros, también contempla un tercer parámetro con la probabilidad de acertar el ítem al azar por parte de personas con nivel de habilidad extremadamente bajo.

Cabe mencionar que en la literatura existen otros tipos de modelos para respuestas no dicotómicas (Nering y Ostini, 2011; Van Der Ark, 2001). Además, la validez interna de una prueba se evalúa en función del ajuste de los ítems al modelo seleccionado. El método más utilizado en la estimación de los modelos es el de *máxima verosimilitud marginal* (ML), y supone que los parámetros de un individuo son variables aleatorias con una determinada distribución (Thissen, 1982).

Los resultados de estos 3 modelos son válidos sólo en la medida en que las, dimensiones son diferentes y claras, es decir, no hay ítems que evalúen diferentes variables al mismo tiempo, y por lo tanto el supuesto de unidimensionalidad es realista. De ahí que hayan aparecido otros modelos, como los modelos multidimensionales de la TRI, que consideran un constructo formado por varios factores. El *modelo multidimensional random coefficient multinomial logit* (MRCML) se presentó como alternativa al *análisis factorial confirmatorio* (AFC) (Reckase, M. D. (2009). El AFC y la TRI multidimensional son métodos muy aplicados para validar una posible organización de la información. El modelo multidimensional de Rasch es el más sencillo de los modelos multidimensionales y supone que todas las cargas de los ítems se fijan en la unidad con el modelo de Rasch (Adams et al., 1997).

Para seleccionar un modelo en concreto, será necesario llevar a cabo una serie de comprobaciones para evaluar su aplicabilidad:

- Comprobar la unidimensionalidad del rasgo a evaluar tal y como exigen los modelos: el método tradicionalmente más utilizado es el *análisis factorial exploratorio* (AFE), en el que se busca comprobar cómo un solo factor da cuenta de un grado elevado de la varianza de los ítems. A pesar de que han surgido mejores alternativas como la propuesta por Lord (2012), la varianza explicada por el primer factor es un indicador muy utilizado e intuitivo de la unidimensionalidad, ya que muestra el grado en el que reduciendo los ítems a un factor explica una parte importante de la varianza explicada por todos los ítems. El AFC es otra opción más aconsejable para evaluar la unidimensionalidad de los datos (Brown, 2015), que posibilita establecer especificaciones sobre la estructura inicial de los datos, haciendo uso de teorías que se manejen y en resultados previos. Una vez aplicado un modelo de análisis factorial, es recomendable verificar los índices de bondad de ajuste para validar

la solución obtenida, como p. ej. un valor de *Chi-cuadrado* no significativo. Existen otros índices de ajuste que también pueden ser examinados de manera conjunta como el *ajuste comparativo* (CFI), el ajuste no normado (TLI) y el error cuadrático medio de aproximación (RMSEA).

- Elección del modelo a utilizar: Los datos con los que contamos son los que nos van a orientar en la elección del modelo con el mejor ajuste posible, p. ej. si los índices de discriminación no son iguales no tiene sentido aplicar el modelo de Rasch.
- Estimar el valor de los parámetros: Después de administrar la prueba a un grupo de participantes por primera vez, se analizan las respuestas para aplicar el modelo que mejor se ajuste y así poder estimar los distintos parámetros requeridos (Hambleton y Swaminathan, 2013). A este paso, comúnmente se le conoce como proceso de calibración. Cabe mencionar que cuantos más parámetros tengamos que estimar para los ítems de acuerdo con el modelo elegido, mayores serán las exigencias en cuanto al tamaño de la muestra de participantes necesario para llevar a cabo la calibración. En concreto, una de las razones por las cuales tradicionalmente el modelo de Rasch ha sido muy utilizado es porque puede ser suficiente un tamaño de muestra de 200 participantes para obtener un buen nivel de precisión (Lord, 1983).
- Comprobar el ajuste del modelo seleccionado: Se comparan los resultados pronosticados con los obtenidos, examinando si estadísticamente difieren o no, mediante distintos procedimientos estadísticos como el análisis de los residuos, la comparación de las distribuciones de las puntuaciones y el uso de chi-cuadrado (Pina y Montesinos, 1996).

Estimación del nivel de rasgo y precisión de las estimaciones

A partir de que el banco de ítems esté calibrado, ya se podrán estimar los niveles de habilidad de los nuevos participantes. Cuando la estimación está condicionada por el conocimiento previo de los parámetros de los ítems, se conoce como *estimación condicional de máxima verosimilitud,* y cuando se desconocen se conoce como *estimación conjunta de máxima verosimilitud*. A parte de las estimaciones de máxima verosimilitud, las *estimaciones bayesianas* están creciendo en uso, las cuales incorporan distribuciones y permiten realizar una mejor estimación de los parámetros (Gifford y Swaminathan, 1990; Lord, 1986; Mislevy, 1986).

La TRI permite estimar el valor del nivel de habilidad ϴ proporcionando una medida de la precisión de las estimaciones mediante el *error típico de estimación*, pero se usará más la *función de información* (Birnbaum, 1968) que no es más que otra forma de expresar el error típico y sirve como indicador de la precisión de una prueba. Más aún, la función de información de una prueba es la suma de las funciones de información de los ítems. En cuanto al error típico de medida para un determinado nivel de rasgo, hay una serie de factores que influyen de manera notable como, por ejemplo, cuanto más discriminativos sean los ítems mayor será la

información, cuanto más bajos sean los índices de azar de los ítems, mayor será la información, cuanto más ítems tenga la prueba mayor será la información, o cuanto más cercanos se encuentren el nivel de rasgo del participante y los parámetros de dificultad de los ítems, mayor será la información.

Diseño del banco de ítems

Para diseñar y construir un banco de ítems una vez definido el constructo a evaluar, hay que desarrollar los ítems de forma adecuada, para posteriormente estimar los parámetros de acuerdo con el modelo seleccionado (Baker y Kim, 2004; Lane et al., 2016; Muñiz, 2018; Zickar, 2020). El uso de la tecnología ofrece una serie de ventajas como, el diseño de ítems innovadores (Cuhadar y Binici, 2022; Drasgow, 2016; Kato y de Klerk, 2017; Sireci y Zenisky, 2015), la generación automatizada de ítems (Attali, 2018; Gierl y Haladyna, 2012; Gierl y Lai, 2018), y la optimización de las pruebas (Van der Linden, 2018).

El proceso de construcción de un banco de ítems debe llevarse a cabo de manera rigurosa basado en unos estándares de calidad (AERA, APA y NCME, 2014), si se pretende validar las inferencias realizadas a partir de las puntuaciones obtenidas por los participantes en las pruebas (Lane et al., 2016). Todas las acciones que se realicen durante el diseño del banco de ítems y del instrumento, facilitarán evidencias que ayudarán a interpretar las puntuaciones obtenidas.

Propiedades psicométricas de la prueba

Después de haber aplicado la prueba a una muestra de participantes, se lleva a cabo un análisis de las propiedades psicométricas de las puntuaciones obtenidas, que consiste en realizar un análisis de los ítems (examinar los índices de dificultad y discriminación, así como el funcionamiento diferencial de los ítems), realizar una estimación de la fiabilidad de las puntuaciones, obtener todo tipo de evidencias de validez (p.ej. estudiar la dimensionalidad de la herramienta para obtener evidencias de validez de la estructura interna) y, por último, la construcción de baremos de puntuaciones. El rigor metodológico es clave en esta fase, con lo cual todos los pasos y decisiones tomadas deben de estar debidamente documentadas y justificadas.

En la tabla 2.18 puede verse los aspectos más destacables de aplicar la TRI.

Campo	Característica
Ventajas de la TRI	Proporciona un fundamento probabilístico al problema de la medición de rasgos y rasgos latentes. Permite obtener evaluaciones invariantes respecto a las herramientas de evaluación utilizadas y de los participantes involucrados (Baker y Kim, 2004; Muñiz Fernández, 1997).

	Permite calibrar y evaluar los ítems de los instrumentos de evaluación, proporcionando formas de evaluar las propiedades de los instrumentos en términos de fiabilidad y validez (Aesaert et al., 2014; Hambleton y Jones, 1993; Muñiz Fernández, 1997; Rasch, 1993; Thissen, 1982; Wilson, 2004).
Supuestos de la TRI	Relación entre los valores del rasgo que evalúan los ítems y la probabilidad de responderlos correctamente (CCI) (Hambleton et al., 1991).
	Unidimensionalidad e independencia local (Embretson y Reise, 2013).
Modelos de la TRI	Modelo logístico de 1 parámetro o modelo de Rasch (Rasch, 1960), de 2 parámetros (Swaminathan y Gifford, 1985) y de 3 parámetros (Birnbaum, 1968).
	Mdelos para respuestas no dicotómicas (Nering y Ostini, 2011; Van Der Ark, 2001).
	Método más utilizado en la estimación de modelos, ML (Thissen, 1982).
	Modelos multidimensionales de la TRI, como el modelo multidimensional random coefficient multinomial logit (MRCML) (Reckase, M. D. (2009) o el modelo multidimensional de Rasch, el más sencillo de los modelos multidimensionales Rasch (Adams et al., 1997).
Aplicando un modelo y comprobaciones correspondientes	Comprobación de la unidimensionalidad del rasgo a evaluar tal y como exigen los modelos Lord (2012), AFE y AFC (Brown, 2015). Índices de bondad de ajuste: Chi-cuadrado, CFI, TLI y RMSEA. Los datos condicionarán el modelo a utilizar.
	Estimación del valor de los parámetros (Hambleton y Swaminathan, 2013). Rasch ha sido muy utilizado porque requiere tamaños de muestra relativamente pequeños para obtener un buen nivel de precisión (Lord, 1983).
	Comprobación del ajuste del modelo seleccionado (Pina y Montesinos, 1996).
Estimación del nivel de rasgo y precisión de las estimaciones	Estimaciones de máxima verosimilitud y estimaciones bayesianas (Gifford y Swaminathan, 1990; Lord, 1986; Mislevy, 1986).
	Estimación del valor del nivel de habilidad θ proporcionando una medida de la precisión de las estimaciones mediante la función de información (Birnbaum, 1968).
Diseño del banco de ítems	Clave el desarrollo de los ítems de forma adecuada (Baker y Kim, 2004; Lane et al., 2016; Muñiz, 2018; Zickar, 2020), basado en unos estándares de calidad (AERA, APA y NCME, 2014).
	La TEA facilita el diseño de ítems innovadores (Cuhadar y Binici, 2022; Drasgow, 2016; Kato y de Klerk, 2017; Sireci y Zenisky, 2015), la generación automatizada de ítems (Attali, 2018; Gierl y Haladyna, 2012; Gierl y Lai, 2018), y la optimización de las pruebas (Van der Linden, 2018).

Tabla 2.18. Aspectos más destacables de aplicar la TRI.

2.4.2. Pruebas lineales y pruebas adaptativas

Las formas más comunes de pruebas son los test lineales (que tradicionalmente han sido la forma de evaluación más utilizada sobre todo en el contexto educativo), los TAI y los test multietapa. A continuación, describimos las dos últimas en detalle:

Test adaptativos informatizados (TAI)

El objetivo principal de un TAI es producir una prueba más eficaz y precisa que una prueba lineal. Los TAI utilizan un enfoque distinto, haciendo uso de distintos algoritmos para administrar los ítems del test. En concreto, los ítems a administrar se adaptan al nivel de capacidad estimado del participante durante el proceso de la prueba, que se actualiza continuamente después de la administración de cada ítem en base a la respuesta mostrada. Es decir, un TAI es un test adaptativo a nivel de ítem, y pueden tener una longitud fija o variable de número de ítems.

Los TAI gracias a que basan su evaluación en el nivel de capacidad estimado para un determinado participante, pueden proporcionar una medición precisa para todas las capacidades de los participantes (Lord, 1974; Wainer et al., 1992). En cambio, las pruebas lineales tienen dificultades para proporcionar mediciones precisas para toda la gama de niveles de capacidad presentes en el grupo de participantes (Hambleton y Swaminathan, 1985; Lord, 1980). Tal y como se muestra en la figura 2.3, los principales elementos de un TAI son: un banco de ítems cuyos ítems han sido previamente calibrados; el proceso de selección de los ítems óptimos en cada momento; el proceso de estimación del nivel de habilidad después de la administración de cada ítem y el procesamiento de su respuesta; los criterios de parada; y la estimación de la capacidad y su puntuación final.

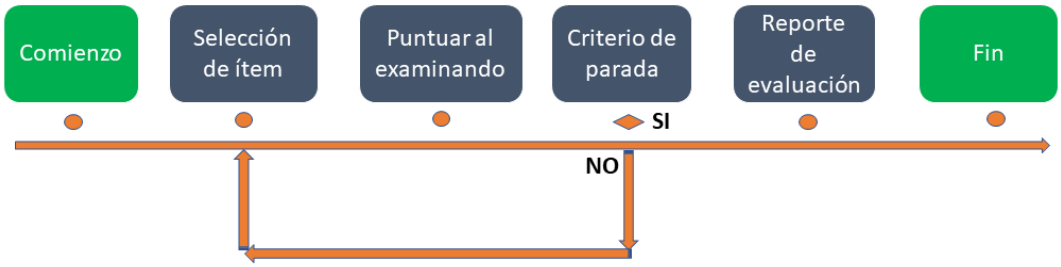


Figura 2.3. Ejemplo de un diseño TAI con adaptación a nivel de ítem

Test multietapa (TME)

Los TME comparten características tanto de los test lineales como de los TAI, minimizando sus desventajas y haciendo que cada vez su aplicación esté más extendida (Hambleton y Swaminathan, 1985; Lord, 1980; Luecht, 1998; Steinfeld y Robitzsch, 2021; van der Linden y Glas, 2010; Wainer, Bradlow, y Wang, 2007; Yamamoto et al, 2018; Yan et al., 2014).

Los TME también implementan la adaptación en base al nivel de capacidad estimado para cada participante, pero la llevan a cabo a nivel de grupos de ítems previamente ensamblados (también llamados módulos). Al igual que los TAI, los TME al basar la evaluación en el nivel de capacidad de un participante dado, pueden proveer una medición precisa para todos los participantes, independientemente de donde se encuentre su nivel de capacidad en la escala de medición (Yan et al., 2014). Con este fin, la prueba es diseñada por etapas, tal y como se muestra en el ejemplo de la figura 2.7. En los TAI tradicionales, cuando se realiza la primera estimación de la capacidad utilizando sólo la primera respuesta tiene un alto sesgo. En cambio, un TME comienza la primera etapa administrando a todos los participantes un conjunto inicial de ítems, normalmente llamado *prueba de enrutamiento*, realizando la adaptación sólo una vez que se ha reunido suficiente información (módulo 1 en la figura 2.4). En función de las respuestas dadas por los participantes en esta etapa, son redirigidos a uno de los diferentes módulos de la siguiente etapa en base al nivel de capacidad estimado (módulos 2 y 3 en la figura 2.4). Este funcionamiento consigue reducir la longitud de la prueba sin tener que perder mucha información, ya que los ítems mostrados tienen niveles de capacidad específicos, consiguiendo una mayor precisión que en una prueba lineal. Además, los diseños de TMEs son flexibles y el número de etapas y módulos por etapa pueden variar. La última etapa de los TMEs suele denominarse *prueba de medición* (módulos 4, 5 y 6 en la figura 2.4).

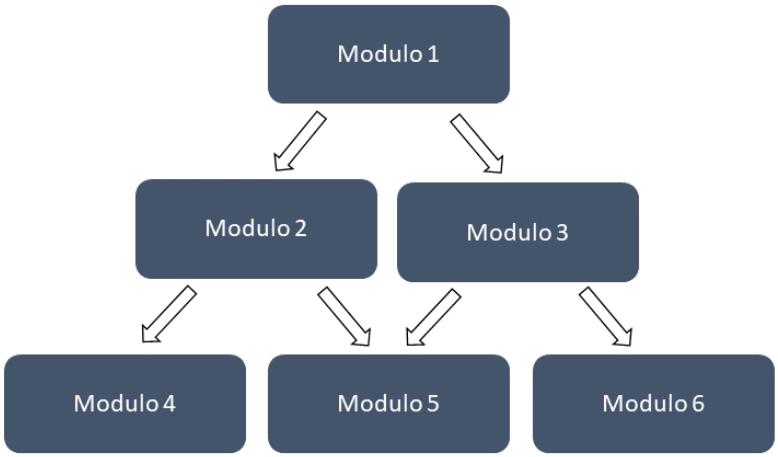


Figura 2.4. Ejemplo de un diseño TME

La arquitectura compuesta por la etapa de enrutamiento, así como sus posteriores etapas y módulos, suele denominarse *panel*. De hecho, una práctica común es diseñar un TME en varios paneles para así poder mejorar la seguridad de la prueba y la tasa de exposición de sus elementos. Los distintos módulos tienen diferentes niveles de dificultad y, además, han de cumplir con los requisitos de la prueba sobre la exposición de los ítems y el balanceo de los contenidos.

Los TME constan de más de un punto de adaptación durante su administración (Yan et al., 2014). La idea fundamental es que cada nuevo módulo concentra más la precisión en la zona de la escala de habilidades en la que se encuentra el participante. Además, este tipo de pruebas permiten a los participantes revisar sus respuestas antes de pasar a la siguiente etapa, sin necesitar modelos complicados para la revisión de las respuestas (Wang et al., 2015).

Wang et al. (2016) sugirieron presentar un grupo de preguntas al principio de la prueba, cuando se dispone de poca información sobre la capacidad del participante, y reducir progresivamente el número de preguntas de cada etapa para aumentar las oportunidades de adaptación.

Más aún, al presentar las preguntas en grupos se facilita llevar a cabo un balance de contenidos en cada etapa.

A la hora de diseñar un TME, existen una serie de puntos clave a examinar y definir. Por ejemplo, el número de ítems de cada módulo, el número de módulos en cada etapa y el número de etapas (este punto vendrá determinado por la duración prevista para la prueba), el contenido de la prueba, y la precisión de la evaluación y clasificación. Al igual que en los TAI, es posible realizar distintos tipos de simulaciones para llevar a cabo este análisis previo durante la fase de diseño, haciendo uso de paquetes de software como *mstR* (Magis et al, 2017). A su vez, también es necesario diseñar el conjunto de ítems y módulos de las distintas etapas, así como establecer los requisitos de contenidos, sus dificultades, y la tasa de exposición de los ítems (Luecht, 2014).

Los dos enfoques más habituales en el diseño de un TME son el centrado en la precisión de la estimación de la capacidad para un rango de niveles de capacidad de los participantes, y el centrado en la mejora de la precisión de la clasificación de los examinados en sus respectivos grupos. Normalmente se basan en el marco de la TRI aplicándose principalmente en dos fases: en el ensamblaje del TME (en base a módulos y especificando las reglas de enrutamiento) y en su posterior aplicación (p. ej. Lord, 1971a; Luecht 2014; Luecht 1998). Sin embargo, no funcionan bien cuando las muestras son pequeñas o cuando se violan supuestos de la TRI como la unidimensionalidad. Como alternativa, Yan et al. (2014) propusieron un algoritmo adaptativo basado en árboles que parece funcionar tan bien como los basados en la TRI.

Una ventaja por destacar de los TME es que la puntuación de un ítem en la sección operativa no depende de puntuaciones de ítems de etapas anteriores y de la etapa en la que se presentó el ítem (Eggen y Verhelst, 2011). Además, las estimaciones del rasgo latente se calculan al finalizar la administración del módulo completo. Más aún, dentro del módulo no son necesarias las estimaciones provisionales de la capacidad de los participantes, y consecuentemente, la fase donde se realiza la estimación es más rápida que en los TAI.

En la tabla 2.19 puede verse las principales referencias identificadas y analizadas respecto a similitudes y diferencias entre los TAI y los TME.

Característica	TAI	TME	Referencia
Evaluación más eficaces aplicando pruebas adaptativas haciendo uso de la tecnología. Más cortas y sin perder precisión para todas las capacidades de los participantes.	✓	✓	(Gardner et al., 2004; Lord, 1974; Troussas et al., 2020; Wainer et al., 1992)
Después de cada nueva respuesta, se actualiza la estimación de la capacidad del participante y se selecciona el siguiente ítem de manera que sea el óptimo para una nueva estimación.	✓	✗	(Van der Linden y Glas, 2010; Wainer et al., 2000)
Adaptación de la dificultad después de cada respuesta, a nivel de bloques de ítems.	✗	✓	(Luecht, 2014; Yan et al., 2014)
Constan de más de un punto de adaptación durante su administración.	✗	✓	(Luecht, 2014; Yan et al., 2014)
Evaluación cada vez más asentada y extendida.	✓	✓	(Hambleton y Swaminathan, 1985; Lord, 1980; Luecht, 1998; Magis y Raîche, 2012; Magis et al, 2017; Scalise y Allen, 2015; Steinfeld y Robitzsch, 2021; van der Linden y Glas, 2010; Wainer, Bradlow, y Wang, 2007; Wietsma, 2016; Yamamoto et al, 2018; Yan et al., 2014; Yan et al., 2016; Zenisky y Luecht, 2016)
Selección de los ítems en el arranque de las pruebas teniendo en cuenta diferentes características de los participantes.	✓	✓	(Fang et al., 2017)
Diseño de soluciones novedosas: combinación de lógica difusa y teorías cognitivas, selección de ítems integrando el conocimiento de los expertos modelado por lógica difusa, aspectos técnicos del diseño, implementación de un sistema de tutoría, y uso de aprendizaje profundo.	✓	✗	(Badaracco y Martínez, 2013; Chrysafiadi et al., 2020; Pan y Lin, 2018; Vie et al., 2017; Wang et al., 2017)
Uso de la TRI y sus principales contribuciones.	✓	✓	(Bernardi y Di Mascio, 2019; Chalmers, 2016; Lord, 1971a; Luecht 2014; Luecht 1998; Oppl et al., 2017)
Modelado de las habilidades de los estudiantes con el objetivo de diseñar pruebas adaptativas más	✓	✗	(Graf et al., 2017; Tseng, 2016)

eficaces			
Posiblidad de revisar las respuestas antes de pasar a la siguiente etapa de manera sencilla.	✗	✓	(Wang et al., 2015)
Selección de un grupo de preguntas al principio de la prueba, y reducir progresivamente el número de preguntas de cada etapa para aumentar las oportunidades de adaptación.	✗	✓	(Wang et al., 2016)
Facilita llevar a cabo un balance de contenidos en cada etapa.	✗	✓	(Wang et al., 2015)
Cuando las muestras son pequeñas, uso de algoritmo adaptativo basado en árboles.	✗	✓	(Yan et al., 2014)
La puntuación de un ítem en la sección operativa no depende de puntuaciones en etapas anteriores, ni de la etapa en la que se presentó el ítem.	✗	✓	(Eggen y Verhelst, 2011)

Tabla 2.19. Principales referencias identificadas respecto a similitudes y diferencias de los TAI y los TME.

2.4.3. Modelos de evaluación adaptativa

Actualmente según Vie et al. (2017), existen dos enfoques predominantes en la evaluación adaptativa: el DC para la evaluación formativa, y los TAI para la evaluación sumativa.

Diagnósticos cognitivos

En un proceso de evaluación, descubrir los conocimientos latentes de los participantes de manera eficaz, puede ser crucial para poder adaptar posteriormente la experiencia de aprendizaje a sus necesidades. La metodología de TAI de diagnóstico cognitivo (DC-TAI) surgió para combinar la eficacia de los TAI con los resultados obtenidos de los modelos de DC (Cheng, Y. ,2009). Consecuentemente, se desarrollaron varios modelos con el objetivo de poder explicar el resultado en una tarea de aprendizaje de un participante en base a los componentes del conocimiento (CC) dominados. Por lo tanto, es necesario especificar para cada ítem cuales son los CC requeridos para poder resolverlo. Para llevar a cabo esta tarea, se suele diseñar una matriz binaria, conocida como q-matrix (Huebner, 2010). En la figura 2.5 se muestra un ejemplo de una posible q-matrix para el AC de IAD.

	Componentes de conocimiento											
	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	CC10	CC11	CC12
Item 1	1	0	0	1	0	0	0	0	1	1	0	0
Item 2	0	0	1	0	0	0	1	0	1	1	0	1
Item 3	0	1	0	1	0	1	0	0	1	1	0	0
Item 4	1	0	0	1	1	0	0	0	1	1	1	0
Item 5	0	0	0	0	0	0	1	1	1	1	0	0
Item 6	0	0	0	1	1	0	1	0	0	1	1	0
Item 7	0	0	1	1	0	0	0	0	1	1	1	0
Item 8	0	0	0	1	1	0	0	0	1	1	1	1
Item 9	1	0	0	1	0	1	0	0	1	1	1	1

CC1: Acceder a un navegador
CC2: Realizar búsquedas sencillas de texto
CC3: Realizar otro tipo de búsquedas sencillas
CC4: Realizar una búsqueda avanzada
CC5: Filtrar los resultados de una búsqueda
CC6: Diferenciar publicidad de los resultados
CC7: Identificar las URLs oficiales
CC8: Acceder a los detalles de la búsqueda
CC9: Almacenar URL en los marcadores del navegador
CC10: Utilizar una extensión del navegador
CC11: Evaluar la fiabilidad de la información encontrada
CC12: Evaluar la fiabilidad del sitio web

Figura 2.5. Ejemplo de q-matrix para el área de Información y alfabetización digital.

Un TAI puede representarse como un autómata en forma de árbol, cuyos estados son los ítems mostrados, y sus aristas se etiquetan con 0 o 1 según la respuesta del examinado (correcta o incorrecta). Por lo tanto, una ejecución del TAI no es más que un camino en el autómata de acuerdo con el rendimiento del participante.

Los dos modelos basados en q-matrix más utilizados son el modelo *Deterministic Input, Noisy And* (DINA) (De La Torre, 2011) y el *Modelo de Jerarquía de Atributos* (Leighton et al. 2004). En el modelo DINA, el participante debe dominar cada CC que interviene en la resolución de un determinado ítem para poder resolverlo. Sin embargo, cuando el número de CC evaluados es elevado, el número de estados posibles es grande y el siguiente ítem a mostrar no se puede calcular de forma eficiente. En el Modelo de Jerarquía de Atributos, lo que se hace es decrementar la complejidad estableciendo prerrequisitos entre las CC, p.ej. estableciendo que para dominar una CC requiera dominar previamente otra CC, reduciendo el número de estados posibles y su complejidad.

La principal diferencia entre un tradicional y los DC-TAI, radica en que en los DC-TAI el modelo de evaluación considera múltiples atributos latentes discretos, mientras que los TAI tradicionales suelen considerar uno o pocos rasgos latentes continuos. Estas diferencias han originado la necesidad de desarrollar la propia CD-TAI. Por ejemplo, el método de selección de ítems más utilizado en los TAI tradicionales es el estadístico de información de Fisher (Lehmann y Casella, 2006), el cual requiere niveles de habilidad continuos. Pero como los atributos en los DC son discretos, el estadístico de Fisher no puede ser utilizado. Afortunadamente, han surgido métodos alternativos para afrontar este hecho (Kaplan et al., 2015), como son la información de *Kullback-Leibler* (KL) (Chang y Ying, 1996; Cheng y Chang, 2007) o el índice de discriminación global (GDI) (De la Torre y Chiu, 2016).

Evaluación sumativa mediante la TRI

Gracias a la TRI, es posible diseñar pruebas a medida partiendo de un banco de ítems, en el que los ítems se agrupan según las estimaciones de los parámetros de los ítems. Como hemos visto previamente, para modelar un banco de ítems, existen tres modelos principalmente: el modelo de un parámetro (1PL) o modelo de Rasch (1993); el modelo de dos parámetros (2PL) (Swaminathan y Gifford, 1985) y el modelo de tres parámetros (3PL) o modelo de Birnbaum (1968). El modelo más sencillo utilizado para los TAI es el modelo de Rasch. Este modelo representa el comportamiento de los participantes con un único rasgo latente, la capacidad, y los ítems con un único parámetro, la dificultad. Por lo tanto, la probabilidad de que un participante resuelva correctamente una tarea sólo dependerá de la diferencia entre la dificultad de la tarea y la capacidad del participante. En concreto, la *función de información de la prueba* se convirtió en la estadística principal utilizada para el ensamblado de pruebas en el contexto de la TRI. Debido a la estrecha relación entre la *función de información de la prueba* y el error estándar de medición de la capacidad, es posible controlar el nivel de error de medición manipulando la *función de información de la prueba* (Zheng et al, 2014).

Una vez ajustado el banco de ítems al modelo seleccionado, los ítems se encuentran organizados según las estimaciones de los parámetros de los ítems. De esta forma, dada la información actualizada sobre la habilidad del participante y los ítems disponibles en el banco, es posible seleccionar los ítems que se consideren más oportunos para un participante dado ofreciéndole una prueba a medida.

El uso de la TRI en los TAI ha generado mucha atención hasta ahora (Bernardi y Di Mascio, 2019; Chalmers, 2016; Gibbons et al., 2016; Magis et al., 2017; Oppl et al., 2017; Reckase, 1974; Van der Linden y Glas, 2010; Wainer et al., 2000; Walker et al., 2010). La TRI ha permitido realizar potentes análisis cuantitativos en términos de mediciones, como el funcionamiento diferencial de los ítems, la vinculación de los parámetros de los ítems, la equiparación de las puntuaciones de las pruebas y los TAI. Actualmente, la TRI es la forma más utilizada de las teorías psicométricas, y puede utilizarse con una gran variedad de algoritmos de selección de ítems y procedimientos de puntuación (Chen y Wang, 2010).

Respecto al formato de los ítems de los TAI, Oppl et al. (2017) indicaron que los ítems de los TAI deberían de requerir opciones de respuesta más complejas y abiertas como, p.ej. incluyendo ítems dinámicos que requieran interacciones del usuario más complejas. Lamentablemente, este tipo de ítems suelen ser muy costoso de implementar y sólo puede administrarse cuando se utilizan soluciones tecnológicas de terceros. Oppl et al. (2017) abordaron esta limitación proporcionando una arquitectura flexible para diseñar TAI con ítems con distintos formatos, integrando la biblioteca *catR* (Magis y Raîche, 2012).

Además, existen modelos IRT tanto para TAI unidimensionales como para TAI multidimensionales donde la prueba estima múltiples atributos de la persona evaluada que representan más de un rasgo. A grandes rasgos, los principales componentes de un TAI son: un banco de ítems, un punto de entrada, un procedimiento para la selección de ítems, un método de puntuación y el criterio de

finalización de la prueba. Numerosos investigadores han demostrado que los componentes del TAI se abordan más fácilmente adoptando un modelo TRI unidimensional (Chalmers, 2016; Thompson y Weiss, 2011; Van der Linden y Glas 2010; Weiss, 2004). En cambio, los modelos multidimensionales requieren más parámetros dificultando su calibración (Desmarais y Baker, 2012).

El modelo de Rasch es el más utilizado en los TAI gracias a su simplicidad, su estabilidad y su fuerte base matemática (Desmarais y Baker, 2012). De esta forma se puede calibrar las dificultades de los ítems y las capacidades de los participantes de forma automática, calculando las estimaciones normalmente mediante una estimación ML. El proceso adaptativo lo que buscará es averiguar qué ítem será el más útil para afinar la estimación alcanzada hasta ese momento. Gracias a la información de Fisher, es posible cuantificar la información que cada ítem proporciona sobre el parámetro de habilidad.

Resumiendo, en una evaluación adaptativa, dada la estimación de la capacidad actual de un participante, se escoge el ítem óptimo que obtenga más información sobre su capacidad. El participante responde a ese ítem y en base al resultado obtenido, se estima de nuevo su habilidad y el intervalo de confianza para la estimación de la habilidad se refina. Y así sucesivamente, hasta que se cumplan las condiciones del criterio de parada y se dejen de administrar más ítems. En la figura 2.6, se muestra un ejemplo del proceso adaptativo llevado para un participante.

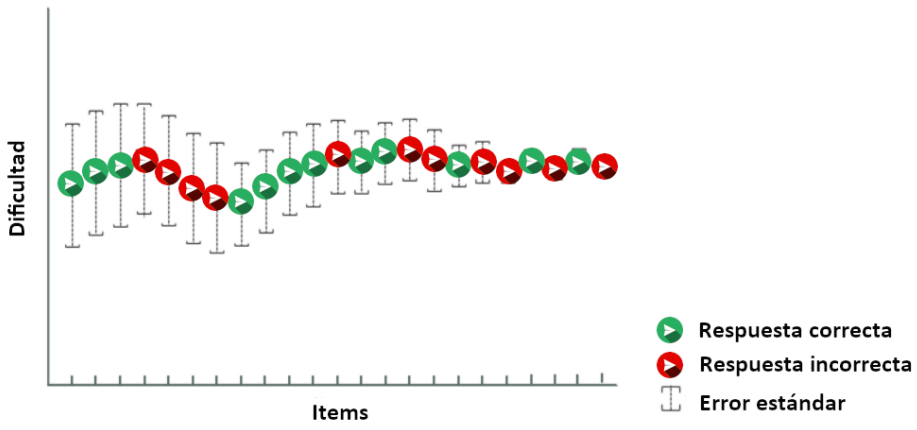


Figura 2.6. Evolución de la estimación de la capacidad a lo largo de un TAI basado en el modelo de Rasch.

Mas aún, si los ítems se agrupan en categorías, es posible mostrar al participante, información útil presentando por cada categoría su respectiva puntuación. Por ejemplo, si consideramos que cada respuesta equivale a un punto en el caso de que sea correcta, y 0 puntos en el caso de una respuesta incorrecta, entonces podríamos calcular las puntuaciones parciales como el número de puntos obtenidos en cada categoría respecto a la puntuación total. Además, si es conocida la puntuación total, podremos hacer uso del modelo de Rasch para obtener la

subpuntuación esperada de cada categoría. Y si además tenemos establecidos los niveles de habilidad esperados por cada categoría, se podría calcular el perfil de desviación para cada participante, mostrando la diferencia entre las puntuaciones parciales observadas y las esperadas. Representaciones como la mostrada en la figura 2.7, son muy útiles para identificar las categorías que requieren mejora.

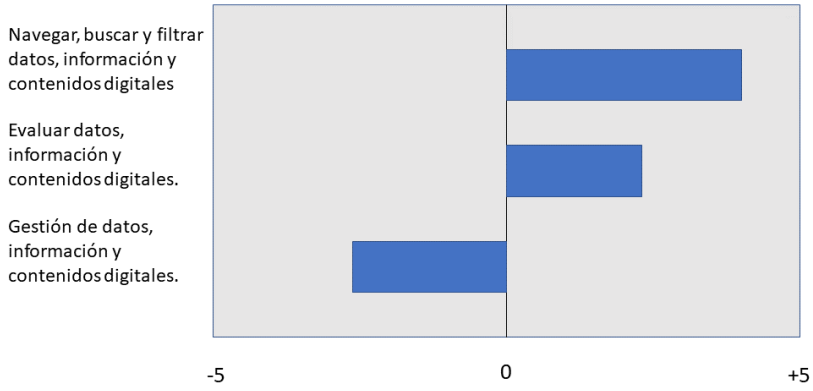


Figura 2.7. Ejemplo de perfil de desviación de un examinado

Hasta ahora, los modelos utilizados para la evaluación adaptativa han sido principalmente sumativos, abarcando un amplio rango de áreas de evaluación y midiendo o clasificando a los examinados, pero sin proporcionar más información como, p. ej. la prueba de nivel de inglés de Cúri y Silva (2019), la prueba para medir la capacidad de pensamiento crítico en física de Abidin et al. (2019), la prueba para evaluar las habilidades de resolución de problemas de física de Istiyono et al. (2018), la prueba para evaluar la personalidad empresarial de Postigo et al. (2020) o la prueba de personalidad de Moore et al. (2018). Cabe destacar que, hoy en día, no tenemos constancia de que se haya diseñado un TAI en el contexto de la evaluación de la CD.

En la tabla 2.20 puede verse los modelos de evaluación adaptativa identificados y analizados, así como sus principales características.

Modelo de evaluación adaptativa	Característica (Referencia/s)
Diagnósticos cognitivos	Combinación de la eficacia de los TAI con los resultados de DC (Cheng, Y. ,2009).
	Diseño de una matriz binaria, o q-matrix, para especificar para cada ítem cuales son los CC requeridos para poder resolverlo (Huebner, 2010).
	modelos basados en q-matrix más utilizados: DINA (De La Torre, 2011) y el Modelo de

	Jerarquía de Atributos (Leighton et al. 2004).
	Como los atributos en los DC son discretos, el estadístico de Fisher no puede ser utilizado (Lehmann y Casella, 2006). Han surgido métodos alternativos (Kaplan et al., 2015): KL (Chang y Ying, 1996; Cheng y Chang, 2007) o GDI (De la Torre y Chiu, 2016).
Evaluación sumativa mediante la TRI	Uso de la TRI en la implementación de los TAI muy extendido (Bernardi y Di Mascio, 2019; Chalmers, 2016; Gibbons et al., 2016; Magis et al., 2017; Oppl et al., 2017; Reckase, 1974; Van der Linden y Glas, 2010; Wainer et al., 2000; Walker et al., 2010).
	El modelo más sencillo utilizado para los TAI es el modelo de Rasch (Desmarais y Baker, 2012; Rasch, 1993).
	Control del nivel de error de medición manipulando la función de información de la prueba (Zheng et al, 2014).
	La TRI proporciona una gran variedad de algoritmos de selección de ítems y procedimientos de puntuación (Chen y Wang, 2010).
	Arquitectura flexible para diseñar TAI con ítems con distintos formatos Oppl et al. (2017), integrando la biblioteca catR (Magis y Raîche, 2012).
	Mejor adoptando un modelo TRI unidimensional (Chalmers, 2016; Thompson y Weiss, 2011; Van der Linden y Glas 2010; Weiss, 2004), ya que los multidimensionales son más dificiles de calibrar (Desmarais y Baker, 2012).
	Los modelos utilizados han sido principalmente sumativos: (Abidin et al., 2019; Cúri y Silva, 2019; Istiyono et al., 2018; Moore et al., 2018; Postigo et al., 2020).
	Hoy en día, no tenemos constancia de que se haya diseñado un TAI en el contexto de la evaluación de la CD, salvo PIX y TOSA, los cuales no ofrecen información al respecto.

Tabla 2.20. Principales características de los modelos de evaluación adaptativa seleccionados.

2.4.4. Características y ventajas de los test adaptativos

Durante las últimas décadas, numerosos autores han destacado las ventajas de los test adaptativos frente a los test convencionales, ya que reducen significativamente el tiempo de las pruebas sin perder precisión, lo que significa que las evaluaciones son más rápidas y precisas (p. ej. Abad et al., 2011; Hambleton et al., 1991; Muñiz, 2018; Rezaie y Golshan, 2015; Wainer et al., 2000; Wu et al., 2017). Además, cuando el diseño de los test adaptativos se apoya en el uso de la tecnología, esta aporta ventajas adicionales a parte del ahorro en papel, como las listadas a continuación:

- Mejora de la aplicación del test, reduciendo el tiempo de aplicación y las posibilidades de copia, así como homogenizando la aplicación del test para todos los participantes (Hambleton et al., 1991).

- Nuevos formatos de ítems, como p. ej. ítems dinámicos y simulaciones interactivas, que permitan RP más complejos que los ofrecidos por los formatos más tradicionales como las preguntas de opción múltiple. Además, posibilitan obtener información sobre el comportamiento de los participantes, gracias a que es posible registrar las interacciones que se han llevado a cabo como, p. ej. los clics de ratón (Greiff et al., 2015; Osborne et al., 2013; Timmis et al., 2016). Más aún, las respuestas se pueden procesar automáticamente mostrando el resultado de su respuesta y proporcionando feedback de manera inmediata (Magis y Raîche, 2012). En los TMEs, incluso los participantes pueden realizar la revisión de sus respuestas a los ítems a nivel de módulos.
- Mejora el control sobre la administración de las pruebas, posibilitando configurar los modos de revisión de los ítems, el orden de entrega de los ítems, así como el tiempo límite de respuesta para toda la prueba.
- Posibilita la administración de pruebas a personas con discapacidad, gracias a la posibilidad de integrar distintos dispositivos de entrada e interfaces, p. ej. adaptando el tamaño de letra y el contraste para personas con problemas de visión (Garrison y Baumgarten, 1986), o introduciendo cambios en el diseño para mejorar la accesibilidad a personas con discapacidad (Hansen et al., 2005).

A su vez, los TAI por el mero hecho de contar con un banco de ítems calibrado y su naturaleza adaptativa, ofrecen una serie de ventajas como:

- Incrementar la seguridad de las pruebas, ya que no todos los participantes van a ver las mismas preguntas, dificultando que estos puedan resolverlas por reconocimiento y memoria.
- Mejorar la calidad de los ítems, ya que, en el proceso de calibración del banco de ítems, se desechan los ítems que no cumplan los requisitos básicos de la TRI.
- Mayor reducción del tiempo de aplicación de la prueba, ya que un TAI administrando la mitad o un tercio de los ítems de una prueba convencional, es capaz de estimar la habilidad de un participante con una determinada precisión (Lan et al. ,2014; Thompson y Weiss, 2011; Weiss, 2004). Este hecho, ayuda a reducir la fatiga y la frustración de los participantes durante la realización de la prueba (Lynch y Howlin, 2014). En el caso de los TME la longitud suele ser ligeramente superior a la de los TAI, pero inferior a la de las pruebas convencionales (Magis et al, 2017).
- Estimaciones más precisas que una prueba lineal en condiciones similares con limitaciones de tiempo requerido y número de ítems a presentar. Esto se debe a que los TAI seleccionan en cada momento el ítem que más información va a darnos respecto a la habilidad estimada para un participante dado (módulos en el caso de un TME).

Sin embargo, a pesar de que el uso de TAI ha sido aceptado en ámbitos educativos, hasta hace poco los diseñadores de TAI no contaban con software flexible y de

código abierto que pudiesen utilizar para realizar estudios de simulación de los TAI y sus distintas configuraciones. Gracias al paquete catR desarrollado por Magis et al. (2017), los diseñadores de TAI han podido abordar esta carencia facilitando la implementación de test adaptativos. En esta tesis usamos el paquete catR para llevar a cabo distintas simulaciones que nos permitieron analizar distintas configuraciones, compilando resultados de las pruebas de forma inmediata, sin generar inconvenientes en los participantes.

En la tabla 2.21 puede verse las principales referencias identificadas y analizadas relacionadas con las características y ventajas de los test adaptativos.

Modelo de evaluación adaptativa	Referencia
Reducción del tiempo de las pruebas sin perder precisión, ofreciendo evaluaciones más rápidas y precisas.	(Abad et al., 2011; Hambleton et al., 1991; Lan et al. ,2014; Muñiz, 2018; Rezaie y Golshan, 2015; Thompson y Weiss, 2011; Wainer et al., 2000; Weiss, 2004; Wu et al., 2017)
Reducción del tiempo de aplicación y las posibilidades de copia. Homogenización de la aplicación del test para todos los participantes.	(Hambleton et al., 1991)
Permiten incorporar formatos de items con RP mas complejos y es posible registrar las interacciones que se han llevado a cabo.	(Greiff et al., 2015; Osborne et al., 2013; Timmis et al., 2016)
Control sobre la administración de las pruebas, posibilitando configurar el orden de entrega de los ítems, así como su revisión.	(Magis y Raîche, 2012)
Posibilita la administración de pruebas a personas con discapacidad.	(Garrison y Baumgarten, 1986; Hansen et al., 2005)
Reducción de la fatiga y la frustración de los participantes durante la realización de la prueba.	(Lynch y Howlin, 2014)
Longitud de los TME suele ser ligeramente superior a la de los TAI, pero inferior a la de las pruebas convencionales.	(Magis et al, 2017)
Posibilidad de realizar estudios de simulación con distintas configuraciones.	(Magis et al., 2017)

Tabla 2.21. Características y ventajas de los test adaptativos.

2.4.5. Puntos que considerar antes de aplicar un test adaptativo

A la hora de desarrollar un TAI, es necesario evaluar una serie de puntos clave, ya que puede resultar que aplicar un enfoque adaptativo no sea el más adecuado. Por

mencionar algunos puntos, es necesario comprobar si el banco de ítems tiene el suficiente tamaño para un número previsto de participantes, o si la longitud de la prueba es la apropiada para lograr el grado de fiabilidad buscado. Además, se deben de dar ciertas circunstancias para que sea aconsejable su uso y se puedan aprovechar las bondades de un enfoque adaptativo (Wainer et al. ,2000). Por ejemplo, si nos encontramos con que la habilidad a medir es muy difícil de evaluar sin un ordenador, o si queremos obtener el nivel de habilidad exacto del participante más allá de saber si es apto o no de acuerdo con cierto umbral de corte establecido. O, por el contrario, se pueden dar ciertas características que al analizarlas puedan hacernos declinar su uso:

- El banco de ítems debería incluir ítems para todo el rango de niveles de dificultad que se desea evaluar, desde ítems fáciles a difíciles. De esta manera, será posible realizar una estimación precisa de toda la gama de niveles de habilidad. Por lo tanto, será necesario contar con un banco de ítems calibrado con un gran número de ítems (Millman y Arter, 1984), lo cual previamente requerirá contar con suficientes muestras de respuestas para poder calibrar correctamente los ítems según el modelo de la TRI escogido. Normalmente, la recolección de muestras es un proceso que requiere un gran esfuerzo, hecho por el cual aplicar un enfoque adaptativo es a menudo es descartado. Incluso, puede resultar necesario garantizar la administración de ciertas proporciones de contenidos en las pruebas, lo cual complica bastante el diseño del TAI.
- Si el objetivo es proporcionar a los participantes un diagnóstico al final de la prueba, los TAI proporcionan una información diagnóstica escasa. Tradicionalmente, los modelos utilizados que siguen un enfoque adaptativo han sido principalmente sumativos, es decir, evalúan o clasifican a los examinados, pero a nivel cualitativo no proporcionan ninguna otra retroalimentación.

Además, habrá que tener presente la gestión del TAI en sí misma, que incluye el equilibrio de los contenidos, el análisis de los ítems, la puntuación de los ítems, el establecimiento de normas, el análisis de las prácticas y la actualización del banco de ítems. Llevar a cabo estudios de simulación pueden ayudar a evaluar las cuestiones previamente listadas, así como apoyar la toma de decisiones (Magis y Von Davier, 2017).

2.4.6. Diseño e implementación del TAI

Para empezar a diseñar la prueba, primero hay que aclarar una serie de puntos clave como la finalidad de esta (lo cual puede suponer la elección de distintos diseños para la prueba), el diseño del banco de ítems y su posterior mantenimiento, el equilibrado de contenidos (si se decide llevar a cabo el mismo), el ensamblado del test, la puntuación y la equiparación de puntuaciones, la fiabilidad y la validez, la seguridad del test y el control de la exposición.

Si el objetivo es diseñar una prueba de aptitud, el diseño del TAI deberá centrarse en la precisión de la estimación de la capacidad para un rango de niveles de aptitud de los participantes. Si el objetivo es diseñar una prueba de clasificación, el diseño del TAI deberá centrarse en la precisión de la clasificación de los examinados en sus grupos correspondientes. Para ambos objetivos, adoptar un enfoque basado en la TRI favorecerá su implementación.

Además, será necesario examinar una serie de cuestiones que son clave en el proceso del diseño del TAI: ¿Qué duración en términos de tiempo debe tener la prueba?, ¿Cuántos ítems se utilizarán (mínimo y máximo) ?, ¿Cómo se llevará a cabo la elección del primer ítem?, ¿Qué algoritmo se aplicará para mostrar el siguiente ítem?, ¿Qué modelo se empleará para la estimación de la capacidad?, ¿Cómo se establecerá el criterio de parada?, ¿Cómo se llevará a cabo la calificación de la prueba?, ¿Se van a establecer requisitos de balanceo de contenidos y cómo?, y por último, ¿Cómo se llevará a cabo el control de la exposición de los ítems?. A estas preguntas se pretende profundizar en las siguientes secciones de este capítulo con el objetivo de poder darles respuesta de la mejor manera posible.

2.4.7. Proceso de aplicación de un TAI

Todas las pruebas son administradas siguiendo unas reglas que indican el orden en que se deben mostrar los ítems al participante. Existe una gran variedad algoritmos para llevar a cabo esta tarea y todos tienen en común el objetivo de seleccionar el siguiente ítem a mostrar (Wainer et al., 2000). Existen claramente 4 situaciones clave a destacar según Magis y Raîche (2012):

- Identificar el primer ítem a administrar (también conocida como *etapa inicial*).
- Una vez evaluada la respuesta, identificar el siguiente a administrar, recalculando la estimación del nivel de capacidad recalcula después administrar de cada ítem.
- Establecer las condiciones de parada, momento en el que se dejará de mostrar más ítems, y se dará la prueba por finalizada.
- Por último, mostrar la estimación final de la capacidad del examinando.

La figura 2.8 muestra el diseño general de un TAI basado en el propuesto por Weiss y Kingsbury (1984), que será explicado en detalle en las siguientes secciones. Para seleccionar el primer ítem a administrar, será necesario disponer de una estimación inicial de la habilidad del participante. Al no contar con información obtenida de respuestas anteriores, habitualmente suele seleccionarse un ítem al azar o elegirse ítems de dificultad media (Veldkamp y Matteucci, 2013). A continuación, se aplicará el procedimiento que nos permitirá seleccionar el ítem que mejor se ajuste a la habilidad dada. Por cada respuesta, se irá recalculando la habilidad estimada del participante, así como la precisión de la medida obtenida. De nuevo, el algoritmo deberá seleccionar y mostrar el ítem que mejor se adapte a la nueva estimación de la habilidad, que normalmente consiste en mostrar el siguiente ítem un poco más

difícil si ha acertado, o un poco más fácil si ha fallado (Hambleton et al., 1991). Este proceso se repetirá hasta que se cumplan los criterios de parada que se hayan definido y que indicarán cuando se tiene que finalizar la administración de ítems (Veldkamp y Matteucci, 2013). Por ejemplo, cuando se trata de pruebas de longitud fija en las que se administra un número fijo de ítems al examinado, el número de ítems a mostrar se suele utilizar como criterio de parada. En cambio, cuando las pruebas son de longitud variable, se suelen utilizar distintos criterios de finalización como, p. ej. cuando la prueba alcance un determinado nivel de precisión de la medición (Segall, 2004), o cuando la prueba alcanza un tiempo máximo establecido de inicio.

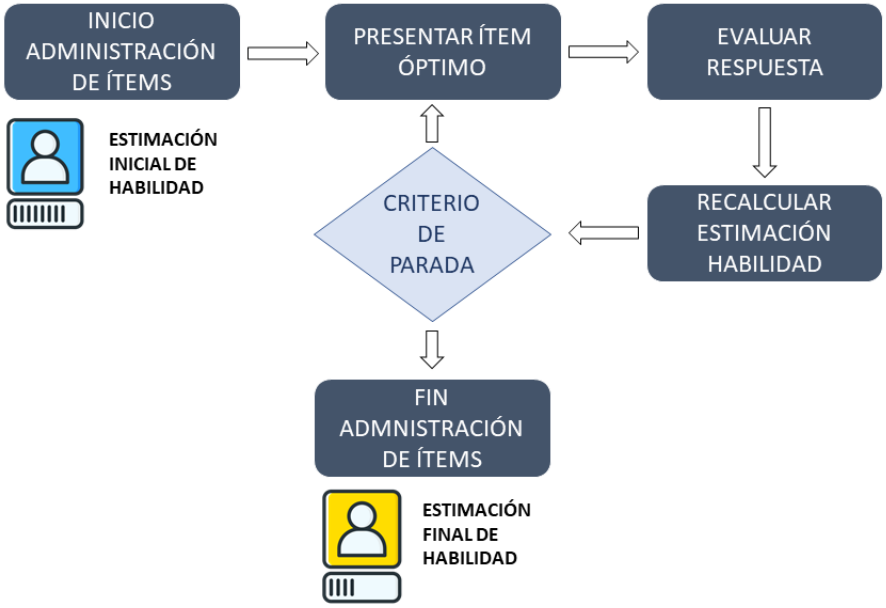


Figura 2.8. Diseño general de un TAI (Weiss y Kingsbury, 1984)

Procedimientos de arranque

El TAI para comenzar a administrar los ítems, obtiene el primer ítem a mostrar en base a una estimación inicial de la habilidad del participante. Con lo cual, cuanta más información dispongamos del participante al inicio, mejor podremos realizar esa estimación inicial, y consecuentemente, antes llegaremos a identificar la habilidad real del participante. Además, como se ha visto, la selección del nivel de dificultad del ítem inicial es clave para mejorar la eficiencia del TAI (Weiss, 1985). Por otra parte, cabe destacar que en algunos casos una elección desafortunada del primer ítem apenas tendrá efectos en la estimación final de la habilidad, siempre y cuando el TAI tenga al menos 20 ítems (Lord, 1980; van der Linden y Pashley, 2000).

De hecho, numerosos autores han buscado distintas opciones con el objetivo de complementar la información de partida para mejorar la estimación inicial. Por ejemplo, Kingsbury y Houser, (1993) hicieron uso de evaluaciones previas del

alumno en la misma prueba, para realizar una estimación de la habilidad mediante regresión estadística, o Sumbling et al. (2007) los cuales utilizaron la estimación obtenida en el primero de los subtest que lo componen para calcular la habilidad inicial en el resto de los subtest.

Una práctica muy extendida cuando no se dispone de ningún tipo de información previa del participante consiste en realizar una pequeña prueba de acceso. De esta forma, es posible obtener una primera estimación de habilidad antes de comenzar con la administración del TAI. Otra alternativa muy utilizada consiste en que el propio participante elija su nivel de habilidad. Se trata de una solución tecnológicamente sencilla, pero la mayoría de los participantes tienden a juzgar mal sus capacidades debido al sesgo positivo intrínseco causado en parte por el exceso de confianza de los propios participantes (Kruger y Dunning, 1999).

Otra alternativa muy común consiste en seleccionar un único primer ítem, seleccionando el más informativo en torno al valor medio de la capacidad del colectivo (casi siempre fijado en cero). El mayor inconveniente es que ese ítem pueda acabar sobreexpuesto. Pero este enfoque común puede mejorarse de muchas maneras como, p. ej.: seleccionando un ítem perteneciente a un intervalo de dificultad determinado (Revuelta et al., 2003), seleccionando un ítem moderadamente fácil para que el participante comience acertando los primeros ítems y así potenciar su motivación (Verschoor y Straetmans, 1999), o seleccionando un ítem cercano al punto de corte (apto y no apto) en la escala de habilidades (Lunz y Bergstrom, 1994). También se han aplicado alternativas más específicas como, p. ej. combinar la TRI con métodos estadísticos bayesianos (Baker y Kim, 2004).

¿Cómo continuar?

Una vez evaluada la respuesta del participante al primer ítem administrado, el TAI deberá estimar de nuevo su habilidad, así como calcular la precisión asociada a dicha estimación (que inicialmente será muy baja). Entonces, el algoritmo de aplicación del TAI deberá seleccionar el siguiente ítem a administrar de acuerdo con el nivel de dificultad apropiado para el nivel estimado. Una vez que el participante responda al nuevo ítem administrado, se volverá a estimar de nuevo el nivel de habilidad y a calcular la precisión con que se ha estimado (que debería de ser mayor), y así sucesivamente hasta que se cumplan las condiciones del criterio de parada. Según el TAI vaya administrando más ítems dentro del test, la precisión de estimación de la habilidad del participante irá aumentando. Por consiguiente, los principales aspectos técnicos implicados son la estimación provisional para el patrón de respuesta actual y el conjunto de ítems administrados, y la selección del siguiente ítem que sea óptimo.

A continuación, describimos los métodos más utilizados para volver a estimar la habilidad del participante después de cada respuesta. Describimos las estrategias

más utilizadas, y a otras cuestiones importantes a tener en cuenta a la hora de continuar, como es el caso del control de la exposición de los ítems.

El TAI por cada respuesta recibida, debe evaluarla y volver a estimar la habilidad del participante para ir ajustándola con mayor precisión. Para llevar a cabo esta acción, se basa en un procedimiento de estimación del nivel de habilidad basado en las respuestas proporcionadas a los ítems que han sido administrados hasta ese momento, para decidir el valor de la escala de habilidad que mejor se ajusta al patrón de respuesta observado. Existen varios estimadores de habilidad, siendo los más populares el estimador ML y los estimadores bayesianos (Magis et al., 2017; Magis y Raîche, 2012; Mislevy, 1986; Swaminathan y Gifford, 1985; van der Linden y Glas, 2010).

El procedimiento de estimación más básico y utilizado es ML (Lord, 1980). Conocidos los parámetros de los ítems, el nivel de habilidad del participante puede ser estimado mediante la *función de verosimilitud*, la cual es el producto de todas las funciones de respuesta al ítem, y se basa generalmente en el patrón de respuesta observado. Principalmente se busca encontrar el nivel de habilidad que maximiza la probabilidad del patrón de respuestas observado. Es decir, si un examinando ha acertado los ítems de dificultad menor o igual que cierto θ, y fallado los que son más difíciles, una primera aproximación dirá que su nivel de habilidad ronda dicho valor θ. Además, ML es recomendable para su uso en pruebas de aptitud, o para realizar comparativas entre grupos de participantes o entre diferentes tipos de pruebas como, p. ej. TAI y pruebas lineales (Abad, Olea et al., 2002). Sin embargo, ML plantea un grave problema cuando se han administrado pocos ítems, y el examinando tiene un patrón constante de respuestas. Así, no es posible acotar el valor de habilidad que hace más creíble que el participante haya acertado (o fallado) todos los ítems planteados, porque la situación más verosímil en este caso es la de que el participante lo sabe absolutamente todo (o no sabe absolutamente nada).

Como alternativa, se propusieron los métodos bayesianos que se basan en la aplicación de la regla de Bayes (1763), que combinan la evidencia reflejada en los patrones de respuesta con el conocimiento previo de la distribución de probabilidad del rasgo latente de los participantes (Bock y Mislevy, 1982; Samejima, 1969). Los métodos bayesianos se basan en dos tipos de distribuciones: la de la probabilidad previa, que es la que se asigna antes de contar con alguna evidencia, y la de probabilidad posterior, que es la que se asigna después de obtener algún resultado. Los procedimientos de estimación bayesianos añaden a la *función de verosimilitud* cierta información acerca de la distribución a priori de la habilidad de la población como, p. ej. usando información de previas evaluaciones a grupos similares (Birnbaum, 1969; Owen, 1975; Weiss, 1982). La *estimación bayesiana Esperada a Posteriori* (EAP) coincide con la media de la distribución de probabilidad posterior de θ (Bock y Mislevy, 1982), mientras que la *estimación bayesiana Máxima A Posteriori* (MAP), se corresponde con la moda de dicha distribución (Lord, 1986). Más aún, Wang y Vispoel (1998) llevaron a cabo un estudio de simulación donde

compararon una selección de métodos de estimación, y encontraron que el método EAP era el que mejor funcionaba entre los métodos bayesianos. Van der Linden y Pashley (2000) mostraron como con más de veinte o treinta ítems apenas habrá diferencias con respecto al ML. Sin embargo, los métodos bayesianos también plantean una serie de problemas, como que la estimación del nivel de habilidad no depende únicamente del rendimiento mostrado en la prueba por el participante, porque los valores estimados también sufren la influencia de los estadísticos (media y varianza) asignados a la distribución a priori del rasgo de toda la población. Otro inconveniente de los métodos bayesianos es que el sesgo en las estimaciones proporcionadas por los métodos bayesianos es ligeramente mayor que el producido por el ML, si bien el error estándar es en la mayoría de los casos menor (Wang y Vispoel, 1998). De ahí que autores como, p. ej. Raîche et al. (2007) hayan presentado nuevas versiones para reducir el sesgo asociado a las estimaciones obtenidas, ajustando la distribución a priori tras la administración de cada ítem.

Una vez estimado el nivel de habilidad provisional del participante en base a las respuestas dadas hasta ese momento, se trata de elegir el siguiente ítem a administrar escogiendo el ítem más informativo entre los restantes. Los algoritmos de selección más habituales son el *criterio de máxima información* y la *selección bayesiana* (Kingsbury y Zara, 1989; Van der Linden y Hambleton, 1997), siendo los primeros los más utilizados por su simplicidad y facilidad de cálculo. Mientras que el *criterio de máxima información* permite a un TAI seleccionar un ítem con la máxima información en el nivel de habilidad actual del participante, es decir, el que minimice el error estándar y maximice la precisión para la estimación actual de la habilidad del participante (Lord, 1980; Weiss, 1982), la selección bayesiana se utiliza para seleccionar el ítem que minimiza la varianza posterior esperada de las estimaciones de habilidad de la capacidad. Sin embargo, el *criterio de máxima información* lo que hace es elegir constantemente los ítems con mayor poder de discriminación, afectando a la seguridad de la prueba sobrexponiéndolos o que no sólo se usen los más discriminativos, siendo las más extendidas: las variantes basadas en la función de información global de *Kullback-Leibler* (Chang y Ying, 1996; Cheng y Chang, 2007), el criterio de información ponderada basado en su función de verosimilitud (Veerkamp y Berger, 1997), y por último, la *función de información de Fisher con distribución posterior* (van der Linden, 1995).

La principal alternativa se basa en los *criterios de selección bayesianos*, que se basan en actualizar cada vez que el examinado responde un ítem, la distribución posterior de θ, la estimación actual de la habilidad del participante, el valor de la información observada en dicho valor (no sólo para el ítem actual, sino también para todas las respuestas anteriores), y la varianza posterior de θ (van der Linden y Pashley, 2000). Además, es posible combinarlos con un procedimiento de estimación MAP o EAP (van der Linden, 1998; van der Linden y Pashley, 2000), siendo el primero la alternativa bayesiana más utilizada, la cual consiste en elegir en cada paso el ítem que proporciona la menor varianza de la distribución a posteriori del nivel de habilidad.

Respecto a otro tipo de restricciones a considerar, algunos TAI incluyen en el algoritmo de administración del test una serie de controles adicionales como, p. ej. para ver que ítems se están sobrexponiendo y cuales están siendo infrautilizados, o para garantizar la seguridad del banco de ítems. Una opción habitual es seleccionar un ítem al azar entre los ítems más informativos, como el *método randomesque* (Green, Bock et al., 1984; Kingsbury y Zara, 1989). Este método selecciona un ítem al azar entre los cinco más informativos en cada momento, destacando por su facilidad de implementación, sin pérdidas apreciables en la precisión del TAI. El enfoque más común es imponer una tasa de exposición máxima que ningún ítem debe superar (Han, 2018; Revuelta y Ponsoda, 1998; Sympson y Hetter, 1985; van der Linden y Veldkamp, 2004), siendo los más utilizados: el método Sympson-Hetter (Sympson y Hetter, 1985), el método restringido (Revuelta y Ponsoda, 1998a), el método multinomial incondicional (Stocking y Lewis, 1998, 2000), o el método multinomial condicional (Stocking y Lewis, 1998). Decantarse por un método en concreto no es tarea fácil, y la decisión dependerá de factores como el propósito de la prueba, los requisitos de seguridad o las características de los ítems, además de la tasa de exposición de cada ítem, la proporción de participantes a los que se administra la prueba y el periodo de tiempo durante el que ocurre la exposición. Además, si no se sigue estrictamente el procedimiento de selección (máximo informativo o bayesiano), supondrá que los ítems administrados no siempre serán los teóricamente óptimos, y podrá verse afectada la precisión de las estimaciones. Con lo cual, deberá existir un compromiso entre la eficiencia de la prueba, su seguridad y la precisión de la medida.

A parte del control de la exposición de los ítems, existen otro tipo de restricciones adicionales como, p. ej. asegurar que la prueba abarca todos los objetivos de aprendizaje con los porcentajes deseados, haciendo uso de un balanceo de contenidos. De hecho, si no se controla, puede influir negativamente en la validez interna del test. Por ejemplo, un TAI evaluación de CD centrado en el AC de IAD, podría administrar los ítems manteniendo una proporción preestablecida de contenido, con un 33% destinado a cada una de las 3 CD que componen esta AC. Asimismo, se puede dar el caso de que los ítems más informativos correspondan todos al mismo objetivo de aprendizaje. El uso de *TAI restringidos* (Kingsbury y Zara, 1989) o los *TAI restringidos con Shadow test* (van der Linden, 2000), pueden ser una solución, ya que dividen el banco de ítems en áreas de contenido de manera que se tienen en cuenta las especificaciones sobre la proporción de ítems de cada partición.

Procedimientos de parada, ¿Cuándo hay que parar?

El TAI continuará seleccionando nuevos ítems y reajustando la habilidad estimada hasta que se cumplan los criterios de parada. La elección de la regla de parada a aplicar en el TAI puede variar según la finalidad de este o las características del banco de ítems, entre otros factores. De las distintas alternativas disponibles a

elegir, las más habituales son aplicar los criterios de longitud fija y de longitud variable (Thissen y Mislevy, 2000). El criterio de longitud fija establece que la prueba terminará cuando se haya utilizado un número prefijado de ítems, y de esta forma el tiempo empleado en la prueba será similar para todos los participantes. En este cao será clave determinar el número de ítems a mostrar ya que probablemente la precisión en los resultados será diferente para cada examinando. Por ejemplo, McBride, Wetzel et al. (1997) demostraron que con 30 ítems las estimaciones son por lo general suficientemente rigurosas. En cambio, el criterio de longitud variable establece que la administración de ítems finalice cuando se alcance una precisión concreta de la estimación de la habilidad.

También es posible establecer un tiempo límite como criterio de parada, pero es necesario establecer tiempos límites lo suficientemente amplios como para que los participantes lentos en responder no sufran una presión adicional. También es posible mezclar diferentes criterios como, p. ej. finalizar la prueba si después de administrar un determinado número de ítems no se ha alcanzado la precisión buscada (Sumbling et al., 2007). En el caso concreto de las pruebas de aptitud, donde los participantes tienen que superar un punto de corte previamente establecido en la escala de habilidades, es habitual establecer como criterio de parada finalizar cuando la probabilidad de asignar al participante en el perfil correcto alcanza un valor aceptable.

Otros criterios ampliamente utilizados son, que el TAI se detenga cuando el nivel de habilidad alcance un nivel de nivel de precisión concreto, que el TAI se detenga cuando el participante alcance un nivel de habilidad mayor o menor que el nivel de capacidad que permita clasificarlo en el dominio de la habilidad, o que el TAI se detenga cuando los restantes ítems elegibles no proporcionen suficiente información para que la información total aumente significativamente (van der Linden y Glas, 2010).

La puntuación en los TAI

Una vez detenida la administración de los ítems, el TAI deberá mostrar la estimación final del nivel de capacidad del participante, basándose en las distintas respuestas recogidas durante la administración del TAI. La estimación de la capacidad se va recalculando durante la realización de la prueba y al final de la administración se calcula la estimación final. Al finalizar la administración, también es posible mostrar el valor final del error estándar y el intervalo de confianza para la capacidad correspondiente. Más aún, de cara a interpretar los resultados de un TAI, basarse en la proporción de aciertos no es un buen indicador de la habilidad de los participantes, y mostrarles el valor numérico correspondiente su estimación de la habilidad en la escala de medida del banco de ítems, no resulta muy informativo para el examinando. Lo que se hace, es inferir otro tipo de puntuación a partir de esa estimación que sea más significativa y valiosa para los participantes. Las prácticas más habituales son: trasladar los valores estimados de la habilidad del participante a otra escala más entendible por el participante, p. ej. acotando los valores al intervalo [0,10], definir una tabla de deciles o centiles, con el objetivo de

indicar al examinado el porcentaje de participantes que han obtenido una puntuación menor, igual o mayor que la suya, o transformar el valor de habilidad estimado para el participante a la métrica de la curva característica del test utilizado (Hambleton y Swaminathan, 1985).

Cabe mencionar que cuando no se ha completado la prueba en el tiempo límite permitido, o se ha diseñado el TAI para permitir la omisión de respuestas, pueden surgir problemas de sesgo en la puntuación prueba. Concretamente, si la cantidad de ítems administrados es pequeña el método de estimación ML puede asignar puntuaciones extremas, mientras que los métodos bayesianos tienden a proporcionar estimaciones muy cercanas a la media de la distribución previa.

En la tabla 2.22 puede verse las principales referencias identificadas y analizadas respecto al proceso de aplicación de un TAI (Magis y Raîche, 2012; Weiss y Kingsbury, 1984).

Situaciones clave en la aplicación de un TAI	Característica (Referencia)
Procedimientos de arranque	Selección de un ítem al azar o ítems de dificultad media (Veldkamp y Matteucci, 2013).
	Nivel de dificultad del ítem inicial es clave para mejorar la eficiencia del TAI (Weiss, 1985).
	Una mala elección del primer ítem apenas tendrá efectos en la estimación final de la habilidad, para TAI con al menos 20 ítems (Lord, 1980; van der Linden y Pashley, 2000).
	Complementar la información de partida para mejorar la estimación inicial, p. ej., evaluaciones previas (Fang et al., 2017) y estimación de la habilidad con regresión estadística (Kingsbury y Houser, 1993), que los propios usuarios indiquen el nivel de habilidad que creen tener (Kruger y Dunning, 1999), o utilizar la estimación del primero de los subtest para calcular la habilidad inicial en el resto de los subtest (Sumbling et al., 2007).
	Alternativas a una de las prácticas más habituales, la de seleccionar un primer ítem en torno al valor medio de la capacidad del colectivo, seleccionando: un ítem perteneciente a un intervalo de dificultad concreto (Revuelta et al., 2003), un ítem muy fácil para potenciar la motivación (Verschoor y Straetmans, 1999), un ítem cercano al punto de corte (Lunz y Bergstrom, 1994), o alternativas más específicas como, p. ej. combinar la TRI con métodos estadísticos bayesianos (Baker y Kim, 2004).
¿Cómo continuar?	Normalmente se muestra un ítem un poco más difícil si ha acertado, o un poco más fácil si ha fallado (Hambleton et al., 1991).
	Estimadores de habilidad más populares: el ML y los bayesianos (Bock y Mislevy, 1982; Lord, 1980; Magis et al., 2017; Magis y Raîche, 2012; Mislevy, 1986; Samejima, 1969;

	Swaminathan y Gifford, 1985; van der Linden y Glas, 2010).
	ML es recomendable para realizar comparativas entre diferentes tipos de pruebas como TAI y pruebas lineales (Abad, Olea et al., 2002).
	Los bayesianos añaden a la función de verosimilitud cierta información acerca de la distribución a priori de la habilidad de la población como, p. ej. usando información de previas evaluaciones a grupos similares (Birnbaum, 1969; Owen, 1975; Weiss, 1982).
	EAP coincide con la media de la distribución de probabilidad posterior de θ (Bock y Mislevy, 1982), y MAP corresponde con la moda de dicha distribución (Lord, 1986).
	EAP es el que mejor funciona entre los métodos bayesianos (Wang y Vispoel, 1998).
	Con más de veinte ítems apenas habrá diferencias entre EAP y ML (Van der Linden y Pashley, 2000).
	El sesgo en las estimaciones proporcionadas por los métodos bayesianos es ligeramente mayor que con ML, si bien el error estándar es en la mayoría de los casos menor (Raîche et al., 2007; Wang y Vispoel, 1998).
	Selección del siguiente ítem a administrar: criterio de máxima información y la selección bayesiana (Kingsbury y Zara, 1989; Lord, 1980; Van der Linden y Hambleton, 1997; Weiss, 1982).
	Variantes al criterio de máxima información: función de información global de Kullback-Leibler (Chang y Ying, 1996; Cheng y Chang, 2007), el criterio de información ponderada basado en su función de verosimilitud (Veerkamp y Berger, 1997), y la función de información de Fisher con distribución posterior (van der Linden, 1995)
	Criterios de selección bayesianos (van der Linden y Pashley, 2000), y posibles combinaciones con es posible combinarlos con MAP o EAP (van der Linden, 1998; van der Linden y Pashley, 2000)
	Selección de un ítem al azar entre los ítems más informativos, como el método randomesque (Green, Bock et al., 1984; Kingsbury y Zara, 1989)
	Tasa de exposición máxima que ningún ítem debe superar (Han, 2018; Revuelta y Ponsoda, 1998; Sympson y Hetter, 1985; van der Linden y Veldkamp, 2004). Métodos más utilizados: Sympson-Hetter (Sympson y Hetter, 1985), restringido (Revuelta y Ponsoda, 1998a), multinomial incondicional (Stocking y Lewis, 1998, 2000), o multinomial condicional (Stocking y Lewis, 1998)
	TAI restringidos (Kingsbury y Zara, 1989) y TAI restringidos con Shadow test (van der Linden, 2000), especificando proporciones de contenidos.
Procedimientos de parada, ¿Cuándo hay que parar?	Al cumplirse los criterios de parada definidos, finalizará la administración de ítems (Segall, 2004; Veldkamp y Matteucci, 2013).
	Más habituales, los criterios de longitud fija y de longitud variable (Thissen y Mislevy, 2000).

	Con criterios de longitud fija McBride, con 30 ítems las estimaciones son suficientemente rigurosas (Wetzel et al., 1997).
	Finalización de la prueba si después de administrar un determinado número de ítems no se ha alcanzado la precisión buscada (Sumbling et al., 2007).
	Otros criterios: cuando el nivel de habilidad alcance un nivel de nivel de precisión concreto, cuando el participante alcance un nivel de habilidad mayor o menor que el nivel de capacidad que permita clasificarlo, o cuando los restantes ítems no proporcionen suficiente información (van der Linden y Glas, 2010).
La puntuación en los TAI	Las prácticas más habituales son: trasladar las estimaciones de habilidad a otra escala más entendible, definir una tabla de deciles o centiles, o transformar el valor de habilidad estimado a la métrica de la curva característica del test (Hambleton y Swaminathan, 1985).

Tabla 2.22. Proceso de aplicación de un TAI.

2.4.8. Propiedades psicométricas del TAI

Hay una serie de factores claves que podrían afectar significativamente a las propiedades psicométricas del TAI y podrían ocasionar efectos negativos en la evaluación. Por ejemplo, permitir a los participantes revisar sus respuestas, ya que contribuye a disminuir los niveles de ansiedad del participante. O permitir las omisiones de respuesta a los ítems, que puede afectar tanto a la estimación de la habilidad del participante como a la selección del siguiente ítem. De hecho, la mayoría de los autores no recomiendan ambas opciones a pesar de que no poder revisar las respuestas a los ítems es un factor negativo (Green, Bock et al., 1984; Wainer, 2000b). Establecer un tiempo límite para llevar a cabo la prueba es otro factor clave para tener en cuenta, ya que automáticamente el banco de ítems pasa a ser multidimensional. Los ítems ya no sólo miden si el participante sabe resolverlas, sino si también es capaz de hacerlo en el tiempo establecido.

Fiabilidad

En la TRI, gracias a la función de información de la prueba contamos con procedimientos precisos para cumplir niveles específicos de objetividad. Dado que la función de información de la prueba puede obtenerse sumando los valores de las funciones de información de los ítems condicionados a la capacidad del participante (θ), muestra lo bien que un test mide a los participantes en cada valor de θ. Luego se trata de un índice de precisión local a nivel del test y es útil para garantizar la objetividad deseable de la prueba. Más aun, la precisión es analizada en base al error típico de las estimaciones, el cual es diferente para cada punto de la escala de habilidades con lo que la precisión de las estimaciones variará según la zona de la escala donde se sitúen.

Gracias al uso de las simulaciones, se puede examinar fácilmente la precisión teórica de un TAI simulando participantes con distintos niveles de habilidad y obteniendo sus estimaciones (Hambleton y Cook, 1983; Olea y Ponsoda, 1996). Por ejemplo, las simulaciones pueden servir para calcular el número de ítems que es necesario administrar para cada nivel de habilidad si queremos obtener una determinada precisión en un TAI de longitud variable. Una vez obtenida una estimación θ de su habilidad original, es entonces posible calcular un indicador de la magnitud del error total de la estimación, como es el índice RMSE (Wang y Vispoel, 1998).

Validez

A la hora de evaluar la validez de un TAI existen distintas alternativas, las cuales a su vez requieren distintos tipos de análisis de datos como, p. ej. el uso del análisis factorial, o de regresiones múltiples.

Para evaluar la validez interna o de contenido, es necesario comprobar que los ítems incluidos en la prueba representan adecuadamente el dominio de conocimiento que se pretende evaluar. Por lo tanto, será necesario incluir métodos de balanceo de contenidos en el método de selección de ítems utilizado.

Para evaluar la validez de constructo es necesario garantizar que la prueba estima únicamente el rasgo esperado. Para llevarlo a cabo, es necesario analizar varios factores como, p. ej. el algoritmo de aplicación del TAI empleado. Suele ser habitual realizar un análisis de la unidimensionalidad del banco de ítems, pero también es recomendable examinar el funcionamiento diferencial de los ítems para evitar, p. ej. que se observen diferencias entre las estimaciones entre distintos géneros, lo cual podría cuestionar la validez de constructo de la prueba. Es decir, si la probabilidad de acierto para niveles similares de habilidad difiere en dos grupos, se dice que ese ítem posee funcionamiento diferencial. El funcionamiento diferencial no supone que existan estimaciones injustas, sino que existe otro factor o dimensión que está influyendo en las estimaciones. Sólo por usar tecnología para la administración de la prueba ya puede causar diferencias en grupos con distinto nivel de CD. Además, en un TAI existe una dependencia entre la secuencia de ítems aplicados y las respuestas dadas a los ítems previos que también puede causar diferencias (Linden et al., 2000).

Para evaluar la validez externa es necesario examinar el potencial de la prueba para predecir el comportamiento futuro del participante. Este análisis puede realizarse comparando los resultados obtenidos en la prueba con resultados de otras pruebas convencionales, o factores asociados al éxito del rasgo evaluado.

Seguridad

La seguridad del banco de ítems depende de muchos factores en un TAI, siendo los más significativos la sobreexposición de los ítems y los ítems sin responder. Si, p. ej. algunos ítems son muy informativos en el nivel medio de habilidad, estos podrían

ser seleccionados con más frecuencia que otros ítems y acabar sobreexpuestos (Revuelta y Ponsoda, 1998). También es habitual las respuestas se vayan haciendo públicas y que los siguientes participantes obtengan falsos resultados. Para abordar estos riesgos, existen distintas metodologías para control la exposición de los ítems (Georgiadou et al., 2007) como ya hemos visto previamente. Respecto al control de los ítems sin respuesta, en el caso de que el TAI haya sido configurado para permitirlo, el participante puede memorizar ciertos ítems que deje sin responder, y una vez conocidas las respuestas, ser capaz de responderlos de nuevo correctamente, comprometiendo la seguridad de la prueba. Soluciones sencillas como, p. ej. presentar aleatoriamente las opciones de respuesta de los ítems puede ayudar a reducir estos riesgos.

2.4.9. Gestión del TAI

La gestión del TAI comprende todas las acciones a llevar a cabo tras al despliegue de la prueba para poder garantizar que las pruebas se lleven a cabo correctamente de acuerdo con las especificaciones previamente definidas (Seo, 2017). En concreto, la gestión del TAI incluye el equilibrio de los contenidos, el análisis y calificación de los ítems, el establecimiento de normas, el análisis de las prácticas, y la actualización del banco de ítems. Si bien quedan fuera del ámbito de esta tesis, es conveniente mencionarlo para ser tenido en cuenta por el esfuerzo que supone.

2.4.10. Consideraciones para este estudio en términos de TAI

Teniendo en cuenta los puntos fuertes de la teoría, para esta disertación utilizamos el modelo de medición de Rasch para investigar la fiabilidad y la validez de las pruebas desarrolladas en nuestro estudio. Dentro del contexto de la TRI, el modelo de medición de Rasch es el modelo más sencillo disponible y facilita la interpretación de las respuestas asumiendo que la respuesta de los participantes a un ítem sólo depende de su competencia y de la dificultad del ítem (Rasch, 1993).

Una de las razones por las cuales elegimos el modelo de Rasch fue debido a que los tamaños de muestra relativamente pequeños pueden ser suficientes, y según Wright y Stone (1979) alrededor de 200 examinados podrían ser suficientes para obtener estimaciones precisas de los parámetros, cantidades similares a las que pudimos recoger en nuestras pruebas.

Por otra parte, aunque la multidimensionalidad del constructo de la CD se ha identificado en varios estudios, los estudios teóricos y empíricos han informado de resultados contradictorios (Reichert et al., 2020). Más concretamente, Vuorikari et al. (2016) describieron teóricamente el constructo de la CD en DigComp, pero las dimensiones identificadas no se han confirmado empíricamente o requieren de más investigación. Como mostramos más adelante, nuestros resultados mostraron que el modelo tridimensional para la prueba de IDL y el modelo cuatridimensional para la prueba de netiqueta se ajustaban mejor a los datos que el modelo unidimensional. Sin embargo, teniendo en cuenta las altas correlaciones obtenidas, parece que todos los ítems se relacionan con un factor fuerte, que puede

interpretarse como CD general, en la misma línea que otros autores han apuntado en sus estudios recientes como, p. ej. Clifford et al. (2020).

Los TAI son un método de medición muy eficaz y flexible que se basa en modelos de TRI y tiene su origen en la psicometría y la evaluación educativa. Paralelamente a la investigación y el desarrollo relacionados con los TAI, el campo de la Inteligencia Artificial (IA), impulsado por las ciencias de la computación, también se ha desarrollado mucho en los últimos años, y concretamente en su uso para la evaluación y las pruebas adaptativas. Como ejemplo, cabe destacar la conferencia IACAT 2022 que se llevó a cabo Frankfurt organizada por la Asociación Internacional de Pruebas Adaptativas Informatizadas (IACAT). En dicha conferencia, se vincularon las dos áreas de forma interdisciplinar, con el objetivo de que ambas áreas aprendiesen la una de la otra.

A pesar de que el diseño de las TAI está, hoy en día, muy extendido a nivel mundial en la evaluación educativa y psicológica, las implementaciones que se han llevado a cabo en el área de la evaluación de la CD son prácticamente nulas. Considerando que en nuestro trabajo ya hemos introducido en la evaluación de la CD un factor muy diferenciador como es la evaluación de habilidades de orden cognitivo alto en un entorno controlado a través de distintos tipos de preguntas como las simulaciones interactivas o las preguntas basadas en imágenes o simulaciones, en nuestra última parte del trabajo nos centraremos en realizar los análisis preliminares para convertir ECTD (una prueba lineal y sumativa) en un TAI basado en la TRI, en la que los ítems se seleccionan gradualmente para el participante, de acuerdo con las respuestas que vaya dando y la estimación de su competencia provisional. Una de las principales ventajas de la TAI es que pueden reducir considerablemente la duración de la prueba sin afectar demasiado a la calidad de las habilidades estimadas. Para justificar la elección de un TAI, el objetivo es ilustrar este hecho comparando las pruebas lineales para ambas pruebas y sus correspondientes TAI, haciendo uso del mismo conjunto de datos simulados.

Al tratarse de un análisis preliminar, llevaremos a cabo el estudio con una configuración sencilla haciendo uso de métodos muy utilizados y extendidos. Como trabajo futuro, nos gustaría continuar evaluando la aplicación de distintas configuraciones y métodos en los distintos elementos clave del TAI, como puede ser el método de arranque, la selección del siguiente ítem o los criterios de parada.

2.5. Conclusiones

En el capítulo 2, hemos llevado a cabo una revisión de las principales áreas en las que hemos desarrollado nuestra investigación. Buscamos diseñar e implementar una herramienta de evaluación de CD, lo suficientemente flexible como para permitir distintos formatos de ítems, como las simulaciones interactivas, que nos posibiliten la evaluación de habilidades tanto de orden cognitivo bajo como de orden cognitivo alto. Tal y como Vuorikari et al. (2022) reflejan en DigComp, cada CD tiene sus propias peculiaridades y para evaluarlas, requerirán distintos formatos de ítems si queremos que los participantes muestren los comportamientos esperados. Revisamos las herramientas de evaluación de CD disponibles y

reflexionamos sobre qué aspectos deberíamos de tener en cuenta en nuestra investigación. Hemos podido constatar un notable y creciente número de herramientas de evaluación de la CD, pero la inmensa mayoría siguen arrastrando las mismas carencias.

Por una parte, la mayoría de las herramientas de evaluación identificadas en las revisiones literarias, no utilizan un marco de referencia común y no cubren todas las dimensiones de la CD. Razón por la cual optamos por realizar nuestra propia implementación haciendo uso de DigComp, el cual facilita el trabajo en el contexto de la CD proporcionando un lenguaje claro y común que posibilita su interpretación por las diferentes partes involucradas y su posterior aceptación. Además, la mayoría de ellas sólo evalúan habilidades de orden cognitivo bajo, probablemente debido al gran esfuerzo que supone la integración de ítems con formatos dinámicos en las pruebas. Más aún, de las pocas identificadas, la mayoría están basadas en software con licencia privativa y no contemplan los dispositivos móviles como medio para el desempeño de la CD. La TEA no sólo ofrece gran variedad oportunidades para mejorar la experiencia de los participantes, sino que también permite desarrollar modos de evaluación más ajustados a las necesidades actuales proporcionando formatos de ítems innovadores y más cercanos a la práctica real (Binkley et al., 2012; Cho et al., 2019; Debuse y Lawley, 2016; Drasgow, 2016; Scherer et al., 2017; Shute y Rahimi, 2017; Stödberg, 2012; Zenisky y Luecht, 2016). Usamos su potencial para implementar simulaciones interactivas, que son especialmente relevantes a la hora de medir constructos cognitivos complejos como la CD en entornos seguros, y otro tipo de formatos dinámicos que nos permitan presentar a los participantes situaciones lo más parecidas posibles a lo que se van a encontrar en la vida real, donde tengan que poner sus conocimientos en práctica para poder resolverlas. Es decir, adoptamos un enfoque basado en el rendimiento. En concreto, encontramos interesante aprovechar el potencial de soluciones como Articulate Storyline para diseñar simulaciones basadas en aplicaciones de todo tipo (con licencia y libres), y para plantear su resolución tanto en estaciones de trabajo como en dispositivos móviles, lo cual refleja más la situación actual. Además, la TEA posibilita la obtención de información sobre el comportamiento y el rendimiento de los examinados durante las pruebas, recogiendo información de diferentes formas como, p. ej. tiempos de respuesta, flujos de clics, etc. (Bartolomé y Garaizar, 2022; Greiff et al., 2015; Osborne et al., 2013; Timmis et al., 2016). En nuestra herramienta de evaluación necesitamos que permita el registro de grano fino de las interacciones de los participantes durante la realización de las pruebas para poder realizar analítica de la evaluación. Para llevarlo a cabo, dotamos a nuestra herramienta de estas capacidades a dos niveles: a nivel de herramienta, registrando los resultados y los tiempos de respuesta, y a nivel de simulaciones, registrando los pasos realizados, los clics erróneos, clics realizados y tiempo empleado. En un contexto en que la actividad del examinando puede ser registrada, es especialmente interesante y viable que la utilicemos para adaptar y mejorar el proceso de evaluación.

Por otra parte, tal y como identificaron en su revisión Siddiq et al. (2016), existe una falta de información coherente sobre los indicadores de fiabilidad y validez. Por

este motivo, lo que hacemos es aplicar una metodología de investigación basada en el diseño (DBR) que se basa en el análisis de diferentes fuentes de información para llevar a cabo el desarrollo de la herramienta de y su posterior validación. La DBR es una metodología muy utilizada en las ciencias del aprendizaje para analizar el desarrollo de soluciones (Herrington et al., 2007; McKenney y Reeves, 2018; Sandoval, 2014). Describimos los principios de diseño aplicados durante los diferentes pasos del desarrollo de las pruebas para evaluar las CD seleccionadas, con el fin de que puedan ser extendidos al resto de los CD incluidos en DigComp. Usamos esta metodología para desarrollar y validar la herramienta de evaluación combinando diferentes estudios que se nutrieron de la información recogida durante la realización de las pruebas.

Más aún, para hacer hincapié en el potencial de TEA, planteamos un estudio exploratorio haciendo uso de tecnología de ET mientras los participantes realizaban las pruebas. De las 5 formas posibles de validar las inferencias realizadas entre las afirmaciones y el comportamiento observado identificadas por Oranje et al. (2017), nos centramos en utilizar los datos registrados con dos fines: 1) generar y probar inferencias sobre el constructo de interés; 2) mejorar los diseños de los ítems analizando los patrones de comportamiento mostrados por los participantes. Las evidencias recogidas de este estudio sobre los RPs de los participantes se utilizaron para complementar las fuentes de evidencia obtenidas del resto de los estudios realizados durante la investigación. A pesar de que la herramienta se ha diseñado con un enfoque sumativo, hemos encontrado de especial interés poder mostrar a los participantes las respuestas correctas del formato de ítems que seleccionamos para este estudio (ítems basados en imágenes o simulaciones, donde el participante tiene que observar y evaluar una situación para posteriormente seleccionar la opción correcta de la una lista de alternativas). Hoy en día no conocemos una herramienta que esté ofreciendo esta opción de revisión de la "respuesta correcta" en la que se le mostrará un modelo de respuesta superpuesto sobre la imagen del ítem, creado en base al comportamiento "tipo" mostrado por los participantes que la han acertado. Esta opción se implementará a corto plazo en BAIT y sólo será consultable cuando un participante haya fallado un ítem.

Además, factores como el aburrimiento (a un participante con nivel alto de CD se le administran demasiadas preguntas muy fáciles para su nivel) y la frustración (a un participante con nivel bajo de CD se le administran demasiadas preguntas muy difíciles para su nivel) deben ser vigilados en la realización de las pruebas de evaluación, si queremos evitar que abandonen la prueba o muestren una falta de compromiso durante la realización de las pruebas. Una posible solución es aplicar TAI (Troussas et al., 2020). En las pruebas adaptativas en lugar de mostrar a cada participante la misma prueba, después de cada nueva respuesta haciendo uso de LA, se actualiza la estimación de la capacidad del participante y se selecciona el siguiente ítem para que sea el que tenga propiedades óptimas de cara a una nueva estimación (Van der Linden y Glas, 2010). En consecuencia, cada participante puede acabar respondiendo a un conjunto diferente de ítems, aquellos que son más informativos con respecto a sus estimaciones de habilidad, y da a lugar a

evaluaciones más cortas sin perder la precisión de las mediciones (Gardner et al., 2004).

En la revisión de la literatura hemos podido constatar que su aplicación en el campo de la evaluación de la CD es casi nula, exceptuando el trabajo realizado por Vie et al. (2017), del cual no se dispone de suficiente información para ser evaluado en profundidad. A la hora de implementar un TAI, es necesario evaluar una serie de puntos clave, ya que puede resultar que aplicar un enfoque de tipo adaptativo no sea el más adecuado. Así que repasamos los principales puntos a tener en cuenta y limitaciones a considerar para la implementación de una prueba adaptativa en un contexto de evaluación de la CD. En concreto, para la implementación de una evaluación sumativa basada en la TRI, teniendo en cuenta que partimos de un banco de ítems calibrado con el modelo de Rasch.

Un aspecto clave es el tamaño del banco de ítems, ya que debería incluir ítems para todo el rango de niveles de dificultad que se desea evaluar, desde ítems fáciles a difíciles, y teniendo en cuenta que el colectivo destinatario es la ciudadanía donde el rango de niveles es muy variado, deberemos de contar con un banco de ítems importante. Más aún, teniendo en cuenta nuestro planteamiento incluyendo ítems con formatos dinámicos, los cuales son más costosos de implementar, se prevé que la construcción del banco de ítems requerirá un gran esfuerzo a en cuenta, incluyendo su calibración y continua revisión. Asimismo, destacar que hemos calibrado el banco de ítems con el modelo de Rasch debido a que no contamos con una cantidad de respuestas tan grande como nos hubiese gustado, para haber podido aplicar un modelo más parámetros. Posiblemente, por los formatos de ítems utilizados el parámetro del pseudoacierto sea muy bajo.

Otro aspecto clave en estos sistemas, es el balanceo de contenidos, que es la parte más importante de la gestión de un TAI sobre todo en un contexto de pruebas de certificación. En este contexto, las habilidades a evaluar se consideran en su mayoría unidimensionales. Sin embargo, también se puede evaluar un dominio relativamente unidimensional, con varios subdominios de contenido subyacentes a la dimensión principal. Pero en este caso, un TAI unidimensional no considera los diversos subdominios de contenido de los ítems como parte del procedimiento de selección de los ítems. Para solucionar este punto, varios autores han propuesto distintos procedimientos para balancear el contenido (p. ej. Kingsbury y Zara, 1991). Por ejemplo, un TAI centrado en el AC de IAD, podría administrar los ítems manteniendo una proporción preestablecida de contenido, con un 33% destinado a cada una de las 3 CD que componen esta AC. Cabe destacar que, al modificar el procedimiento de selección de ítems en su búsqueda del ítem óptimo, el balanceo de contenidos afectará a la eficacia del TAI, resultando en pruebas más largas que un TAI no balanceado en contenidos. Más aún, el procedimiento de balanceo de contenidos no provee al participante un nivel de habilidad estimado en cada subdominio de contenido, sino una única estimación de la habilidad general basada en la prueba (Weiss y Kingsbury, 1984). Considerando estos aspectos relacionados con el balanceo de contenidos, para este primer aproximamiento que realizamos en nuestra investigación priorizamos la simplicidad, y los pospusimos para futuras líneas de investigación.

Además, hay dos posibles enfoques que podrían ser aplicados en nuestro contexto: el TAI tradicional y el TME. Un TME facilita llevar a cabo el balanceo de contenidos a nivel de módulo, y consecuentemente, a nivel de prueba, en la cual la habilidad del participante se estima mediante módulos adaptativos, lo que garantiza una precisión elevada y similar. Además, los TMEs aportan ventajas como permitir a los participantes revisar sus respuestas antes de pasar a la siguiente etapa, sin necesitar modelos complicados para la revisión de las respuestas (Wang et al., 2015). Sin embargo, su potencial es directamente proporcional al grado de complicidad a la hora de implementarlo, p. ej. cada módulo ha de ser ensamblado cumpliendo una serie de características. Razón por la cual nos decantamos por un TAI tradicional, apuntando la línea basada en un TME para futuras líneas de investigación.

Después del análisis realizado en esta parte de la investigación, hemos filtrado los diferentes aspectos identificados y abordarlos en esta investigación. Para ello, nos hemos centrado en los puntos que más nos interesan de cara a nuestra experimentación, teniendo presente en todo momento que el objetivo final es la mejora continua del servicio BAIT, que se nutre de los resultados de esta tesis. Razón por la cual la herramienta de evaluación ETCD comparte muchas similitudes con la de BAIT. Por ejemplo, ambas comparten la misma arquitectura de software y hacen uso de la tecnología Articulate para el diseño e integración de las simulaciones interactivas. En la figura 2.9 se muestran las principales áreas abordadas en la investigación, donde hemos marcado con cruces rojas las principales áreas donde hemos identificado carencias de interés para nuestra investigación, y con cruces verdes desde que perspectiva las hemos abordado. En los siguientes capítulos describimos como hemos abordado los principales aspectos identificados en esta investigación.

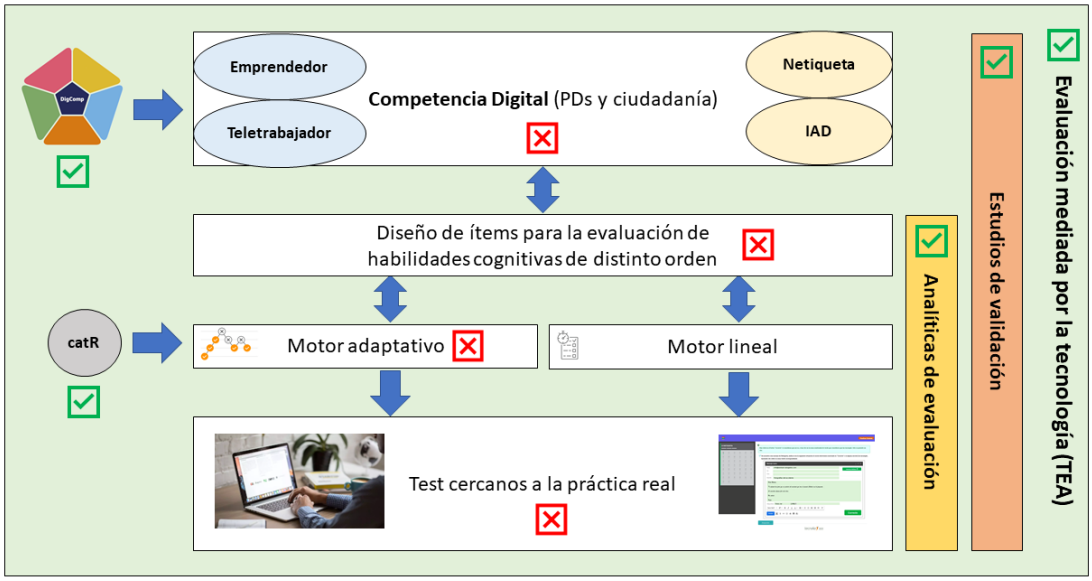


Figura 2.9. Principales carencias identificadas en la investigación y cómo se han abordado.

No estamos en una era de cambios, sino en un cambio de era caracterizado por la digitalización de prácticamente todo lo que nos rodea.

Emérito Martínez

3.

Evaluación y Acreditación de Perfiles Digitales (P4E)

Europa sigue careciendo de la mano de obra digitalmente cualificada que se necesita, y el reconocimiento de la CD como componente transversal que apoya otras competencias clave es un punto crucial. En consecuencia, están surgiendo muchos sistemas para la evaluación y certificación, utilizando diferentes enfoques, la mayoría basados en autoevaluaciones, sin cubrir todas las CD, y centradas en habilidades cognitivas de orden bajo (Kluzer y Priego, 2018; Siddiq et al., 2016).

Además, no existe un sistema que permita evaluar el nivel de CD de los trabajadores en función de sus puestos y funciones, es decir, sus diferentes perfiles digitales (PD), y así, poder mejorar la CD de los distintos PD. Ni tampoco se ha abordado directamente la definición de ciertos PD como los teletrabajadores o los emprendedores de manera que puedan satisfacer sus necesidades futuras.

Asimismo, es necesario que las diferentes partes interesadas contribuyan al diseño de un ecosistema para la recapacitación de adultos, incluyendo elementos clave como una evaluación inicial para reconocer las habilidades existentes, o un análisis de la demanda de habilidades actual y reflejarla en el diseño de PD (Foro Económico Mundial, 2017). La demanda de competencias está evolucionando rápidamente, afectando de manera pronunciada a las CD requeridas para las distintas categorías de trabajo y ocupaciones. Sin embargo, no existen especificaciones para diseñar sistemas de recapacitación de adultos basados en un lenguaje y una terminología comunes. La mayoría de los marcos de alfabetización digital adoptados se basan en cursos de formación y marcos de evaluación utilizados por empresas comerciales (Law et al., 2018). Y, en consecuencia, las herramientas de formación y evaluación de CD están fuertemente influenciadas por el marco elegido.

Afortunadamente, la llegada de DigComp facilitó el desarrollo de implementaciones personalizadas, tal y como Kluzer y Priego (2018) listaron en su revisión. Entre las implementaciones listadas, aparece la herramienta de evaluación y acreditación de CD de PD P4E, cuyos detalles describimos en este capítulo. Para su diseño e implementación, utilizamos el marco DigComp siguiendo un enfoque pragmático. Este tipo de enfoques se adaptan bien al propósito de la acreditación profesional, sin embargo, es bastante difícil lograr una herramienta de calidad en términos de fiabilidad y validez. Un enfoque pragmático tiende a penalizar la validez interna, pero en cambio, ofrece una mejor validez externa, ya que la herramienta es mejor comprendida y aceptada por terceras partes interesadas. Por consiguiente, afrontamos el reto de equilibrar la validez interna y externa, tanto a través de consideraciones metodológicas como del diseño del instrumento de evaluación, siguiendo un enfoque similar al sugerido por Laanpere (2019) para el DLGF de la UNESCO.

Cabe mencionar que este estudio fue parcialmente financiado por la Comisión Europea (ERASMUS+ 2016-1-ES01 KA204-024983). En el mismo participó el consorcio del proyecto compuesto por: Tecnalia Research & Innovation (ES), The Institute of Entrepreneurship Development (GR), All Digital (BE), FIT-The ICT Pipeline (IR), KZgunea (ES), Consorcio Fernando de Los Ríos (ES) y SIA Data Media Group (LV).

3.1. Objetivos de investigación

El objetivo de esta investigación es ayudar a dar respuesta a la siguiente pregunta de investigación, para lo cual diseñamos y desarrollamos P4E, y la cual propusimos a personas interesadas dotarse de las CD necesarias para teletrabajar o ser emprendedoras:

PI_K1: ¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?

3.2. Metodología

Para responder a nuestra pregunta de investigación, llevamos a cabo este estudio siguiendo un enfoque pragmático, tomando DigComp como referencia. Nos basamos en una metodología mixta contando con diferentes fuentes de información (consulta a expertos, cuestionarios y resultados de pruebas) para el diseño y validación de la herramienta que presentamos en este capítulo. La metodología cualitativa desempeñó un papel fundamental en términos de validez. Recogimos comentarios y sugerencias muy valiosos del intercambio de opiniones con expertos y de los cuestionarios en línea recibidos, que complementaron el análisis estadístico. La figura 3.1 se muestran los diferentes pasos llevados a cabo en el desarrollo de la herramienta de evaluación.

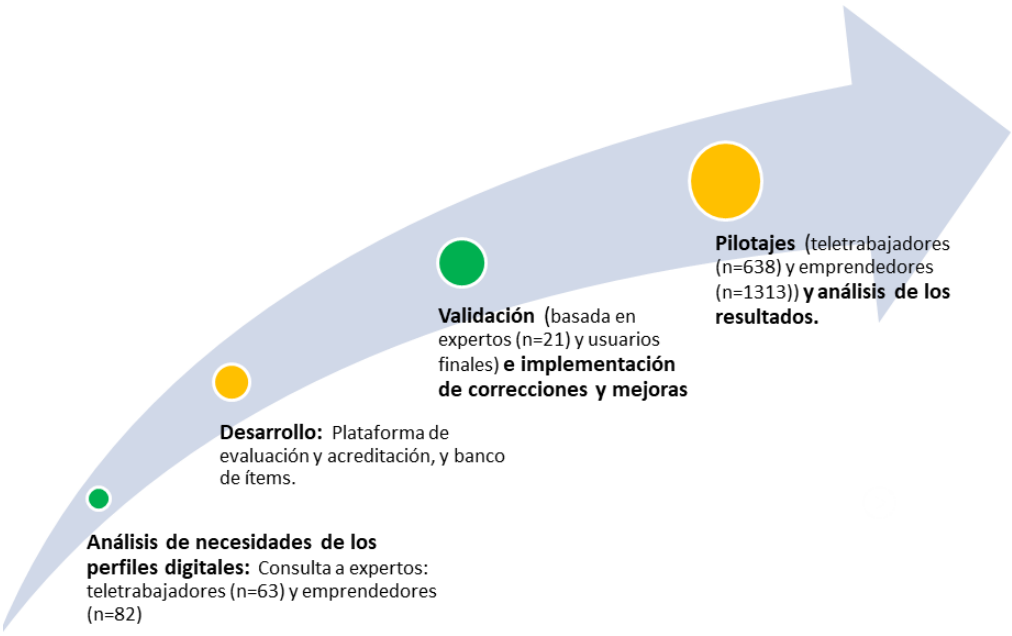


Figura 3.1. Metodología seguida para el desarrollo de la plataforma de evaluación y acreditación.

En este capítulo, describimos nuestro enfoque específico y los resultados de cada fase, presentando los resultados de cada una de ellas en la sección Resultados. Cabe comentar que llevamos a cabo las distintas fases en orden, aunque algunas de ellas podían gestionarse simultáneamente.

3.2.1. Análisis de necesidades de los PD.

Participantes

En esta fase participaron facilitadores pertenecientes a distintas redes de telecentros nacionales y de ámbito europeo, y usuarios finales. Analizamos las necesidades de CD de los adultos interesados en el teletrabajo o el emprendimiento, con la colaboración de profesores especializados en formación. Contactamos con teletrabajadores y diferentes actores implicados en el área del emprendimiento.

De los 82 encuestados para el PD del emprendedor, 39% fueron formadores, profesores o facilitadores, el 19% fueron gestores o coordinadores de proyectos, el 16% fueron emprendedores, directores o propietarios de PYMES, y el resto fueron empleados en PYMES, consultores y otros. De los 63 encuestados para el PD del teletrabajador, el 26% fueron directores o coordinadores de proyectos, el 18% teletrabajadores, el 15% trabajadores autónomos del sector TIC, el 13% formadores, profesores o facilitadores electrónicos, y otras profesiones menos frecuentes como investigadores 11,3% y consultores 8,1%.

Además, el 55% de los encuestados fueron hombres y el 45% mujeres. El 59% tenía entre 35 y 55 años y el 33% entre 25 y 34 años. Los encuestados procedían de

15 países europeos diferentes, con un 52% procedentes de España, el 19% de Letonia, el 13% de Grecia y el resto de otros países.

Materiales

Utilizamos cuestionarios en línea implementados en Smartsurvey[46] en los cuales hicimos uso de los marcos DigComp y EntreComp (para el caso particular del PD del emprendedor) para la elaboración de las preguntas.

Diseño

El objetivo de este estudio es comprender qué CD necesita una persona interesada en teletrabajar o convertirse en emprendedor independientemente del campo de actividad. En la primera parte del cuestionario les pedimos que indicasen la relevancia y el nivel necesario para un buen desempeño que consideraban para cada CD. Adicionalmente, en el cuestionario del emprendedor, les pedimos que indicaran cuáles de las 15 competencias definidas en EntreComp podrían ser habilitadas mediante las CD y en concreto, con cuáles. En el cuestionario del teletrabajador, adicionalmente les preguntamos por aquellas competencias a nivel general que deberían poseer los teletrabajadores (como la gestión del tiempo, la comunicación, la conciliación de la vida laboral y familiar, etc.) y qué CD las facilitarían. Además, para ambos PD les pedimos ejemplos de tareas realizadas con esas CD, que nos sirviesen de orientación en el diseño los ítems de evaluación.

Procedimiento

La recolecta de información vía cuestionarios en línea se llevó a cabo tanto a nivel nacional (España, Letonia, Grecia e Irlanda) como a nivel europeo, por parte de All Digital. A continuación, calculamos dos indicadores, la Media de Relevancia y de Nivel para cada CD, y el porcentaje de respuestas totales para cada CD en relevancia y nivel. Para tener una visión global de las CD de ambos PD, examinamos la posición de las CD atendiendo a la Media de Relevancia y Nivel. Esta información mostró la importancia de cada CD, y la Media de Relevancia la usamos para categorizar las competencias de ambos PD.

3.2.2. Desarrollo de la plataforma de evaluación y acreditación, y del banco de ítems

Participantes

El desarrollo fue llevado a cabo por un equipo multidisciplinar de 21 profesionales del consorcio del proyecto P4E. El 33% eran investigadores o técnicos, el 43% eran gestores o coordinadores de proyectos y el 23% eran directores o propietarios de

[46] https://www.smartsurvey.com/

PYMES. El 48% eran hombres y el 52% mujeres. Los profesionales procedían de 5 países europeos diferentes, el 33% de España, el 10% de Letonia, el 19% de Grecia y el 14% de Bélgica. Además, el 43% pertenecía a redes de telecentros de CD.

Materiales

Se creo una lista de escenarios para cada PD en los que las CD eran facilitadoras (ver tablas 3.1 y 3.2). Esta lista se creó a partir de la revisión de la literatura, la experiencia de los socios y los expertos consultados (ver punto 3.2.1).

Competencias	Escenarios
• Formación a distancia • Búsqueda de información • Instalación de software • Gestión de la identidad digital	Como actividad inicial, deberá solicitar el alta en organismos de la administración pública, como Hacienda Pública (alta como autónomo) y Seguridad Social (cotización). Para ello, primero deberá habilitar un DNI electrónico, que le permitirá el acceso, consulta y realización de trámites (por ejemplo, la herramienta BILA-hacienda).
• Formación a distancia • Gestión de la identidad digital	Periódicamente realiza los trámites necesarios con la hacienda pública para informar de su actividad (declarar trimestralmente los impuestos, la declaración anual del IRPF y el registro de operaciones económicas).
• Búsqueda y evaluación de información • Identificación de necesidades • Herramientas de gestión • Formación a distancia	Para llevar la contabilidad, debe buscar un programa informático adaptado a sus necesidades, bien sea un servicio web o un proveedor externo.
• Búsqueda de información • Guardado y recuperación de información • Formación a distancia	Para mantenerte al día en cuanto a información legislativa y técnica se refiere, realiza la suscripción al boletín periódico de información legislativa (reales decretos, BOE, etc.) y, por otro lado, busca aplicaciones o bases de datos disponibles en la web que tengan información útil para su profesión.
• Gestión de licencias • Búsqueda y evaluación de información • Formación a distancia • Herramientas de gestión	Toda la información y documentación externa que necesita junto con la que genera (documentos técnicos, informes, facturas, planos, etc.) debe gestionarla a través de una herramienta de gestión documental que facilite la gestión de cambios y versiones.
• Interacción tecnológica • Compartir información y contenidos • Protección de la información	Se comunica con clientes utilizando las herramientas oportunas, y comparte información, documentos y noticias sobre su negocio. Para ello, utiliza distintos tipos de aplicaciones desde distintos dispositivos.
• Desarrollo de contenidos • Programación	Publicita su actividad utilizando inicialmente tres medios. Primero, diseña un cartel publicitario con los principales productos o servicios que ofrece su negocio con una herramienta ofimática. Además, registra su perfil en una red

• Herramientas de diseño • Participación ciudadana en línea • Colaboración a través de medios digitales	social (ej. LinkedIn). Por último, utiliza un dominio público para publicar su sitio web diseñado con una herramienta de diseño.

Tabla 3.1. Escenarios seleccionados para el emprendedor.

Competencias	Escenarios
• Protección de dispositivos • Búsqueda de información • Conectividad • Ergonomía	Accede a un lugar público con WIFI para trabajar, y busca el lugar más adecuado para trabajar teniendo en cuenta aspectos de ergonomía y de calidad de la señal. Al ser una red WIFI abierta, utiliza herramientas de encriptación seguras.
• Privacidad • Navegación segura • Seguridad de los dispositivos	Va a un cibercafé porque tiene que utilizar un ordenador. Al ser un equipo público, tiene en cuenta que las medidas de seguridad aplicadas son desconocidas y que la huella digital debe ser protegida.
• Conectividad • Ahorro de energía • Identificación de necesidades	Va a trabajar en un lugar donde no sabe si hay conexión WIFI disponible y, por precaución, lleva un módem USB. Como no sabe si tendrás acceso a una fuente de alimentación, aplica medidas ahorro energético. Al usar la conexión 4G, tiene en cuenta el tamaño de los archivos.
• Herramientas de comunicación • Herramientas de almacenamiento • Herramientas de planificación • Conectividad • Configuración de dispositivos	Organiza una reunión con varias personas. Para la comunicación, hará uso de herramientas de videollamada, que te permiten compartir el escritorio, etc. Toda la información estará ubicada en un servicio de almacenamiento en la nube que mantiene la información sincronizada.
• Protección de los dispositivos • Protección de la información	Accede a la información desde distintos dispositivos, e implementa medidas de seguridad. Limita el acceso a los dispositivos y a la información, así como, supervisa las conexiones hacia y desde el ordenador. Utiliza antivirus y realiza copias de seguridad.
• Presentaciones en línea • Vídeos en línea • Búsqueda de información • Geolocalización	Visita las oficinas de un cliente para realizar una presentación. Como no sabe la ubicación, usa una app de geolocalización desde su móvil para llegar al destino.

Tabla 3.2. Escenarios seleccionados para el trabajador en movilidad.

Diseño

Llevamos a cabo el diseño realizando nuestra propia implementación basada en DigComp, un marco multidimensional estructurado en cinco AC. Aunque cada AC tiene sus propias particularidades, existen referencias cruzadas a nivel de CD con otras AC, tal y como se ha indicado en el propio DigComp. Teniendo en cuenta la complejidad de esta cuestión, decidimos desarrollar una prueba para cada AC y PD

(10 pruebas en total), sin considerar las relaciones a nivel de AC y CD, o el diseño se habría complicado notablemente. Más aún, si una CD o un nivel no era necesario para un PD en conceto, los ítems correspondientes no los incluimos en la prueba.

Las pruebas evalúan los 3 componentes de la CD (conocimiento, habilidad y actitud), para cada una de las 5 AC, y proporciona a los participantes una foto de sus niveles de CD. Asimismo, nos basamos en un enfoque basado en el rendimiento e incorporamos diferentes formatos de ítems.

Diseñamos los ítems de las pruebas en términos de narrativa y los contextualizaron de acuerdo con situaciones reales de la lista de escenarios. Además, diseñamos la herramienta de evaluación y acreditación abordando los 6 niveles de CD de acuerdo con DigComp 2.1, que son los niveles de CD más requeridos para la mayoría de los ciudadanos europeos y para la mayoría de los PD. Desarrollamos el banco de ítems en 5 idiomas (inglés, español, euskera, griego y letón), con el objetivo de no añadir una dificultad adicional por el idioma utilizado. En concreto, desarrollamos 3 tipos de ítems de evaluación diferentes según el componente de la CD que pretendían evaluar. Todos ellos eran ítems dicotómicos excepto la pregunta de actitud compuesta por una escala Likert de 5 puntos con varias afirmaciones. Para los ítems de conocimiento, creamos uno para cada nivel según los niveles de DigComp 2.1: básico (1 y 2), intermedio (3 y 4) y avanzado (5 y 6). Nos basamos en el formato de opción múltiple con una única respuesta correcta posible. Además, algunos ítems son dependientes del contexto, es decir, incluyen un escenario, una tabla o material visual, que los participantes deben analizar para responder al ítem. Este formato es muy útil para medir habilidades complejas como el pensamiento crítico (Haladyna et al. 2002).

Para evaluar la habilidad, usamos dos tipos de ítems:

- Simulaciones, basadas en imágenes interactivas, en las que los participantes tienen que interactuar para resolver la tarea demandada en el enunciado. El área de los clics se evalúa automáticamente antes de decidir si se debe mostrar o no un nuevo paso. Diseñamos las simulaciones basándonos en los programas más extendidos y utilizados.
- Tareas prácticas, en las que los participantes tienen que interactuar con la estación de trabajo y sus programas, para llevar a cabo la tarea solicitada en el enunciado (ver ejemplo en la figura 3.2). Los participantes tienen que interactuar con el ordenador y sus programas, para llevar a cabo la tarea solicitada y subir los resultados a la plataforma. Integramos este tipo de ítems haciendo uso de APIs como Apache POI47 para evaluar automáticamente los resultados.

[47] https://poi.apache.org/

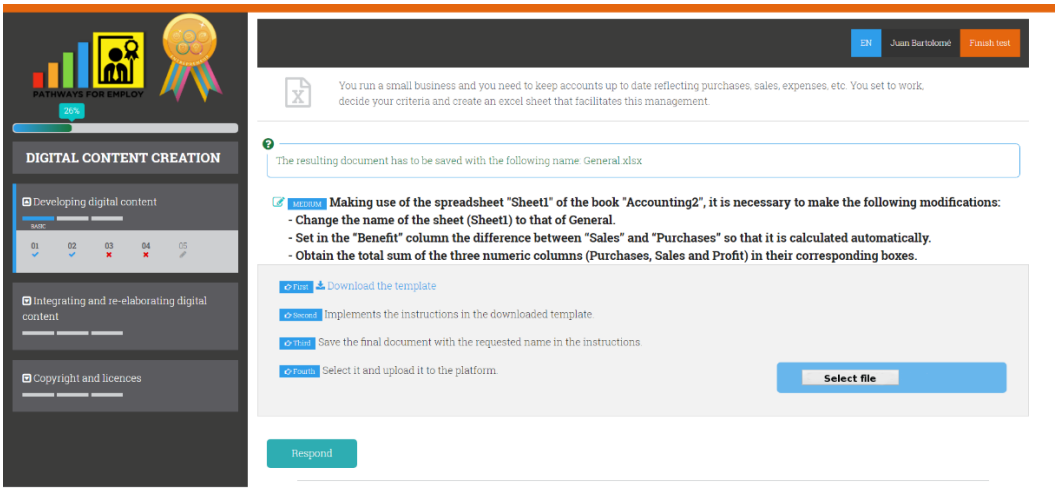


Figura 3.2. Ejemplo de un ítem basado en una tarea práctica.

En cuanto a las preguntas de actitud, intentamos alinear las afirmaciones con el resto de los ítems de la misma CD. Éramos conscientes de que estas escalas debían ser validadas para inferir conclusiones válidas. Sin embargo, decidimos incluirlas para concienciar a los usuarios de que este componente de la CD es de vital importancia, a pesar de que los resultados no se utilizaron.

Por último, diseñamos una serie de funcionalidades necesarias para la herramienta P4E como el registro de usuarios, el repositorio de PD, el PD personal, el historial de pruebas realizadas y, por último, la posibilidad de poder comparar los niveles alcanzados por CD con los niveles recomendados en los PD por los expertos. El desarrollo se basó en software de código abierto: Java Platform Enterprise Edition (J2EE), Struts, iBATIS y MySQL.

Procedimiento

El desarrollo de la herramienta de evaluación, así como el banco de ítems, comenzó con un diseño inicial, que fue probado y perfeccionado de forma iterativa con los resultados obtenidos en las fases de validación (ver punto 3.2.3) y pilotajes con usuarios finales (ver punto 3.2.4).

3.2.3. Validación basada en expertos y usuarios finales, e implementación de correcciones y mejoras.

Participantes

En esta fase, contrastamos información con 21 expertos de diferentes países (3 de Letonia, 3 de Grecia, 3 de Irlanda, 9 de España y 3 de la red internacional de All Digital). Cada socio del proyecto seleccionó al menos 3 formadores de CD, profesores, investigadores o expertos en emprendimiento, basándonos en los criterios: experiencia en CD, experiencia con herramientas de evaluación, dominio

del inglés, y conocimiento general sobre los perfiles y tareas de los emprendedores y teletrabajadores.

Materiales

Los expertos realizaron las pruebas disponibles en la plataforma y aportaron sus comentarios rellenando un cuestionario en línea implementado en Smartsurvey y proporcionando comentarios por correo electrónico, o en conversaciones con ellos durante la realización de las pruebas. Toda esta información la clasificamos bajo cinco categorías: Lengua, Aspectos técnicos, Contenido, Estructura y Diseño.

Diseño

Los expertos de manera individual eligieron un PD y realizaron las 5 pruebas de evaluación proporcionando su grado de satisfacción con los siguientes aspectos:

- La adecuación del contenido de las preguntas, en términos de conocimientos, habilidades y actitudes, para esa AC.
- La adecuación de los criterios de evaluación y los niveles de dificultad asignados.
- La calidad de la información proporcionada al finalizar la prueba.
- La satisfacción general con la prueba de evaluación.
- La satisfacción general con la plataforma de evaluación y acreditación.

Procedimiento

Esta fase tuvo lugar entre abril y mayo (2018) en 5 países (Letonia, Bélgica, España, Irlanda y Grecia), donde se probaron la plataforma online, los PD y la herramienta de evaluación en inglés. Contactamos con la mayoría de ellos vía correo electrónico y les hicimos llegar los detalles del proyecto, los objetivos, instrucciones, así como los enlaces a los cuestionarios en línea. En esta fase de validación nos basamos en dos pilares fundamentales:

- La realización de nuestra implementación basada DigComp siguiendo la estructura de AC y CD, para garantizar la estructura interna de la prueba.
- La realización de una validación basada en el panel de expertos utilizando un enfoque de diseño top-down y revisando exhaustivamente las fuentes bibliográficas, así como instrumentos similares ya existentes (Sireci y Faulkner-Bond 2014). Analizamos la relación entre el contenido de las pruebas y el constructo que se pretendía medir, y prestamos especial atención a la redacción, al formato de los ítems, las preguntas, los niveles y las tareas incluidas. Cada experto revisó las 5 pruebas para el PD seleccionado: Teletrabajador (n=11) y Empresario (n=10).

A raíz de los comentarios y sugerencias recibidos por los expertos, llevamos a cabo una revisión del banco de ítems, que inicialmente contaba con 120 ítems de conocimiento y 44 de habilidad, con los siguientes criterios:

- Pertenecer al menos a un escenario.
- Ser lo más cortos y sencillos posible.
- Abordar la CD seleccionada, el componente (conocimiento, habilidad o actitudes) y el nivel (básico, intermedio o avanzado) específico.

A la hora de elegir el número de ítems a incluir en las pruebas, se tuvo en cuenta el tiempo necesario para completar la prueba, deseablemente 10 minutos, con el fin de minimizar los abandonos. Como resultado, una versión mejorada de la plataforma P4E y las pruebas quedaron listas para el siguiente paso.

3.2.4. Pilotajes con usuarios finales y análisis de los resultados

Participantes

Participaron 525 usuarios finales de 23 países diferentes, siendo España (72,53%), Grecia (13,18%) y Letonia (6,5%) los más presentes (ver tabla 3.3).

Género (%)		Situación laboral (%)			Rango de edad (%)			Nivel educativo (%)		
M	H	Empleado	Desempleado	Autónomo	16-24	25-54	>=55	Alto	Medio	Bajo
55.1	44.9	82.3	7.4	6.2	5.7	87.8	6.1	69.1	28.8	2.1

Tabla 3.3. Información de los participantes.

Además, según los resultados de las encuestas, los participantes estaban bastante seguros de su nivel de CD (39% avanzado, 43% intermedio, 8% básico y 10% inicial).

Materiales

Los participantes completaron las pruebas disponibles en P4E, y además dieron su opinión a través de cuestionarios en línea que se administraron al final de las pruebas. Tanto las pruebas como los cuestionarios estuvieron disponibles en letón, griego, español e inglés.

Diseño

Recogimos la siguiente información durante el registro de los participantes: sexo, situación laboral, rango de edad y nivel educativo. Además, a partir de las pruebas

realizadas registramos las respuestas para cada ítem y calculamos automáticamente los niveles alcanzados en cada CD.

A partir de los cuestionarios en línea recogimos información sobre los siguientes aspectos con el fin de analizar y mejorar la plataforma:

- Rendimiento técnico de cada una de las pruebas.
- Satisfacción general con la herramienta P4E.
- Satisfacción general con el contenido de la herramienta.
- Comprensibilidad de los resultados mostrados al finalizar las pruebas.
- Satisfacción con el diseño de las pruebas.

Además, también calculamos algunos indicadores de la fiabilidad global y la calidad de cada ítem. Para comprobar la consistencia interna de las pruebas calculamos el alfa de Cronbach. Examinamos las medias y la desviación estándar para mostrar la distribución de los ítems, y calculamos los índices de dificultad y de discriminación de los ítems.

Procedimiento

Llevamos a cabo esta fase entre junio (2018) y enero (2019) con el objetivo de probar P4E con usuarios finales. Invitamos a los participantes usando 2 estrategias, mediante grupos controlados asociados a procesos formativos relacionados, y mediante convocatorias abiertas.

La población objetivo es personas de 16 a 65 años, con diferentes niveles de CD a nivel europeo. Seleccionamos los países de los pilotajes de manera que representasen países con diferentes porcentajes de población cualificada a nivel de sociedad. Participaron 525 usuarios finales y se realizaron 1.951 pruebas, representando ambos PD: emprendedores (n=1.313) y teletrabajadores (n=638). Los idiomas del pilotaje fueron inglés, español, letón, griego y euskera. Se les invitó a evaluarse de las 5 AC para un PD específico, pero era posible realizar sólo algunas pruebas de un PD. Los resultados se guardaban automáticamente y el PD de los participantes se actualizaba automáticamente. Todos los ítems eran dicotómicos y se ponderaron según los siguientes criterios:

- Los ítems de conocimiento: nivel básico 1 punto, nivel intermedio 2 puntos, y nivel avanzado 3 puntos.
- Los ítems de habilidad 5 puntos.
- Los ítems de actitud no se calificaron.

Decidimos asignar una mayor puntuación a los ítems de habilidad por considerarlos más discriminatorios, ya requerían habilidades de orden superior para resolverlas. Los umbrales de cada CD se definieron en función de la puntuación obtenida: Nivel inicial (< 9%); básico (>= 9% y < 36%); intermedio (>= 37% y < 53%); y avanzado (> 54%).

Por último, recogimos la opinión de los participantes mediante cuestionarios en línea administrados al final de las pruebas en varios idiomas: 328 español, 60 inglés, 80 letón, 108 griego y 24 vasco (emprendedores (n=255) y teletrabajadores (n=343)).

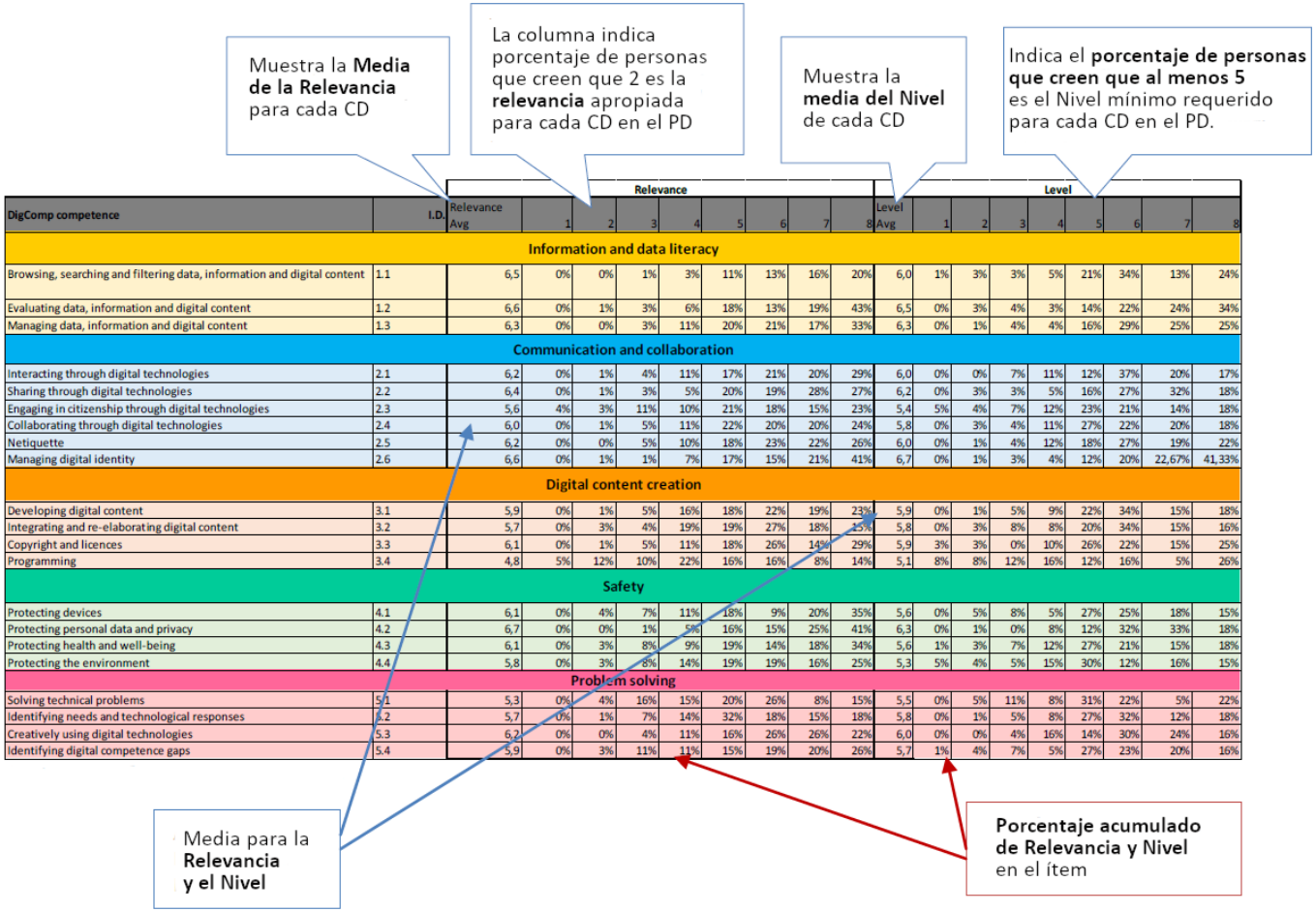
Analizamos las respuestas como parte del proceso de validación de la herramienta, con el fin de obtener evidencias relevantes para las interpretaciones de las puntuaciones obtenidas y su relevancia para el uso propuesto (Aesaert et al. 2014). En cuanto a la fiabilidad, consideramos que los resultados de las pruebas podrían ser corroborados por información de otras fuentes o que los resultados erróneos podrían ser fácilmente rectificados. Por lo tanto, las puntuaciones con una fiabilidad modesta podrían aceptarse teniendo en cuenta nuestras limitaciones de tiempo y presupuesto. También examinamos las propiedades psicométricas de los ítems. Hay que tener en cuenta que esta herramienta no pretende medir la CD actual de los individuos con el objetivo de obtener una certificación o fines similares. En ese caso, tendríamos que haber seguido un enfoque diferente, realizando las pruebas en un entorno controlado.

En primer lugar, analizamos la fiabilidad de las pruebas y realizamos un análisis de los resultados del paso anterior (n=1.951) basado en la TCT. Para comprobar la consistencia interna de las pruebas usamos el alfa de Cronbach, que indica indirectamente el grado en que un conjunto de ítems mide un único constructo latente unidimensional. Previamente, eliminamos las respuestas de los participantes que no respondieron al menos 2/3 de la prueba, suponiendo que no hicieron el examen con verdadero interés, quedando finalmente 994 pruebas. Además del alfa de Cronbach, también calculamos los estadísticos de los ítems, como la media y la desviación estándar, para mostrar la distribución de los ítems. Los estadísticos de los ítems pueden indicar qué ítems disminuyen la fiabilidad y, por tanto, debe considerarse su sustitución o eliminación. También calculamos el índice de dificultad y el índice de discriminación.

3.3. Resultados

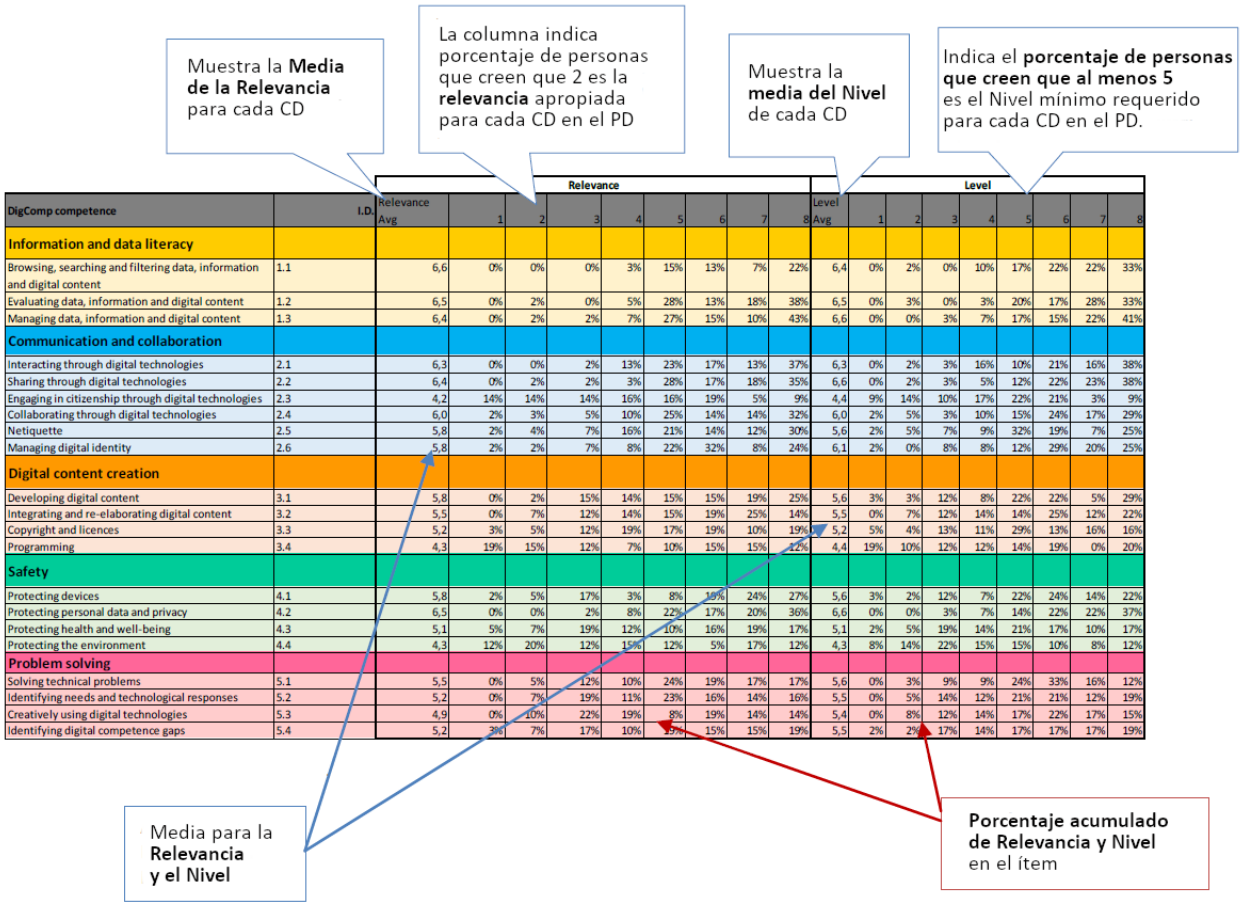
3.3.1. Análisis de necesidades de los PD

Para cada PD, obtuvimos una tabla que muestra la Media de Nivel y la Media de Relevancia para cada CD (ver figuras 3.3 y 3.4).



		Relevance									Level								
DigComp competence	I.D.	Relevance Avg	1	2	3	4	5	6	7	8	Level Avg	1	2	3	4	5	6	7	8
Information and data literacy																			
Browsing, searching and filtering data, information and digital content	1.1	6,5	0%	0%	1%	3%	11%	13%	16%	20%	6,0	1%	3%	3%	5%	21%	34%	13%	24%
Evaluating data, information and digital content	1.2	6,6	0%	1%	3%	6%	18%	13%	19%	43%	6,5	0%	3%	4%	3%	14%	22%	24%	34%
Managing data, information and digital content	1.3	6,3	0%	0%	3%	11%	20%	21%	17%	33%	6,3	0%	1%	4%	4%	16%	29%	25%	25%
Communication and collaboration																			
Interacting through digital technologies	2.1	6,2	0%	1%	4%	11%	17%	21%	20%	29%	6,0	0%	0%	7%	11%	12%	37%	20%	17%
Sharing through digital technologies	2.2	6,4	0%	1%	3%	5%	20%	19%	28%	27%	6,2	0%	3%	3%	5%	16%	27%	32%	18%
Engaging in citizenship through digital technologies	2.3	5,6	4%	3%	11%	10%	21%	18%	15%	23%	5,4	5%	4%	7%	12%	23%	21%	14%	18%
Collaborating through digital technologies	2.4	6,0	0%	1%	5%	11%	22%	20%	20%	24%	5,8	0%	3%	4%	11%	27%	22%	20%	18%
Netiquette	2.5	6,2	0%	0%	5%	10%	18%	23%	22%	26%	6,0	0%	1%	4%	12%	18%	27%	19%	22%
Managing digital identity	2.6	6,6	0%	1%	1%	7%	17%	15%	21%	41%	6,7	0%	1%	3%	4%	12%	20%	22,67%	41,33%
Digital content creation																			
Developing digital content	3.1	5,9	0%	1%	5%	16%	18%	22%	19%	23%	5,9	0%	1%	5%	9%	22%	34%	15%	18%
Integrating and re-elaborating digital content	3.2	5,7	0%	3%	4%	19%	19%	27%	18%	15%	5,8	0%	3%	8%	8%	20%	34%	15%	16%
Copyright and licences	3.3	6,1	0%	1%	5%	11%	18%	26%	14%	29%	5,9	3%	3%	0%	10%	26%	22%	15%	25%
Programming	3.4	4,8	5%	12%	10%	22%	16%	16%	8%	14%	5,1	8%	8%	12%	16%	12%	16%	5%	26%
Safety																			
Protecting devices	4.1	6,1	0%	4%	7%	11%	18%	9%	20%	35%	5,6	0%	5%	8%	5%	27%	25%	18%	15%
Protecting personal data and privacy	4.2	6,7	0%	0%	1%	5%	16%	15%	25%	41%	6,3	0%	1%	0%	8%	12%	32%	33%	18%
Protecting health and well-being	4.3	6,1	0%	3%	8%	9%	19%	14%	18%	34%	5,6	1%	3%	7%	12%	27%	21%	15%	18%
Protecting the environment	4.4	5,8	0%	3%	8%	14%	19%	19%	16%	25%	5,3	5%	4%	5%	15%	30%	12%	16%	15%
Problem solving																			
Solving technical problems	5.1	5,3	0%	4%	16%	15%	20%	26%	8%	15%	5,5	0%	5%	11%	8%	31%	22%	5%	22%
Identifying needs and technological responses	5.2	5,7	0%	1%	7%	14%	32%	18%	15%	18%	5,8	0%	1%	5%	8%	27%	32%	12%	18%
Creatively using digital technologies	5.3	6,2	0%	0%	4%	11%	16%	26%	26%	22%	6,0	0%	0%	4%	16%	14%	30%	24%	16%
Identifying digital competence gaps	5.4	5,9	0%	3%	11%	11%	15%	19%	20%	26%	5,7	1%	4%	7%	5%	27%	23%	20%	16%

Figura 3.3. Media de la Relevancia y Media del Nivel para el emprendedor.



		Relevance									Level								
DigComp competence	I.D.	Relevance Avg	1	2	3	4	5	6	7	8	Level Avg	1	2	3	4	5	6	7	8
Information and data literacy																			
Browsing, searching and filtering data, information and digital content	1.1	6,6	0%	0%	0%	3%	15%	13%	7%	22%	6,4	0%	2%	0%	10%	17%	22%	22%	33%
Evaluating data, information and digital content	1.2	6,5	0%	2%	0%	5%	28%	13%	18%	38%	6,5	0%	3%	0%	3%	20%	17%	28%	33%
Managing data, information and digital content	1.3	6,4	0%	2%	2%	7%	27%	15%	10%	43%	6,6	0%	0%	3%	7%	17%	15%	22%	41%
Communication and collaboration																			
Interacting through digital technologies	2.1	6,3	0%	0%	2%	13%	23%	17%	13%	37%	6,3	0%	2%	3%	16%	10%	21%	16%	38%
Sharing through digital technologies	2.2	6,4	0%	2%	2%	3%	28%	17%	18%	35%	6,6	0%	2%	3%	5%	12%	22%	23%	38%
Engaging in citizenship through digital technologies	2.3	4,2	14%	14%	14%	16%	16%	19%	5%	9%	4,4	9%	14%	10%	17%	22%	21%	3%	9%
Collaborating through digital technologies	2.4	6,0	2%	3%	5%	10%	25%	14%	14%	32%	6,0	2%	5%	3%	10%	15%	24%	17%	29%
Netiquette	2.5	5,8	2%	4%	7%	16%	21%	14%	12%	30%	5,6	2%	5%	7%	9%	32%	19%	7%	25%
Managing digital identity	2.6	5,8	2%	2%	7%	8%	22%	32%	8%	24%	6,1	2%	0%	8%	8%	12%	29%	20%	25%
Digital content creation																			
Developing digital content	3.1	5,8	0%	2%	15%	14%	15%	15%	19%	25%	5,6	3%	3%	12%	8%	22%	22%	5%	29%
Integrating and re-elaborating digital content	3.2	5,5	0%	7%	12%	14%	15%	19%	25%	14%	5,5	0%	7%	12%	14%	14%	25%	12%	22%
Copyright and licences	3.3	5,2	3%	5%	12%	19%	17%	19%	10%	19%	5,2	5%	4%	13%	11%	29%	13%	16%	16%
Programming	3.4	4,3	19%	15%	12%	7%	10%	15%	15%	12%	4,4	19%	10%	12%	12%	14%	19%	0%	20%
Safety																			
Protecting devices	4.1	5,8	2%	5%	17%	3%	8%	19%	24%	27%	5,6	3%	2%	12%	7%	22%	24%	14%	22%
Protecting personal data and privacy	4.2	6,5	0%	0%	2%	8%	22%	17%	20%	36%	6,6	0%	0%	3%	7%	14%	22%	22%	37%
Protecting health and well-being	4.3	5,1	5%	7%	19%	12%	10%	16%	19%	17%	5,1	2%	5%	19%	14%	21%	17%	10%	17%
Protecting the environment	4.4	4,3	12%	20%	12%	15%	12%	5%	17%	12%	4,3	8%	14%	22%	15%	15%	10%	8%	12%
Problem solving																			
Solving technical problems	5.1	5,5	0%	5%	12%	10%	24%	19%	17%	17%	5,6	0%	3%	9%	9%	24%	33%	16%	12%
Identifying needs and technological responses	5.2	5,2	0%	7%	19%	11%	23%	16%	14%	16%	5,5	0%	5%	14%	12%	21%	21%	12%	19%
Creatively using digital technologies	5.3	4,9	0%	10%	22%	19%	8%	19%	14%	14%	5,4	0%	8%	12%	14%	17%	22%	17%	15%
Identifying digital competence gaps	5.4	5,2	3%	7%	17%	10%	15%	15%	15%	19%	5,5	2%	2%	17%	14%	17%	17%	17%	19%

Figura 3.4. Media de la relevancia y media del nivel para el teletrabajador.

A continuación, obtuvimos un gráfico que muestra la posición de las CD para ambos PD teniendo en cuenta la Media de Relevancia y la Media de Nivel (ver figuras 3.5 y 3.6).

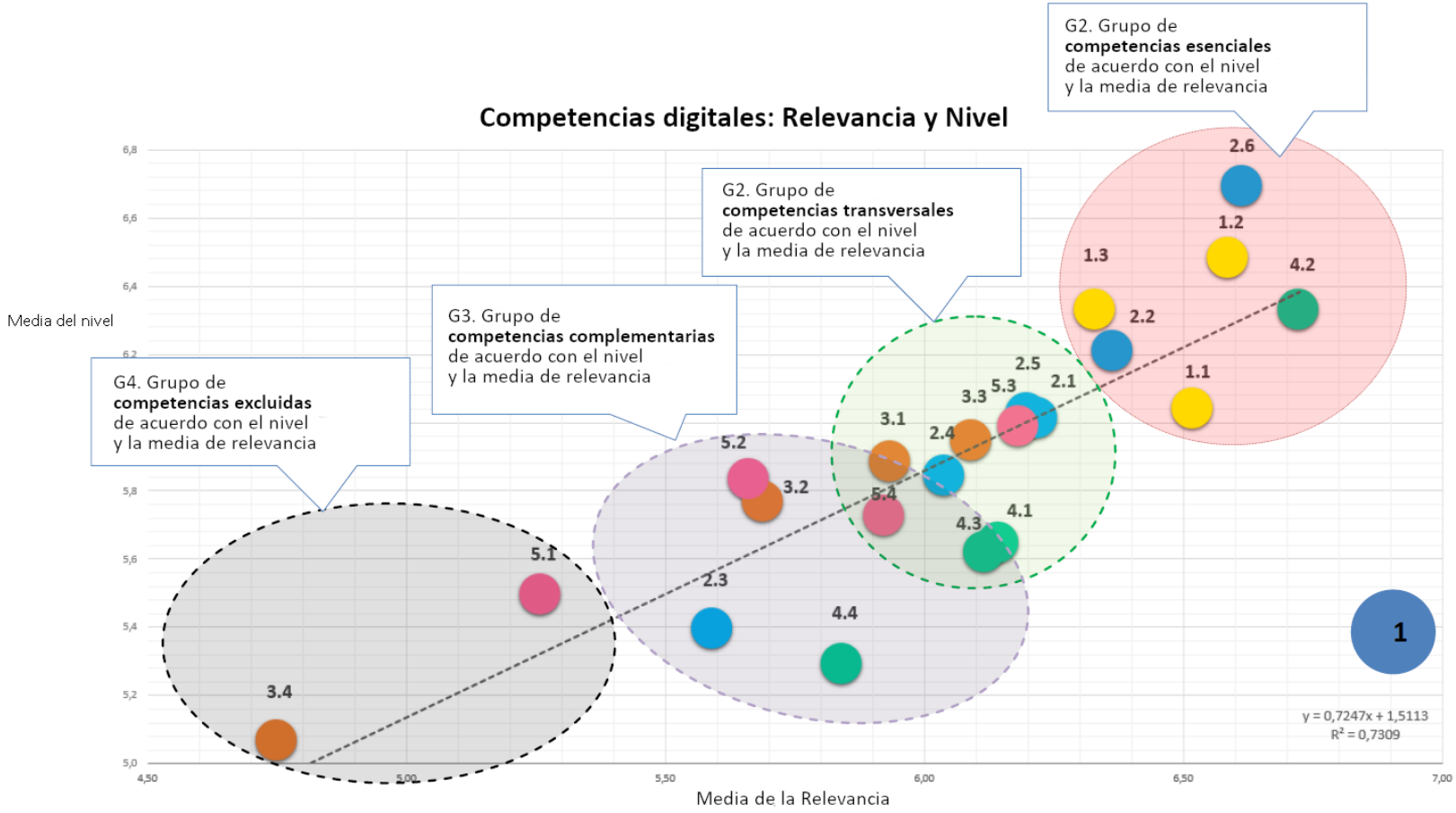


Figura 3.5. Representación gráfica de la media de la relevancia y media del nivel para el emprendedor.

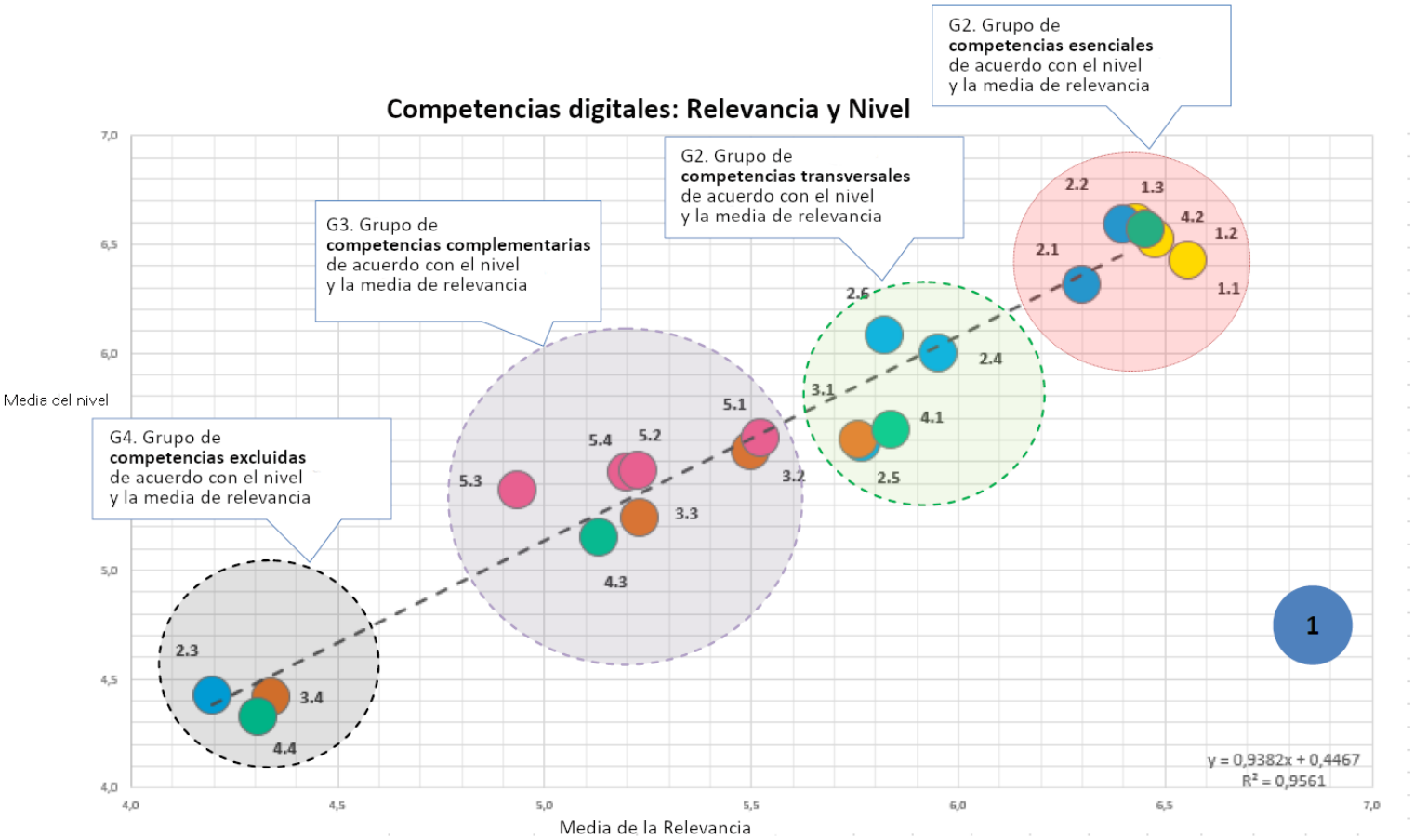


Figura 3.6. Representación gráfica de la media de la relevancia y media del nivel para el teletrabajador.

Las CD esenciales fueron las más importantes (las del primer grupo), por lo que debían ser de nivel avanzado, y, por lo tanto, tenían asociado un nivel alto en la media. Las competencias transversales fueron las CD con una relevancia media en la encuesta, consideradas como necesarias en cualquier tarea para este PD. Consideramos el nivel de las CD transversales menos exigente que el del grupo de CD esenciales, ya que mejoraban el PD, pero no eran esenciales, con un valor medio en relevancia y nivel. Además, identificamos un cuarto grupo que contenía competencias con baja relevancia, que fueron excluidas de ese PD. En cuanto al nivel de competencia, consideramos el porcentaje de respuestas y los siguientes criterios: (1) las CD esenciales se definieron por los percentiles 35. Por lo tanto, el 65% de los encuestados creyeron que este nivel era el mínimo requerido para cada competencia. Por ejemplo, en la primera CD del PD de emprendedor, al menos el 71% de los encuestados pensaba que el nivel mínimo exigido era el seis; (2) las competencias transversales y complementarias siguieron el mismo criterio, pero se eligió un percentil menor, el 25.

Finalmente, elaboramos ambos PD (ver figuras 3.7 y 3.8).

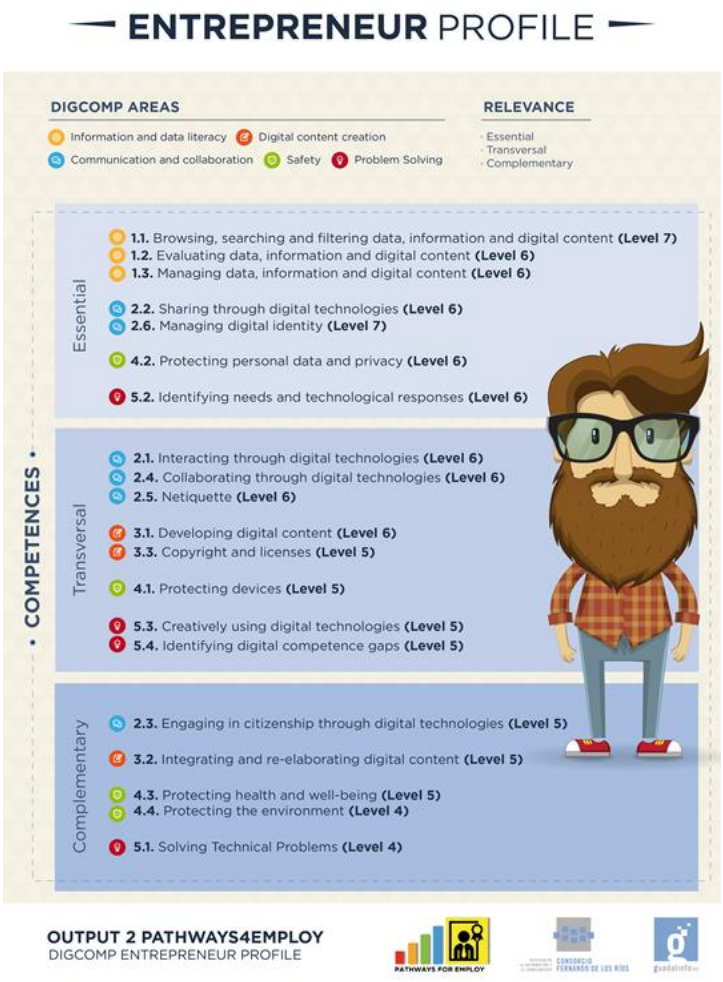


Figura 3.7. PD del emprendedor

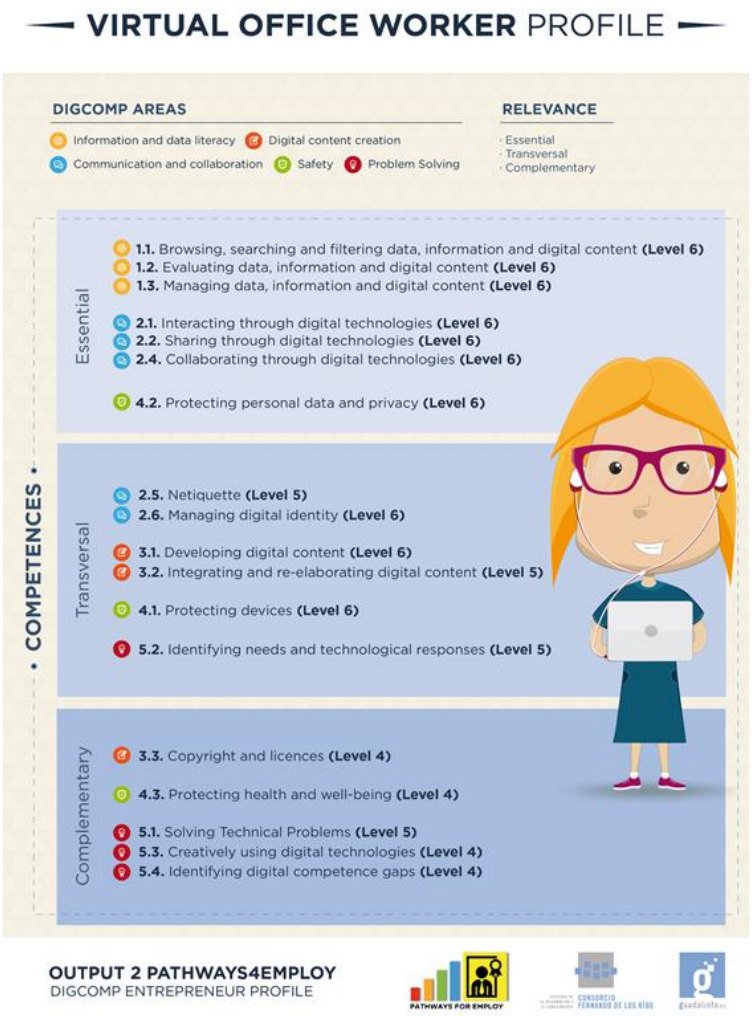


Figura 3.8. PD del teletrabajador

3.3.2. Desarrollo de la plataforma de evaluación y acreditación, y del banco de ítems

Desarrollamos el banco de ítems incluyendo 20 ítems por CD y los integramos P4E. Tras desarrollar y probar 204 ítems de conocimiento y 88 ítems de habilidad, finalmente seleccionamos 120 de conocimiento y 34 de habilidad. Las pruebas y los ítems pueden ser accedidos en http://pathwaysforemploy.com/.

3.3.3. Validación basada en expertos y usuarios finales, e implementación de correcciones y mejoras

El 71,4 % de los encuestados evaluó positivamente su satisfacción con la plataforma, tal y como se muestra en la tabla 3.4, y se identificaron varios aspectos a mejorar.

1 (No está nada bien)	2	3	4	5	6 (Excelente)	Análisis (M y SD)
0.0%	9.5%	19.0%	28.6%	33.3%	9.5%	M: 4.14 SD: 1.12

Tabla 3.4. Satisfacción general con la plataforma de evaluación y acreditación.

Los expertos coincidieron mayoritariamente en que el contenido de las preguntas reflejaba los conocimientos, habilidades y actitudes de la competencia (desde el 66,6 % en el módulo de *Creación de Contenido Digital* hasta el 80,9 % en la prueba *IAD*), tal y como se muestra en las tablas 3.5 y 3.6.

Prueba	1 (No está nada bien)	2	3	4	5	6 (Excelente)	Análisis (M y SD)
Área 1: Alfabetización en información y datos	0.0%	9.5%	23.8%	23.8%	38.1%	4.8%	M: 4.05 SD: 1.09
Área 2: Comunicación y colaboración en línea	0.0%	9.5%	9.5%	33.3%	38.1%	9.5%	M: 4.29 SD: 1.08
Área 3: Creación de contenido digital	9.5%	9.5%	9.5%	23.8%	42.9%	4.8%	M: 3.95 SD: 1.4
Área 4: Seguridad	0.0%	9.5%	9.5%	28.6%	42.9%	9.5%	M: 4.33 SD: 1.08
Área 5: Resolución de problemas	0.0%	4.8%	9.5%	28.6%	38.1%	19.0%	M: 4.57 SD: 1.05

Tabla 3.5. Satisfacción general con las pruebas de evaluación.

Prueba	Muy en desacuerdo	En desacuerdo	Ni de acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo	Análisis (M y SD)
Área 1: Alfabetización en información y datos	4.8%	9.5%	4.8%	61.9%	19.0%	M: 3.81 SD: 1.01
Área 2: Comunicación y colaboración en línea	4.8%	4.8%	14.3%	66.7%	9.5%	M: 3.71 SD: 0.88
Área 3: Creación de contenido digital	0.0%	4.8%	28.6%	57.1%	9.5%	M: 3.71 SD: 0.7
Área 4: Seguridad	0.0%	14.3%	9.5%	61.9%	14.3%	M: 3.76

						SD: 0.87
Área 5: Resolución de problemas	4.8%	9.5%	9.5%	61.9%	14.3%	M: 3.71 SD: 0.98

Tabla 3.6. Satisfacción general con el contenido de los ítems de evaluación.

En cuanto al contenido de los ítems de las pruebas, los expertos evaluaron si los elementos de evaluación estaban bien redactados, bien contextualizados y eran fáciles de entender. Su opinión puso de manifiesto la necesidad de mejorar la traducción y la formulación de los ítems. En cuanto a la dificultad y adecuación de los criterios de los ítems, la mayoría pensaba que los criterios eran adecuados aun existiendo diferencias a nivel de prueba, tal y como se muestra en la tabla 3.7.

Prueba	Muy en desacuerdo	En desacuerdo	Ni de acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo	Análisis (M y SD)
Área 1: Alfabetización en información y datos	0.0%	28.6%	4.8%	42.9%	23.8%	M: 3.62 SD: 1.13
Área 2: Comunicación y colaboración en línea	4.8%	4.8%	23.8%	52.4%	14.3%	M: 3.67 SD: 0.94
Área 3: Creación de contenido digital	0.0%	14.3%	19.0%	66.7%	0.0%	M: 3.52 SD: 0.73
Área 4: Seguridad	0.0%	9.5%	19.0%	61.9%	9.5%	M: 3.71 SD: 0.76
Área 5: Resolución de problemas	0.0%	14.3%	19.0%	52.4%	14.3%	M: 3.67 SD: 0.89

Tabla 3.7. Nivel de dificultad y adecuación de los ítems de las pruebas.

Los expertos respondieron positivamente sobre la información proporcionada al final de las pruebas, como se muestra en la tabla 3.8. Hubo una importante cantidad de respuestas neutras que consideramos como un aspecto a mejorar.

Prueba	Muy en desacuerdo	En desacuerdo	Ni de acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo	Análisis (M y SD)
Área 1: Alfabetización en	0.0%	28.6%	4.8%	42.9%	23.8%	M: 3.62

información y datos						SD: 1.13
Área 2: Comunicación y colaboración en línea	4.8%	4.8%	23.8%	52.4%	14.3%	M: 3.67 SD: 0.94
Área 3: Creación de contenido digital	0.0%	14.3%	19.0%	66.7%	0.0%	M: 3.52 SD: 0.73
Área 4: Seguridad	0.0%	9.5%	19.0%	61.9%	9.5%	M: 3.71 SD: 0.76
Área 5: Resolución de problemas	0.0%	14.3%	19.0%	52.4%	14.3%	M: 3.67 SD: 0.89

Tabla 3.8. Nivel de dificultad y adecuación de los ítems de las pruebas.

Además, analizamos la relación entre la satisfacción global con P4E, la satisfacción con las pruebas y los PD seleccionados. Para ello, calculamos el coeficiente de Pearson, obteniendo que no existe una correlación significativa con el PD seleccionado. Además, los resultados muestran que la satisfacción general con P4E correlaciona significativamente con la satisfacción con las pruebas de las AC Área1: *Alfabetización en información y datos* r (21) = .677, p < .01, Área2: *Comunicación y colaboración en línea* r (21) = .566, p < .01, y Área5: *Resolución de problemas* r (21) = .615, como se muestra en la tabla 3.9.

		PD	Satisf. Area1	Satisf. Area2	Satisf. Area3	Satisf. Area4	Satisf. Area5	Satisf. Global
PD	r (21)	1	,440*	,277	,374	,170	,127	,008
	Sig.		,046	,225	,095	,461	,582	,972
Satisf. Area1	r (21)	,440*	1	,747**	,625**	,567**	,617**	,677**
	Sig.	,046		,000	,002	,007	,003	,001
Satisf. Area2	r (21)	,277	,747**	1	,872**	,873**	,902**	,566**
	Sig.	,225	,000		,000	,000	,000	,008
Satisf. Area3	r (21)	,374	,625**	,872**	1	,888**	,857**	,473*
	Sig.	,095	,002	,000		,000	,000	,031
Satisf. Area4	r (21)	,170	,567**	,873**	,888**	1	,954**	,537*
	Sig.	,461	,007	,000	,000		,000	,012
Satisf.	r (21)	,127	,617**	,902**	,857**	,954**	1	,615**

Area5	Sig.	,582	,003	,000	,000	,000		,003
Satisf. Global	r (21)	,008	,677**	,566**	,473*	,537*	,615**	1
	Sig.	,972	,001	,008	,031	,012	,003	

Tabla 3.9. Relación entre la satisfacción con las pruebas de las AC, la satisfacción global de las pruebas y los PD seleccionados. Correlaciones de Pearsons (* p < 0,05 (bilateral); ** p < 0,01 (bilateral)).

A continuación, resumimos las recomendaciones y sugerencias recibidas agrupadas en las cinco categorías que establecimos previamente:

- Lenguaje: algunos ítems debían reformularse de forma más sencilla, manteniendo una terminología coherente.
- Aspectos técnicos: se sugirieron varias mejoras y aspectos de usabilidad en las simulaciones y tareas abiertas.
- Contenido: los criterios de evaluación deben ser más específicos, claros y estar mejor formulados en algunos ítems. Además, algunas preguntas son más fáciles de responder si se está acostumbrado al programa en el que se basa el ítem.
- Estructura: hay que formular las preguntas de forma más breve y sencilla. Además, también es necesario reestructurar la página de resultados que se muestra al final de las pruebas para hacerla más comprensible.
- Diseño: diferenciar claramente las instrucciones, del enunciado y el contexto, para que esté todo más claro de manera homogénea.

Como resultado de esta fase, eliminamos y modificamos los ítems identificados como débiles, y decidimos aplicar la siguiente configuración en las pruebas para cada CD: 3 ítems de conocimientos (1 básico, 1 intermedio y 1 avanzado), 1 ítem de habilidad, y 1 ítem de actitud con varias afirmaciones. Finalmente, contamos con 80 ítems (conocimiento y habilidad) para cada PD, además de los ítems de actitud, que son las que actualmente se encuentran disponibles en P4E.

3.3.4. Pilotajes con usuarios finales y análisis de los resultados

Respecto a la satisfacción general de los encuestados con P4E, el 66% se mostró "muy satisfecho o satisfecho", lo que se consideró un buen resultado general, el 9% no estaba satisfecho y el 25% estaba parcialmente satisfecho. En cuanto a los resultados de las pruebas y la información recibida, el 64% declaró que los resultados de las pruebas eran comprensibles y el 30% afirmó que los resultados eran parcialmente comprensibles. Sólo el 6% contestó que no había entendido los resultados mostrados. Los encuestados sugirieron que se añadiese al informe de resultados las AC que faltan por cubrir o en las que no alcanzan el nivel suficiente, y no sólo los comentarios positivos. Además, el 49% de los encuestados afirmó que la prueba les ayudó a comprender sus carencias y su nivel de CD, y el 31% respondió

que en parte. Sin embargo, el 20% de los encuestados no vio que la herramienta les ayudara en esta cuestión. Como ya comentamos previamente, identificamos como futura línea de trabajo mejorar la información suministrada a los participantes al final de las pruebas para que esta sea lo más clara y práctica posible y les sirva de utilidad para continuar mejorando sus CD.

A nivel de los resultados obtenidos por los participantes, hemos visto como incluso dentro de cada AC, los resultados para cada CD pueden diferir. Esto puede explicarse por el carácter multidimensional de DigComp, en particular por las diferentes tecnologías, áreas de aplicación (escenarios), niveles de dificultad y tipos de preguntas. Los resultados por AC y PD se muestran en las figuras del 3.9 al 3.18.

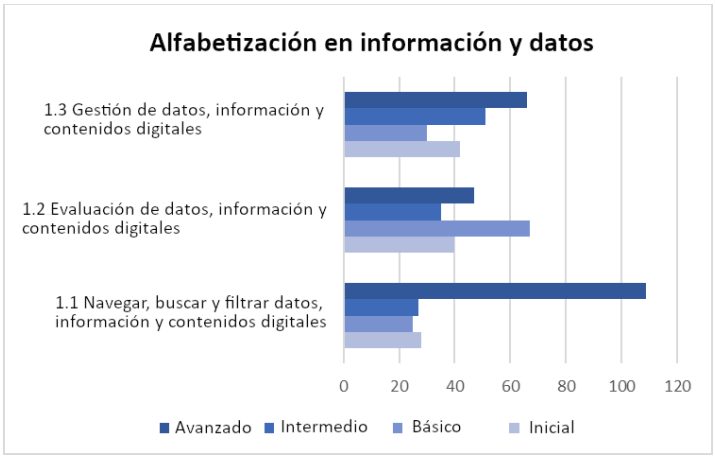


Figura 3.9. Emprendedor - Alfabetización en información y datos (n=189)

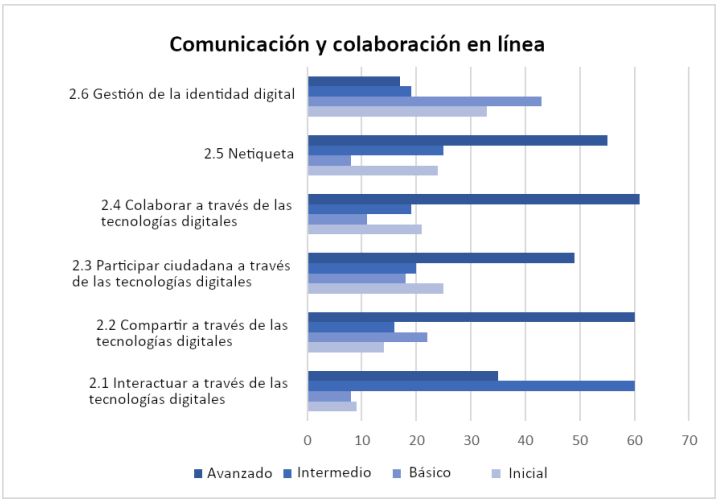


Figura 3.10. Emprendedor – Comunicación y colaboración en línea (n=112)

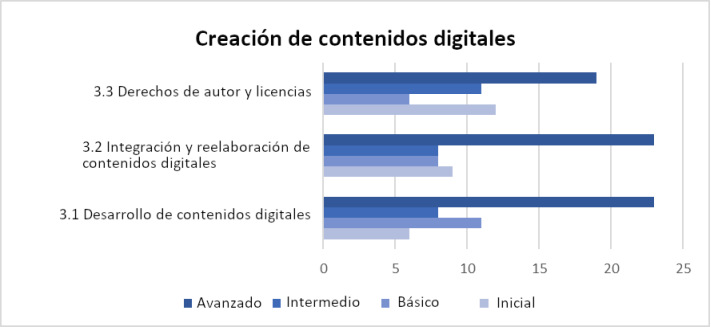


Figura 3.11. Emprendedor – Creación de contenidos digitales (n=48)

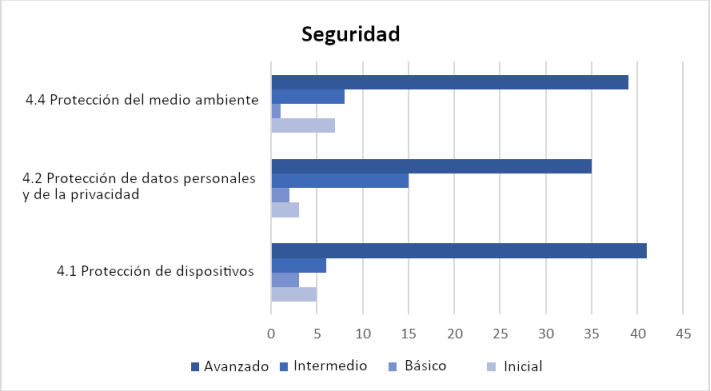


Figura 3.12. Emprendedor - Seguridad (n=55)

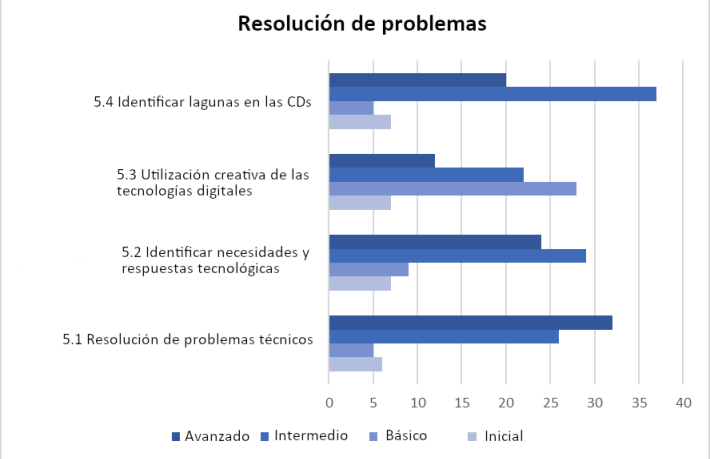


Figura 3.13. Emprendedor – Resolución de problemas (n=69)

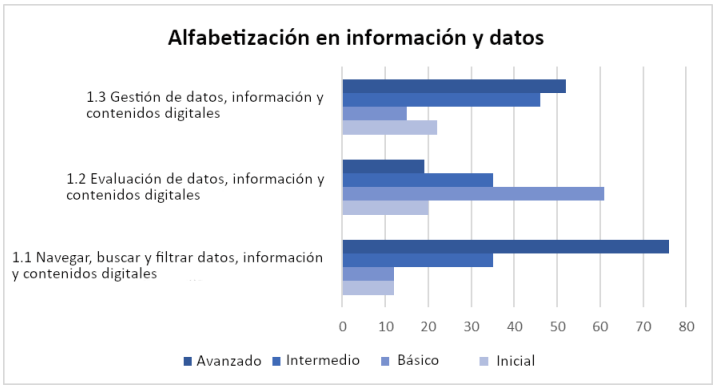


Figura 3.14. Teletrabajador - Alfabetización en información y datos (n=135)

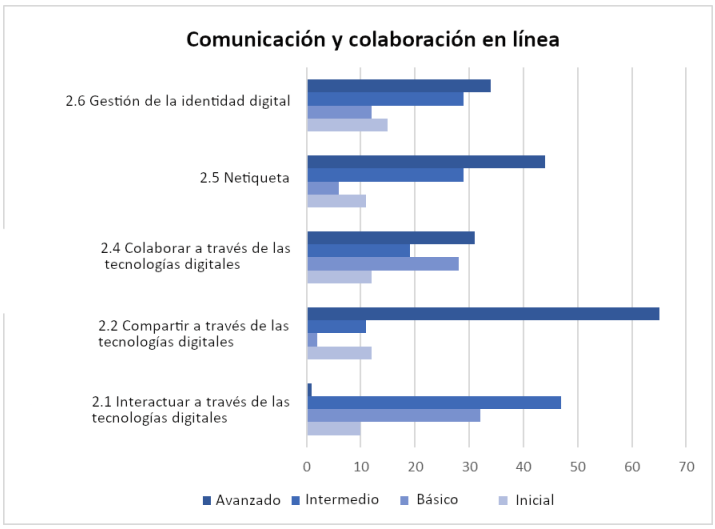


Figura 3.15. Teletrabajador - Comunicación y colaboración en línea (n=90)

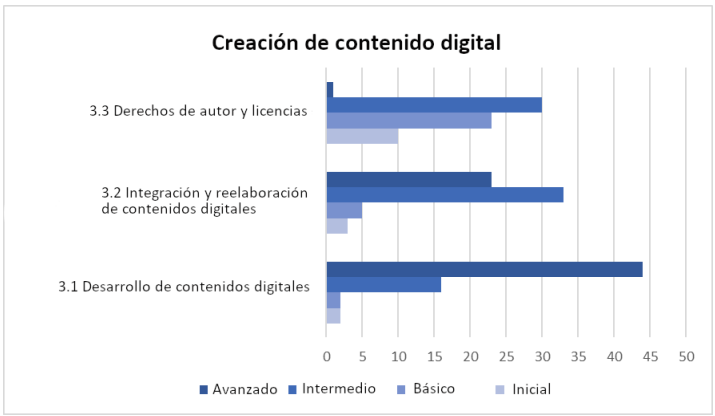


Figura 3.16. Teletrabajador – Creación de contenidos digitales (n=64)

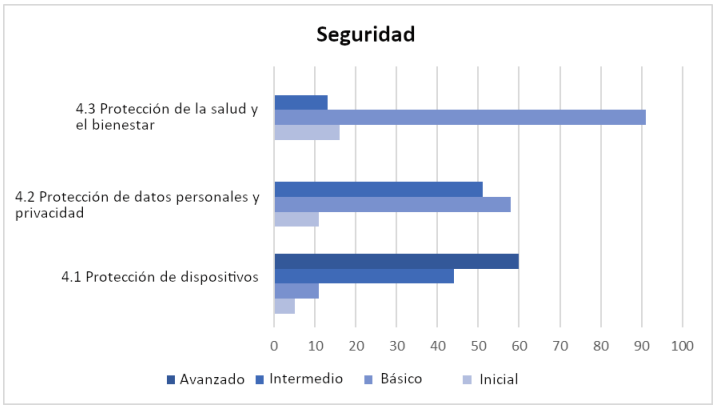


Figura 3.17. Teletrabajador - Seguridad (n=120)

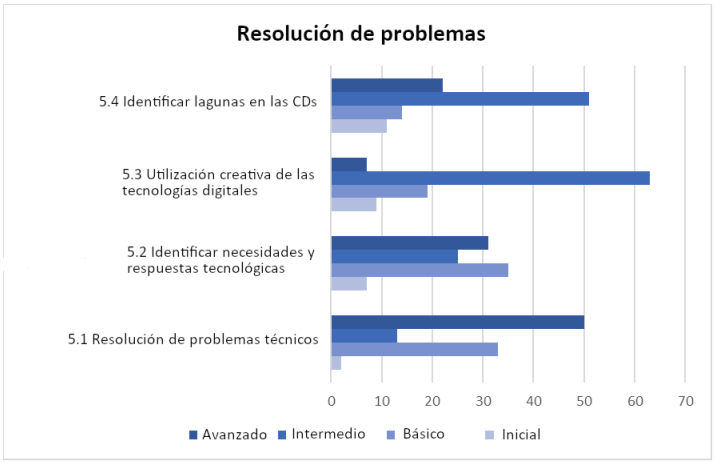


Figura 3.18. Teletrabajador – Resolución de problemas (n=98)

Viendo los resultados obtenidos, dedujimos que las pruebas disponibles en P4E pueden utilizarse para obtener el nivel de CD de los usuarios, ya que existen grupos de usuarios que obtienen resultados mejores o peores que la media con significación estadística. Al examinar los resultados, decidimos modificar los umbrales de las puntuaciones a nivel de CD. En las tablas 3.10 y 3.11 se puede examinar un ejemplo para IAD. Cabe señalar que en las CD 3.1 y 4.3 los ítems de habilidad estaban sesgados y los participantes no pudieron obtener resultados reales. Una vez solucionados los problemas con estos dos ítems, decidimos mantener los intervalos iniciales para estas dos CD y esperar a disponer de nuevos resultados para volver a analizarlos.

CD	1.1	1.2	1.3
Nº de participantes	310	310	310
M	7.34	5.44	7.37
SD	3.28	3.10	3.70

Límite 1 (M -2*SD)	0.77	-0.07	-0.03
Límite 2 (M -1*SD)	4.06	2.33	3.66
Límite 3 (M + 1*SD)	10.63	8.54	11.07
Límite 4 (M + 2*SD)	13.99	11.65	14.77

Tabla 3.10. IAD – análisis estadístico de los resultados.

Nivel de CD	Intervalos iniciales	Rango CD 1.1	Rango CD 1.2	Rango CD 1.3
Inicial	(0-1)	(0-1)	0	0
Básico	(2-3)	(2-3)	(1-2)	(1-3)
Intermedio	(4-6)	(4-10)	(3-9)	(4-10)
Avanzado	(7-11)	11	(10-11)	11

Tabla 3.11. Umbrales de puntuación iniciales y revisados para IAD.

Para examinar la fiabilidad nos basamos en la TCT calculando el alfa de Cronbach para cada prueba, incluyendo otros indicadores como los índices de dificultad, los índices de discriminación, las medias y las desviaciones estándar. Un intervalo ampliamente aceptado para describir la consistencia interna mediante el alfa de Cronbach es el siguiente: >= 0,9 excelente; >=0,8 y <0,9 bueno; >=0,7 y <0,8 aceptable; >=0,6 y <0,7 cuestionable; >=0,5 y <0,6 pobre; > 0,5 inaceptable. Como se muestra en la tabla 3.12, antes de aplicar los cambios y las correcciones, dos de las pruebas ya eran aceptables.

	Alfabetización en información y datos	Comunicación y colaboración en línea	Creación de contenido digital	Seguridad	Resolución de problemas
Alpha de Cronbach	0.64	0.73	0.70	0.59	0.61
Otros indicadores	Nº Items: 12 Respuestas: 310 M: 8,0452 SD: 2,3372	Nº Items: 24 Respuestas: 209 M: 16,5120 SD: 3,7762	Nº Items: 12 Respuestas: 159 M: 7,4717 SD: 2,3839	Nº Items: 12 Respuestas: 310 M: 8,0452 SD: 2,3372	Nº of items: 15 Respuestas: 150 M: 11,5933 SD: 2,2423

Tabla 3.12. Alpha de Cronbach para las 5 pruebas del emprendedor.

En la tabla 3.13 se muestran los principales índices de dificultad y discriminación calculados para los ítems del PD del emprendedor.

Alfabetización en información y datos			Comunicación y colaboración en línea			Creación de contenido digital			Seguridad			Resolución de problemas		
Ítem	Dif.	Disc.	Ítem	Dif.	Disc.	Ítem	Dif.	Disc.	Ítem	Dif.	Disc.	Ítem	Dif.	Disc.
C1	0,87	0,24	C1	0,81	0,23	C1	0,88	0,24	C1	0,66	0,11	C1	0,79	0,43
C2	0,77	0,30	C2	0,86	0,22	C2	0,48	0,20	C2	0,87	0,31	C2	0,98	0,10
C3	0,55	0,14	C3	0,93	0,41	C3	0,76	0,53	C3	0,82	0,36	C3	0,92	0,42
S1	0,65	0,26	S1	0,43	0,32	S1	0	0	S1	0,72	0,38	S1	0,61	0,31
C4	0,85	0,31	C4	0,50	0,26	C4	0,76	0,35	C4	0,87	0,18	C4	0,99	0,26
C5	0,63	0,15	C5	0,94	0,17	C5	0,84	0,29	C5	0,88	0,09	C5	0,95	0,45
C6	0,26	0,02	C6	0,61	0,35	C6	0,90	0,28	C6	0,93	0,33	C6	0,87	0,35
S2	0,51	0,44	S2	0,61	0,36	S2	0,67	0,58	S2	0,89	0,29	S2	0,58	0,46
C7	0,92	0,41	C7	0,67	0,30	C7	0,71	0,28	C7	0,64	0,10	C7	0,21	-0,25
C8	0,72	0,38	C8	0,45	0,19	C8	0,67	0,44	C8	0,81	0,17	C8	0,98	0,25
C9	0,73	0,31	C9	0,77	0,06	C9	0,50	0,35	C9	0,31	0,06	C9	0,62	0,15
S3	0,56	0,50	S3	0,73	0,25	S3	0,30	0,31	-	-	-	S3	0,14	0,04
			C10	0,94	0,18				C10	0,87	0,31	C10	0,92	0,31
			C11	0,87	0,26				C11	0,88	0,25	C11	0,66	0,19
			C12	0,63	0,50				C12	0,84	0,33	C12	0,91	0,09
			S4	0,88	0,45				S4	0,62	0,27	S4	0,47	0,46
			C13	0,78	0,44									
			C14	0,83	0,25									
			C15	0,68	0,24									
			S5	0,83	0,44									
			C16	0,35	0,18									
			C17	0,54	0,15									
			C18	0,71	0,32									
			S6	0,27	0,13									

Tabla 3.13. Índices a de dificultad y discriminación de los ítems de las pruebas del emprendedor (En rojo los ítems que no cumplen los mínimos requeridos).

A partir de los resultados podemos ver cómo los ítems de habilidad (Sx) muestran mejores índices de discriminación que los ítems de conocimiento (Cx). Además, en algunas CD hay algunos ítems con niveles de dificultad más difíciles o fáciles de lo esperado, de acuerdo con el nivel asignado inicialmente. De acuerdo con los estándares internacionales, la distribución de los resultados de los coeficientes de discriminación es < 0 negativo, (0 - 0,14) malo, (0,15 - 0,25) regular, (0,26 - 0,35) bueno, y > 0,35 excelente. Realizamos los siguientes cambios:

- Corregimos los problemas técnicos de S1 en el área de Creación de Contenidos Digitales.
- El ítem C7 en el área de Resolución de problemas fue sustituido porque estaba mal redactado y era muy difícil de entender.
- 9 ítems mostraron índices de discriminación malos (entre 0.00 y 0.14), de los cuales sólo el S6 del área Comunicación y colaboración en línea, y el S3 del área Resolución de problemas eran ítems de habilidad. Concretamente, los problemas se debían a que los pasos de las simulaciones no eran claros y tuvimos que rediseñarlos de nuevo.
- En cuanto a las preguntas de conocimiento, se reformularon para mejorar la comprensión y se examinaron las respuestas disponibles para detectar opciones con pocas posibilidades de ser elegidas.
- 15 ítems mostraron coeficientes de discriminación regulares (entre 0.15 y 0.25), y todos ellos eran preguntas de conocimiento. Se revisaron para mejorar la comprensión de las preguntas y las posibles respuestas, sustituyendo las opciones con bajas posibilidades de ser seleccionadas.
- Por último, 20 y 21 ítems mostraron índices de discriminación clasificados como buenos (entre 0.26 y 0.35) y excelentes (entre 0.36 y 0.58) respectivamente. Estos ítems se mantuvieron sin aplicar cambios. Salvo 3 simulaciones que se modificaron, 6 se identificaron como buenos discriminadores y 9 como excelentes discriminadores.

De acuerdo con el enfoque adoptado, decidimos modificar o sustituir los ítems débiles en lugar de simplemente eliminarlos para mejorar la consistencia interna, con el fin de mantener la estructura y la comprensión de las pruebas.

3.4. Conclusiones y discusión

Los resultados presentados en este capítulo ayudan a entender los pasos llevados a cabo y las consideraciones y decisiones que se han tomado en cada paso. Pueden servir de consideración para diseñar y validar una herramienta de evaluación y acreditación de PD, llevando a cabo una implementación personalizada basada DigComp. Revisamos las conclusiones asociadas a nuestro objetivo de investigación, al cual seguimos dándole respuesta a lo largo del capítulo 3.4.

PI_K1. *¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?*

A lo largo de este capítulo hemos mostrado los pasos que llevamos a cabo haciendo uso de una metodología mixta (cuantitativa y cualitativa) para el diseño y validación de una herramienta de evaluación y acreditación de PD. Decidimos llevar a cabo la implementación siguiendo un enfoque pragmático, que facilitase el diseño e implementación, a pesar de que la fiabilidad y la validez de una prueba basada en el rendimiento es un gran reto, y aún más cuando la prueba mide un constructo complejo como la CD (Laanpere, 2019). Por tanto, decidimos evaluar cada CD como un constructo independiente, sin tener en cuenta los puntos de solapamiento a nivel de CD identificados en DigComp. Con este enfoque, priorizamos la validez externa de la herramienta, pero nos exigió detallar todos los pasos y las decisiones tomadas para ser mejor comprendidos, aceptados y utilizados por un público más amplio (Van Deursen et al., 2016).

Incorporamos DigComp en la definición de los PD y, posteriormente, en el diseño y desarrollo de las pruebas de evaluación. En el desarrollo del banco de ítems fue crucial su uso para identificar el contenido requerido y asignar los niveles de los ítems según los descriptores. Los descriptores de DigComp se han definido utilizando la taxonomía de Bloom, por lo que analizamos los enunciados de los ítems, identificamos los verbos utilizados y propusimos un nivel de acuerdo con el marco de referencia. Más aún, tal y como pudimos constatar de los comentarios y sugerencias recibidos de los expertos, tuvimos que prestar especial atención a los ítems que requerían una observación profunda y exigían esfuerzos de comprensión lectora, ya que la complejidad lingüística es un factor clave, tal y como Plath y Leiss (2018) indicaron, mostrando como el aumento de la complejidad lingüística de las tareas reducía la tasa de éxito de las respuestas.

La adopción de DigComp, nos proporcionó criterios claros y bien definidos que se adaptaron a nuestras necesidades en todas las fases del desarrollo. Además, la credibilidad y fiabilidad del marco debido a su origen y aval de la UE, contribuyó a la difusión de los resultados. Una de las conclusiones a las que se llegó fue que DigComp apenas era conocido en el ámbito profesional cuando empezó el proyecto, aspecto que ha cambiado notablemente los últimos años. A pesar de este hecho, la adopción de DigComp contribuyó a crear un entendimiento común para debatir sobre la CD entre los diferentes actores implicados en el proceso, como un componente transversal necesario para una gran variedad de PD. Esta contribución fue esencial para la evaluación positiva recibida sobre las pruebas de evaluación, el contenido de los ítems, su nivel de dificultad y adecuación, y la información proporcionada al final de las pruebas.

Además, hay que tener en cuenta algunas limitaciones a la hora de interpretar los resultados de este estudio. Por ejemplo, sería recomendable elaborar un banco de ítems más grande. De hecho, aumentar el número de ítems de un test, es una forma habitual de mejorar la fiabilidad. Más aún, desde el principio tuvimos claro

que no podíamos esperar obtener un mayor grado de fiabilidad y precisión debido a las decisiones tomadas, como elaborar un número corto de ítems por CD. A pesar de esta limitación, consideramos que las decisiones derivadas de los resultados no eran tan críticas como para buscar un mayor grado de precisión y fiabilidad.

Del análisis de los ítems pudimos comprobar que los ítems identificados como de habilidad mostraban mejores índices de discriminación que los ítems de conocimiento, lo cual está alineado con las puntuaciones que asignamos a cada tipo de pregunta. Desafortunadamente, el desarrollo de este tipo de ítems (simulaciones o tareas reales) requiere más tiempo y esfuerzo. También pudimos comprobar que no conseguimos alinear correctamente los niveles de algunos ítems en varias CD (por ejemplo, un ítem del nivel intermedio parecía más difícil que un ítem del nivel avanzado). Engelhardt et al. (2017) mostraron cómo cambiando ciertas características del ítem es posible influir en la dificultad de este sin modificar el constructo representado (por ejemplo, podríamos disminuir el número de intentos permitidos en una simulación para aumentar su nivel de dificultad). Además, si tuviésemos un banco de ítems suficientemente grande, podríamos seleccionar los ítems con los coeficientes de discriminación más adecuados según los niveles requeridos.

Como ya mencionamos previamente, si buscamos alcanzar una visión más amplia no sólo centrada en la evaluación objetiva de las CD, deberíamos mejorar la presentación de los resultados a los participantes al finalizar las pruebas. Por ejemplo, podríamos enriquecer los resultados añadiendo los conocimientos y habilidades necesarios para pasar al siguiente nivel en las CD que no completó.

Asimismo, hay que mencionar que tratamos de diseñar los ítems de las pruebas no basándonos en aplicaciones o dispositivos específicos como sugiere Law et al. (2018). Sin embargo, las preguntas de simulación las tuvimos que diseñar seleccionando una aplicación, así que tratamos de seleccionar las herramientas más utilizadas con comportamientos ampliamente aceptados en términos de usabilidad. No pretendimos evaluar las propias herramientas per se.

Otro punto que merece la pena mencionar es que los expertos que conocían DigComp tenían serias dudas sobre ciertas CD como p. ej. Programar. Esta falta de claridad podría ser una de las razones por las que la CD de programación no fue seleccionada como necesaria en ninguno de los PD. De hecho, en DigComp 2.2, Vuorikari et al. (2022) trataron de esclarecer este punto incorporando nuevos ejemplos de conocimientos, habilidades y actitudes.

En futuros estudios se podría considerar la identificación de SCs tal y como proponen en IKANOS, identificando tareas de trabajo específicas relevantes para una CD concreta, que deben destacarse en un PD.

Este capítulo contribuye a la definición y evaluación de la CD de los PD aportando resultados empíricos. Hemos seguido un enfoque pragmático basado en DigComp que hace uso de un lenguaje y unos criterios comunes que han facilitado el

desarrollo de una herramienta como la presentada aquí. Esperamos que las explicaciones sobre los pasos y las decisiones tomadas puedan servir de inspiración a los interesados en llevar a cabo sus propias implementaciones.

La evaluación es el motor del aprendizaje, ya que de ella depende tanto qué y cómo se enseña, cómo el qué y el cómo se aprende.

Neus Sanmamrtí

4.

Herramienta ETCD

Después de examinar el creciente número y variedad de herramientas existentes para evaluar la CD en un contexto general orientado a la ciudadanía, y cómo se ha llevado a cabo la documentación de las propiedades psicométricas de las pruebas, consideramos que era importante diseñar una nueva herramienta de evaluación de CD abordando las principales carencias detectadas.

Por este motivo, comenzamos en 2019 el desarrollo en una herramienta de evaluación de CD que posibilitara el registro detallado de todas las interacciones realizadas por los participantes durante la realización de las pruebas, para poder analizarlas con posterioridad y nutrir los diferentes estudios llevados a cabo como parte del proceso de validación de las pruebas. Llevamos a cabo el desarrollo de la herramienta siguiendo una metodología de investigación basada en el diseño (DBR) siguiendo diferentes fuentes de información. La metodología seguida nos sirve para describir los principios de diseño aplicados durante los diferentes pasos del desarrollo de las pruebas de CD seleccionadas, con el fin de que puedan ser extendidos al resto de CD de DigComp, y para recabar las evidencias de validez y fiabilidad que nos ayuden a soportar la calidad de la herramienta desarrollada.

4.1. Objetivos de investigación

Con el objetivo de dar respuesta a las siguientes preguntas de investigación, diseñamos y desarrollamos ETCD, la cual propusimos a dinamizadores de KZgunea, empleados de Tecnalia y público general, es decir, ciudadanos, y analizamos sus registros:

- PI_K1. *¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?*

- PI_K2. *¿Qué propiedades psicométricas tienen las pruebas? ¿Qué evidencias se pueden presentar que soporten las inferencias realizadas de las puntuaciones obtenidas?*

Los siguientes dos objetivos, que podrían ir enmarcados en el objetivo PI_K2 y viendo la relevancia que han tenido nuestra investigación, los describimos en detalle en el capítulo 5:

- PI_K3. *¿Permite el análisis de las interacciones del examinando durante la prueba entender su comportamiento, de manera que podamos generar y probar inferencias sobre el constructo de interés?*
- PI_K4. *¿Permite el análisis de las interacciones del examinando durante la prueba entender su comportamiento, de manera que podamos utilizar esta información para mejorar los diseños de los ítems?*

4.2. Metodología

Para responder a nuestras preguntas de investigación, desarrollamos una herramienta de evaluación de la CD orientada a la ciudadanía seleccionando dos casos de estudio: la Información y Alfabetización Digital (IAD) y la netiqueta.

Los casos de estudio seleccionados representan dos enfoques diferentes que están siendo adoptados por algunas de las iniciativas relevantes identificadas como casos de éxito en el estudio de implementaciones basadas en DigComp llevado a cabo por Kluzer y Priego (2018). Los autores identificaron dos enfoques habitualmente adoptados: pruebas basadas en un AC o pruebas basadas en una CD. Para la selección del AC, elegimos una de las tres principales AC de DigComp, *IAD*, pero se podría haber elegido perfectamente otra AC como la *Comunicación y la Colaboración en línea* debido a su relevancia. Para seleccionar una CD, nos encontramos en la misma situación y elegimos una CD que no suele evaluarse en profundidad, normalmente sólo habilidades de orden cognitivo bajo. Más aun, seleccionamos Netiqueta porque los resultados de incluir formatos de ítems dinámicos serían más notables que en otras CD.

Aplicamos una DBR que se basó en el análisis de diferentes fuentes de información para llevar a cabo el desarrollo de la herramienta de evaluación y su posterior validación. La DBR es una metodología muy utilizada en las ciencias del aprendizaje para analizar el desarrollo de soluciones como la que presentamos (Herrington et al., 2007; McKenney y Reeves, 2018; Sandoval, 2014).

En nuestra investigación hemos combinado diferentes métodos durante el proceso de diseño iterativo del instrumento de evaluación para dos implementaciones. Según Reeves (2006), cada ciclo consta de diferentes fases (véase la figura 4.1).

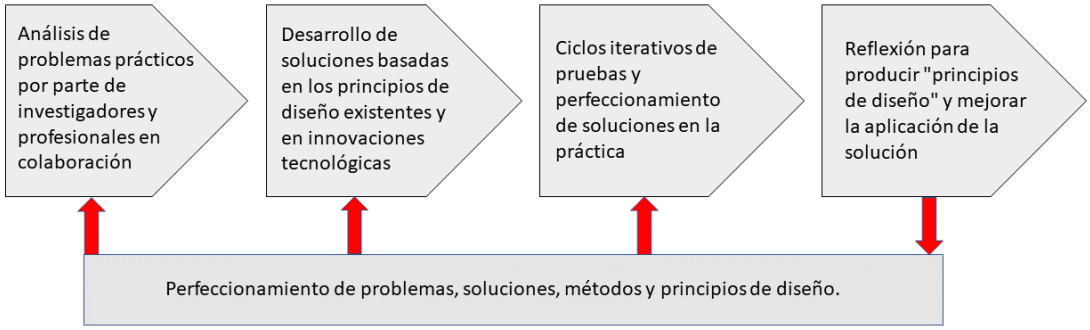


Figura 4.1. Enfoque DBR en la investigación tecnológica según Reeves (2006).

En resumen, pretendemos dar respuesta a las preguntas de investigación PI_K1 y PI_K2, con los siguientes objetivos:

- Diseñar una herramienta para la evaluación de la CD que admita formatos dinámicos como las simulaciones interactivas, que son especialmente relevantes a la hora de medir constructos cognitivos complejos como la CD.
- Describir los principios de diseño aplicados durante los diferentes pasos del desarrollo de las pruebas con el fin de que puedan extenderse al resto de las CD incluidas en DigComp.

El enfoque específico que seguimos se guio por los diferentes elementos, directrices y consideraciones sugeridas por Herrington et al. (2007) para cada fase de la DBR propuesta por Reeves (2006) (ver tabla 4.1).

Fase	Elemento
FASE 1: análisis del problema por parte de investigadores y profesionales en colaboración	• Planteamiento del problema • Consulta con investigadores y profesionales • Preguntas de investigación • Revisión de la literatura
FASE 2: desarrollo de soluciones marco teóricas basadas en los principios de diseño existentes y en innovaciones tecnológicas	• Marco de referencia teórico • Elaboración de un borrador de principios para guiar el diseño de la solución • Descripción de la solución propuesta

FASE 3: ciclos iterativos de pruebas y perfeccionamiento de la solución en la práctica	• Implementación de la intervención (primera iteración con los facilitadores del centro de CD y segunda iteración con los ciudadanos) • Participantes • Recogida de datos • Análisis de los datos
FASE 4: reflexión para producir "principios de diseño" y mejorar la aplicación de la solución	• Principios de diseño • Artefacto diseñado

Tabla 4.1. Fases de la metodología DBR basada en el trabajo de Herrington et al. (2007) adaptada a nuestra propuesta de investigación.

En este capítulo, describimos nuestro enfoque específico y los resultados de cada fase, incluidos dos ciclos iterativos para la fase 3. Presentamos los resultados de los dos ciclos interactivos en la sección Resultados. Cabe comentar que llevamos a cabo las distintas fases en orden, aunque algunas de ellas podían gestionarse simultáneamente.

4.2.1. Análisis del problema por parte de investigadores y profesionales en colaboración

Al principio de la fase 1, planteamos el problema. Para ello, examinamos las soluciones disponibles realizando una revisión de la literatura, como hemos mostrado a lo largo del capítulo 2. Identificamos una falta de instrumentos adecuados para evaluar la CD de las personas. Es necesario incluir en las pruebas formatos de ítems innovadores y más cercanos a la práctica real que permitan evaluar las habilidades de orden superior según los niveles medio y avanzado de DigComp. De lo contrario, se evalúa principalmente el componente de conocimiento de la CD. El propósito de nuestra investigación está guiado por este problema identificado, y el objetivo es el desarrollo y la validación de una posible solución.

Tras la revisión de los conocimientos y las prácticas actuales, definimos nuestros objetivos basándonos en los principios de diseño identificados en varios estudios clave. Paralelamente, hemos estado participando en *The DigComp Community of Practice* (DigComp CoP)[48], que fue lanzado a finales de 2019 por All Digital[49] para promover la adopción y apoyar el desarrollo de DigComp. Hemos participado en

[48] https://all-digital.org/invitation-to-digcomp-cop/

[49] https://all-digital.org/

los grupos de trabajo, intercambiando material y experiencia, accediendo a buenas prácticas, e informándonos sobre los últimos avances respecto a DigComp y las implementaciones que se han llevado a cabo.

4.2.2. Desarrollo de soluciones marco teóricas basadas en los principios de diseño existentes y en innovaciones tecnológicas

El diseño de una intervención basada en una comprensión detallada del problema se guía por principios de diseño que son argumentos teóricos prescriptivos (McKenney y Reeves, 2018). Por lo tanto, definimos una solución inicial basada en los estudios clave identificados (BAIT; Bartolomé y Garaizar, 2022; Bartolomé, Garaizar y Larrucea, 2022) y en otros estudios seleccionados de la revisión bibliográfica (Law et al., 2018; Reichert et al., 2020; Santos y Serpa, 2017; Siddiq et al., 2016). Una vez establecida la nueva propuesta de diseño, es necesario examinarla y mejorarla tras las pruebas y el análisis (McKenney y Reeves, 2018).

Para el diseño de la herramienta de evaluación, tomamos DigComp 2.1 como marco de referencia (Carretero et al., 2017). DigComp ofrece una visión clara de los diferentes componentes de la CD (conocimientos, habilidades y actitudes) a la hora de utilizar diferentes dispositivos y servicios digitales, que son necesarios para lograr una participación plena en nuestra sociedad y que pueden adaptarse a muchos ámbitos de la vida. En concreto, nos centramos en la "*mejora de la empleabilidad*" como escenario de aplicación, ya que este estudio está estrechamente relacionado BAIT y es su escenario de aplicación. Inicialmente, seleccionamos 4 CD como casos de estudio y los seis primeros niveles según DigComp 2.1 para ser implementados y evaluados (básico, intermedio y avanzado). Se puede considerar que estos son los niveles de CD más demandados por los ciudadanos para mejorar su empleabilidad. Asimismo, consideramos cada CD como un constructo independiente y desarrollamos una prueba para cada una de ellas con el fin de que se midieran de forma independiente. A partir de la revisión de la literatura, identificamos una serie de subcompetencias (SC) que decidimos incluir en cada prueba (véase la tabla 4.2). Los descriptores definidos para cada CD, las SC y los niveles correspondientes pueden examinarse en la tabla 4.3. A continuación, partiendo de estos descriptores, desarrollamos los ítems de evaluación.

AC	CD	SC
Comunicación y colaboración en línea	Netiqueta	SC1: aplicar las pautas básicas de Netiqueta al utilizar el correo electrónico (p.ej., uso de la copia oculta (CCO), reenviar un correo o contenido, etc.).
		SC2: aplicar normas de escritura en línea sencillas (no usar mayúsculas, respetar la ortografía, referirse a los demás por sus

		alias o apodos, etc.) y utilizar adecuadamente los emoticonos al comunicarse en línea.
		SC3: reconocer los comportamientos adecuados en redes sociales, como recibir el permiso de los demás antes de publicar (especialmente cuando hay niños de por medio); evitar el SPAM (p.ej., enviar invitaciones u otros mensajes a todo el mundo); evitar utilizar palabras o lenguaje poco claro que pueda ser malinterpretado.
		SC4: reconocer comportamientos inapropiados en línea, como el acoso, el trolling o el ciberacoso. Ser capaz de hacer frente a los comportamientos incorrectos, como señalar las publicaciones irrespetuosas o notificar a la policía.
IAD	Navegar, buscar y filtrar datos, información y contenidos digitales	SC5: analizar las necesidades de información, buscar datos e información en entornos digitales, filtrar y localizar.
		SC6: definir la estrategia de búsqueda necesaria en cada momento.
	Evaluación de datos, información y contenidos digitales	SC7: examinar y evaluar la credibilidad y fiabilidad de las fuentes de datos e información.
		SC8: examinar y evaluar los contenidos, datos e información digitales.
	Gestión de datos, información y contenidos digitales	SC9: organizar, almacenar y procesar datos, información y contenidos en entornos digitales.

Tabla 4.2. Casos de estudio seleccionados basados en DigComp y SCs seleccionadas.

(AC) CD: SCs	Nivel Básico	Nivel Intermedio	Nivel Avanzado
(IAD) Navegar, buscar y filtrar datos, información y contenidos digitales: • Articular las necesidades de información, buscar datos, información y contenidos en entornos digitales, acceder a ellos y	En el nivel básico y con autonomía y orientación adecuada cuando sea necesario, puedo: • Identificar mis necesidades de información. • Encontrar datos, información y contenidos a través	De forma independiente, según mis propias necesidades, y resolviendo problemas bien definidos y no rutinarios, puedo: • Ilustrar las necesidades de información: Identificar diferentes tipos de fuentes y	En un nivel avanzado, según mis propias necesidades y las de los demás, y en contextos complejos, puedo: • Evaluar las necesidades de información. • Adaptar mi estrategia de

navegar por ellos. • Crear y actualizar estrategias de búsqueda propias.	de una búsqueda sencilla en entornos digitales, también en mí móvil utilizando una app o no. Identificar el tipo de búsqueda necesaria. • Encontrar cómo acceder a estos datos, información y contenidos y navegar entre ellos, también en mi móvil utilizando una app o no. • Navegar en un sitio web para acceder a la información requerida o realizar la acción solicitada. • Identificar estrategias sencillas de búsqueda personal. • Seleccionar la fuente adecuada e identificar las palabras clave que son útiles. Uso de marcadores para almacenar las URL.	motores de búsqueda; identificar diferentes áreas en una página de resultados de búsqueda; uso de la navegación privada; y acceso a marcadores. • Organizar las búsquedas de datos, información y contenidos en entornos digitales: Guardar un vídeo para verlo más tarde. • Describir cómo acceder a estos datos, información y contenidos, y navegar entre ellos: Información en páginas web; vídeos en YouTube; productos en tiendas online; noticias; y redes sociales. • Organizar estrategias de búsqueda personales: utilizando palabras clave y encontrar aplicaciones adecuadas en mi móvil que se ajusten a mis necesidades; uso de diferentes tipos de filtros para buscar y filtrar información, búsquedas basadas	búsqueda para encontrar datos, información y contenidos más adecuados en entornos digitales: Uso de los filtros de búsqueda avanzada; uso de las búsquedas rápidas de Google; y búsqueda de información en una hoja de cálculo. • Explicar cómo acceder a los datos, información y contenidos más adecuados y navegar entre ellos. • Variar las estrategias de búsqueda personales: Acceso a los marcadores y contraseñas almacenados en su cuenta de Google desde diferentes dispositivos.

		en imágenes, búsquedas en Google Maps.	
(IAD) Evaluación de datos, información y contenidos digitales: • Analizar, comparar y evaluar críticamente la credibilidad y fiabilidad de las fuentes de datos, información y contenidos digitales. • Analizar, interpretar y evaluar críticamente datos, información y contenidos digitales.	A nivel básico y con autonomía y orientación adecuada cuando sea necesario, puedo: • Detectar la credibilidad y fiabilidad de las fuentes habituales de datos, información y contenido digital: elementos clave en sitios web; correos electrónicos; tiendas online; y redes sociales como Twitter, YouTube, WhatsApp).	De forma independiente, según mis propias necesidades, y resolviendo problemas bien definidos y no rutinarios, puedo: • Realizar el análisis, comparación y evaluación de fuentes de datos, información y contenidos digitales: sitios web, correos electrónicos, tiendas online y redes sociales. • Realizar el análisis, la interpretación y la evaluación de datos, información y contenidos digitales: noticias, publicaciones en redes sociales como Twitter y YouTube, e información compartida a través de WhatsApp.	En un nivel avanzado, según mis propias necesidades y las de los demás, y en contextos complejos, puedo: • Evaluar críticamente la credibilidad y fiabilidad de las fuentes de datos, información y contenidos digitales: páginas web, correos electrónicos, tiendas online y redes sociales. • Evaluar críticamente datos, información y contenidos digitales: noticias, publicaciones en redes sociales como Twitter y YouTube, e información compartida vía WhatsApp.

(IAD) Gestión de datos, información y contenidos digitales: • Organizar, almacenar y recuperar datos, información y contenidos en entornos digitales. • Organizarlos y procesarlos en un entorno estructurado.	A nivel básico y con autonomía y orientación adecuada cuando sea necesario, puedo: • Identificar cómo organizar, almacenar y recuperar datos, información y contenidos de forma sencilla en entornos digitales: entornos basados en Windows: carpetas, ordenar por, abrir con, etc. • Reconocer dónde organizarlos de forma sencilla en un entorno estructurado.	De forma independiente, según mis propias necesidades, y resolviendo problemas bien definidos y no rutinarios, puedo: • Organizar la información, datos y contenidos para almacenarlos y recuperarlos fácilmente: gestionar los marcadores y los archivos adjuntos en correos electrónicos. • Organizar información, datos y contenidos en un entorno estructurado: gestionar la información en aplicaciones y crear un documento en Google Docs para compartirlo con otros.	En un nivel avanzado, según mis propias necesidades y las de los demás, y en contextos complejos, puedo: • Adaptar la gestión de información, datos y contenidos para facilitar su recuperación y almacenamiento. • Adaptarlos para que se organicen y procesen en el entorno estructurado más adecuado.

(Comunicación y colaboración en línea) Netiqueta: • Aplicar los principios básicos de la Netiqueta en el uso del correo electrónico: uso del CCO, reenvío, etc. • Aplicar las normas básicas de escritura en línea (no usar mayúsculas, cuidar la ortografía, referirse a los demás por sus alias o apodos, etc.) y utilizar adecuadamente los emoticonos al comunicarse en línea. • Reconocer comportamientos adecuados que se deben adoptar en redes sociales, como pedir permiso antes de publicar o compartir fotos de otras personas, evitar el SPAM, utilizar con cuidado el sarcasmo, la ironía o palabras que puedan ser malinterpretadas por los demás. • Reconocer comportamientos social y éticamente inapropiados, como	A nivel básico y con autonomía y orientación adecuada cuando sea necesario, puedo: • Diferenciar normas de comportamiento sencillas y saber cómo utilizar las tecnologías digitales e interactuar en entornos digitales: al utilizar el correo electrónico, etiquetar a personas en una foto en las redes sociales, etc. • Elegir modos y estrategias de comunicación sencillos y adaptados a un público: identificar los malos comportamientos, cómo responder a situaciones desagradables en redes sociales, etc.	De forma independiente, según mis propias necesidades, y resolviendo problemas bien definidos y no rutinarios, puedo: • Discutir las normas de comportamiento y los conocimientos prácticos durante el uso de las tecnologías digitales y la interacción en entornos digitales: correo electrónico, medios sociales como Twitter, participación en foros, etc. • Discutir las estrategias de comunicación adaptadas a un público concreto: correos electrónicos formales/informales, participación en foros y en las redes sociales.	En un nivel avanzado, según mis propias necesidades y las de los demás, y en contextos complejos, puedo: • Adaptar las normas de comportamiento y los conocimientos técnicos más adecuados al utilizar las tecnologías digitales e interactuar en entornos digitales: correo electrónico, medios sociales como Twitter, participación en foros, etc. • Adaptar las estrategias de comunicación más adecuadas en entornos digitales a una audiencia específica: correos electrónicos formales/informales, participación en foros y en medios sociales, etc. • Aplicar diferentes aspectos de diversidad cultural y generacional en entornos digitales.

la incitación al odio, el trolling, el ciberacoso, etc., y saber utilizar métodos básicos para hacer frente a las interacciones negativas, p.ej., señalando las publicaciones a los propietarios de los servicios, a la policía, etc.).			

Tabla 4.3. Descriptores definidos para cada CD, SC y niveles correspondientes.

Los ítems se distribuyeron en cada SC de forma no uniforme, es decir, consideramos que algunas SC requerían más ítems para ser evaluadas correctamente. Aunque las CD descritas en DigComp pueden parecer estables a corto plazo, en el contexto actual, en el que la tecnología continuamente provoca profundos cambios de hábitos en las personas, la construcción de las CD requiere una revisión constante (Aesaert et al., 2014; Siddiq et al., 2016). Por ejemplo, en la selección de SC para el CD de Netiqueta, no incluimos nada relacionado con su aplicación en videoconferencias. Meses más tarde, debido a la pandemia, estas cuestiones pasaron a ser relevantes en este ámbito debido a un aumento sin precedentes del uso de este tipo de herramientas.

Cabe mencionar que, respecto a los tres componentes de la CD, optamos por excluir del ámbito de nuestro estudio la evaluación del componente actitudinal. El componente actitudinal es complejo y no hay consenso sobre cómo evaluarlo y, además, no se va a evaluar directamente en BAIT.

A continuación, de acuerdo con el análisis realizado durante la revisión de la literatura, decidimos aplicar de un enfoque basado en el rendimiento, en el que los participantes son monitorizados en un entorno de evaluación por ordenador (CBA). Diseñamos los ítems para evaluar si los participantes son capaces de entender cualquier entorno digital de forma efectiva en lugar de evaluar sus conocimientos sobre aplicaciones específicas. Los participantes tienen que poner en práctica sus conocimientos y de esta manera, se pueden activar y evaluar las habilidades cognitivas de orden superior. Esta forma de evaluar permite obtiene la situación más realista de los niveles de CD de los participantes.

Con este objetivo, diseñamos una herramienta de evaluación web online siguiendo la misma arquitectura que BAIT para facilitar la transferencia de resultados. Otros aspectos que tuvimos que considerar durante el diseño fueron: el modo de administración de la prueba (bajo condiciones controladas), la cantidad y el tipo de contenido y preguntas (se necesitaría un número significativo de preguntas de conocimiento y habilidades para evaluar las SC seleccionadas para cada CD) y el tiempo necesario para realizar cada prueba. También decidimos incluir diferentes formatos dinámicos, como simulaciones interactivas diseñadas a medida con el requisito de que pudieran ser utilizadas en un entorno segura. El número de ítems para cada prueba fue de 41, 40 y 30 ítems respectivamente para cada una de las CD de la prueba de *IAD* y 44 ítems para la prueba de *Netiqueta*.

Para el diseño de los ítems se siguieron los criterios de diseño que se exponen a continuación:

- Cuanto más cortos y sencillos, mejor.
- Relacionados con situaciones prácticas y comunes, especialmente en escenarios del mundo real.
- Neutral con respecto a marcas comerciales y soluciones tecnológicas específicas. Si no es posible, utilizamos como base para las simulaciones las soluciones más utilizadas. En las simulaciones proporcionamos mensajes "*alt*" (texto alternativo a las imágenes) al pasar por encima de las diferentes opciones como ayuda.
- Abordar los elementos de CD seleccionados (conocimientos y habilidades) y hacer referencia a los tres niveles (básico, intermedio y avanzado).
- Equilibrar el número de preguntas de conocimientos y habilidades de cada prueba: 22/22 para la prueba de *Netiqueta* y 25/35 para la prueba de *IAD*.
- Todos los ítems son dicotómicos para todos los formatos (correcto 1 e incorrecto 0). No se tuvo en cuenta la complejidad de los ítems ni las respuestas parciales durante la resolución de un ítem. Tomamos esta decisión para simplificar su comprensión, o de otra forma, los criterios de evaluación de cada ítem habrían sido más complicados.

Utilizamos diferentes formatos de ítems: ítems de opción múltiple, simulaciones interactivas, ítems basados en una imagen o simulación, y tareas abiertas.

Intentamos que todos los ítems se mostrasen en una sola pantalla para minimizar desplazamientos del navegador.

En las simulaciones interactivas representamos situaciones de la vida real, donde los participantes tienen que resolver las tareas demandadas realizando las acciones requeridas, como compartir un documento almacenado en la nube o localizar la farmacia abierta más cercana desde el teléfono móvil. En el diseño de las simulaciones, seleccionamos escenarios que pueden encontrarse comúnmente en el contexto de los CD seleccionadas. Buscamos evaluar las habilidades cognitivas en una situación en la que se debe aplicar la tecnología digital en diferentes dispositivos (p.ej., dispositivos móviles, ordenadores portátiles o estaciones de trabajo). No nos interesa evaluar las herramientas per se. Este enfoque se adapta mejor al rápido cambio tecnológico. Desarrollamos las simulaciones utilizando una solución comercial llamada Articulate Storyline (ASL)[50] para el diseño de simulaciones interactivas basadas en escenarios ramificados, en los que diferentes elecciones llevan a al examinando por diferentes caminos, pudiendo trazar el rendimiento de los participantes y evaluando realmente su aptitud. Las simulaciones pueden diseñarse teniendo en cuenta los comportamientos que se suelen realizar en un contexto real (clics, dobles clics, introducción de texto, clic derecho, teclas de acceso directo, al hacer clic con el botón derecho se muestra el menú contextual, al hacer clic en un enlace subrayado en color azul se redirige a la página enlazada, etc.). Además, para la resolución de los ítems, consideramos diferentes caminos y determinamos un límite de clics erróneos permitidos para ajustar la dificultad del ítem. Con este enfoque, los participantes pueden explorar los programas y las situaciones hasta cierto punto, utilizando el juicio y la toma de decisiones, en lugar de determinar de memoria la ubicación de todas las funcionalidades.

Para cada ítem, recogimos el tiempo de respuesta individual y el resultado (ítem resuelto o no). Además, ASL mediante scripts incrustados en las simulaciones permite la creación de variables para recoger más información sobre el rendimiento de los participantes. Así, registramos adicionalmente el número de clics en cada paso (correctos o no) y el último al que llegó. Un ejemplo del diseño de una simulación basada en un dispositivo móvil en ASL puede verse en la figura 4.2.

[50] https://articulate.com/360/storyline

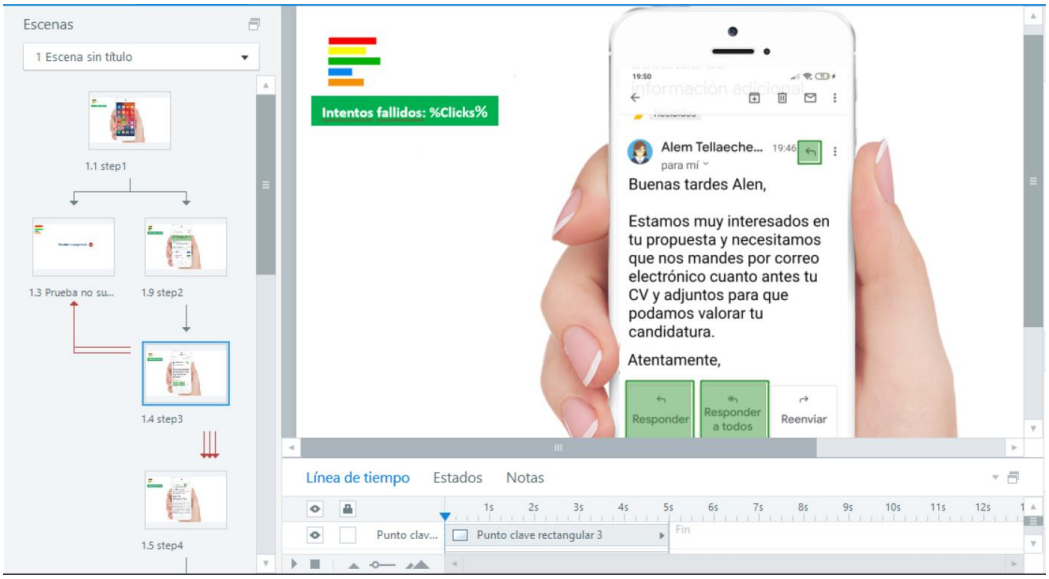


Figura 4.2. Ejemplo de diseño de una simulación basada en un dispositivo móvil en ASL.

También utilizamos ítems basados en una imagen o simulación, en los que se presenta una situación en una imagen o en una simulación a los participantes y éstos tienen que evaluarla críticamente poniendo en práctica sus conocimientos, para posteriormente seleccionar la opción correcta. Más aun, aplicamos los mismos principios de diseño que para las simulaciones interactivas, con la salvedad de que este formato no tiene un límite de clics erróneos, es decir, los participantes pueden examinar las diferentes situaciones planteadas libremente.

Otro formato utilizado fueron los ítems de tareas abiertas, donde los participantes tienen que interactuar con el ordenador y sus aplicaciones. Por ejemplo, abrir una hoja de cálculo y aplicar filtros para localizar cierta información o acceder a un portal de ofertas de empleo simulado "*Lanbila*" para resolver una serie de tareas. Implementamos este tipo de ítems realizando desarrollos a medida e integrándolos en ECTD para que los resultados se calculasen automáticamente (ver figuras 4.3 y 4.4).

Figura 4.3. Ejemplo de un ítem de tarea abierta.

Figura 4.4. Web de "*Lanbila*".

Por último, todos los diferentes formatos de ítems seleccionados se integraron en ETCD. Un ejemplo de la interfaz de la prueba puede verse en la figura 4.5. La herramienta de avaluación registra las respuestas a los ítems, los resultados obtenidos por cada ítem y prueba, los tiempos de respuesta por ítem y prueba, y adicionalmente para las simulaciones, el número de clics erróneos y el último paso alcanzado (para conocer el camino seguido en sus soluciones).

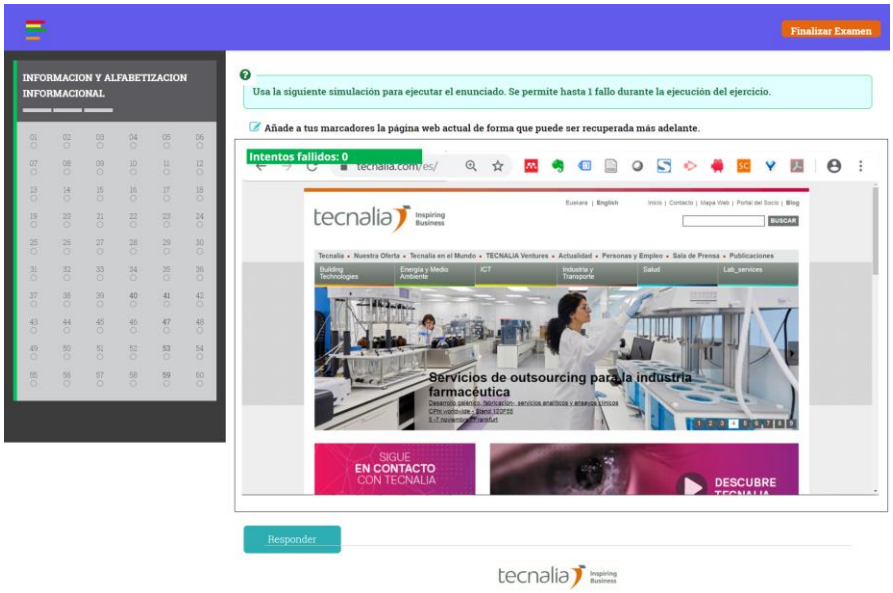


Figura 4.5. Ejemplo de la interfaz de la prueba, mostrando un ítem basado en una simulación

Tras diseñar ETCD, pasamos a la fase de pruebas para analizar la solución diseñada. Posteriormente, basándonos en los resultados de la fase de pruebas, revisamos y redefinimos los elementos y principios del diseño junto con los procesos de implementación.

4.2.3. Ciclos iterativos de prueba y perfeccionamiento de la solución en la práctica

En esta sección, describimos los resultados de poner en práctica y evaluar la solución propuesta en los ciclos iterativos, detallando la metodología seguida, dado que representa principalmente las fases de recogida y análisis de datos del estudio.

Nos gustaría señalar que debido a la pandemia tuvimos que introducir importantes modificaciones en el proceso de recogida de datos de la segunda iteración. No pudimos recoger los resultados en un entorno controlado bajo una sesión supervisada como habíamos previsto inicialmente. Tuvimos que organizar una convocatoria abierta dentro de la All Digital Week[51], en la que la mayoría de los participantes realizaron las pruebas desde sus casas y trabajos en diferentes condiciones. De hecho, varios comentarios y sugerencias recibidos tras realizar las pruebas estaban relacionados con el entorno de la prueba. Así que tuvimos que examinar a los comentarios recibidos para descartar los causados por no haberse administrado en un entorno controlado.

[51] https://www.alldigitalweek.eu/

Primera iteración con facilitadores de centros de CD

Durante el mes de marzo de 2020, los facilitadores de los centros de CD de la red de telecentros KZgunea (KZgunea)[52] completaron individualmente las cuatro pruebas disponibles en ETCD y nos enviaron sus comentarios rellenando una plantilla que les enviamos previamente por correo electrónico junto a los objetivos del estudio y las instrucciones a seguir. Las pruebas se mostraron en español. Los participantes podían navegar por los ítems y cambiar el orden de sus respuestas. Las acciones realizadas durante las pruebas y el orden de las respuestas por parte de los facilitadores eran registradas la plataforma, y los resultados se generaban de forma automática. En esta primera iteración pretendíamos investigar el contenido y la redacción de los ítems, y las sugerencias de los facilitadores para mejorar los ítems y las pruebas.

Segunda iteración con ciudadanos

Durante los días del 22 a 28 de marzo de 2021, se celebró la *All Digital Week*, ofreciendo diversas actividades en línea con el objetivo de promover la adquisición de CD. Decidimos apoyar la acción organizando una actividad online en la que se invitaba a los ciudadanos a evaluar sus CD completando las pruebas disponibles en ETCD. Dirigimos la invitación principalmente a ciudadanos que conocían IT Txartela y/o personas interesadas en mejorar su CD. Para acceder a nuestro grupo objetivo, utilizamos varias estrategias: 1) pusimos un banner de invitación en la página web de IT Txartela; 2) hicimos difusión del evento en las redes sociales; y 3) accedimos a contactos personales como amigos, familiares y colegas para llegar a un mayor número de participantes. Para hacer más atractiva la participación en el estudio, decidimos sortear una serie de dispositivos digitales entre todos los participantes.

Para esta segunda interacción, decidimos implementar una prueba basada en el AC *IAD*, formada por una selección de ítems de las tres CD de esa área, y una prueba basada en una CD, con una selección de ítems de la CD de *Netiqueta*. Nótese que descartamos las relaciones cruzadas entre las CD seleccionadas y medimos cada CD individualmente. Esta decisión favoreció la validez externa de las pruebas al simplificarlas y facilitar su comprensión. Además, al explicar en profundidad todos los pasos y las decisiones tomadas en cada fase, buscamos facilitar que fueran adoptadas por un público más amplio (Van Deursen et al., 2016). Los principales factores que tuvimos en cuenta para el desarrollo de las pruebas finales fueron:

- El tiempo necesario para completar cada prueba debía ser inferior a 30 minutos, para disminuir la probabilidad de que los usuarios abandonasen demasiado pronto. Así, la prueba de *IAD* incluyó 60 ítems y la de *Netiqueta* 44 ítems.

[52] https://www.kzgunea.eus/es/inicio

- La distribución de los ítems en la prueba de *IAD* fue similar para cada DC. En la prueba de *Netiqueta*, la distribución la realizamos asegurando que todas las SC estuvieran presentes. La distribución de las SC la realizamos de acuerdo con la revisión bibliográfica realizada al inicio del estudio y con las sugerencias de los facilitadores (ver tabla 4.4).
- Respecto a los niveles de CD abarcados, se consideró la siguiente proporción de ítems para cada prueba: 25% de nivel básico, 50% de nivel intermedio y 25% de nivel avanzado. Asignamos los niveles de CD a los ítems mapeando los verbos de los enunciados con la taxonomía de Bloom (Krathwohl, 2002).

Prueba	N° de ítems	SC
Netiqueta	10	SC1
	11	SC2
	16	SC3
	7	SC4
IAD	10	SC5
	10	SC6
	10	SC7
	10	SC8
	20	SC9

Tabla 4.4. Distribución del número de ítems para cada SC.

Las pruebas finales quedaron disponibles para su uso en ETCD. Antes de iniciar las pruebas, se proporcionó a los examinados en la misma herramienta una ayuda interactiva para que se familiarizaran con el entorno de la prueba. Los ítems se cargaron en las pruebas en el mismo orden. Sin embargo, los participantes podían navegar por los ítems y cambiar el orden de sus respuestas. Las acciones realizadas durante las pruebas y el orden de las respuestas de los participantes quedaron registradas, y los resultados se generaron automáticamente. Resumiendo, en el segundo ciclo iterativo pretendimos:

- Evaluar las pruebas con usuarios finales.
- Analizar las respuestas de los participantes e investigar la idoneidad de distintos modelos TRI examinando diferentes indicadores de ajuste del modelo.
- Examinar la calidad de las pruebas (AERA, APA, y NCME, 2014; Messick, 1995), analizando la fiabilidad y la validez de las pruebas.

4.2.4. Reflexión para definir "principios de diseño" y mejorar la implementación de soluciones

Seguimos una DBR porque esta metodología es adecuada para describir el proceso iterativo de diseño y desarrollo de los principales resultados de nuestro estudio y para especificar los principales aspectos considerados y las decisiones tomadas. Los principios de diseño descritos en el estudio contienen el conocimiento de los procedimientos, los resultados y el contexto seguidos durante los diferentes pasos. Implementamos nuestra solución tomando como referencia DigComp, que fue creado para ser utilizado para el desarrollo de iniciativas a medida, proporcionando una terminología común adaptable a nuestros requisitos en materia de CD. DigComp no depende de la tecnología y describe las CD en términos generales. Con lo cual, las organizaciones interesadas en desarrollar su propia implementación deben identificar qué conocimientos y habilidades son relevantes y si algunas aplicaciones o dispositivos digitales específicos son elementos clave según sus peculiaridades. Así pues, especificamos qué conocimientos y habilidades eran de interés para nuestro grupo objetivo y posteriormente, diseñamos la herramienta de evaluación. Para implementar los ítems más adecuados, utilizamos diferentes formatos como simulaciones interactivas y otros formatos dinámicos. Hay que tener en cuenta que para otras CD podrían ser otros formatos más adecuados. Los lectores podrán decidir qué aspectos pueden ser de interés para sus propias implementaciones en función de sus contextos específicos.

4.3. Resultados

4.3.1. Fase 3: Ciclos iterativos de pruebas y perfeccionamiento de la solución en la práctica

Primera iteración con facilitadores de centros de CD

Los participantes fueron facilitadores (n = 93) de 75 centros diferentes de KZgunea. No se registró ninguna información personal de los facilitadores ni se realizó ninguna selección adicional. Los servicios prestados por KZgunea incluyen la formación en CD y apoyo a IT Txartela. Su experiencia es de gran valor, ya que apoyan diariamente las necesidades de los ciudadanos en materia de CD.

Enviamos la invitación con los detalles del estudio por correo electrónico al coordinador de KZgunea, incluyendo una plantilla para recoger la información. La participación era voluntaria. Una vez realizadas las pruebas, el coordinador nos envió las plantillas rellenadas por los facilitadores. Posteriormente, analizamos la información. Primero, identificamos los ítems que obtuvieron más comentarios y sugerencias. Examinamos los ítems que obtuvieron al menos tres menciones

relacionadas con la dificultad de comprensión o que presentaban dificultades técnicas, analizando detalladamente sus comentarios y sugerencias. Así mismo, también analizamos todas las sugerencias de mejora, así como los comentarios relacionados con el grado de dificultad de los ítems. Si los comentarios sobre su nivel se alejaban demasiado del nivel inicialmente seleccionado, se revisaba la pregunta en profundidad. En concreto, para las simulaciones interactivas, aumentamos o disminuimos el límite de clics erróneos permitidos para ajustar el nivel de dificultad del ítem. Como resultado, se revisaron varios ítems y se modificaron algunos de ellos. Los detalles pueden ser consultados en la tabla 4.5.

CD	Ítem	Formato	Comentarios y sugerencias	Acciones implementadas
Netiqueta	Item3	Simulación	No está muy clara la acción solicitada en la tarea.	Complete el enunciado indicando mejor la acción que debe esperar el usuario.
Netiqueta	Item5	Opción múltiple	El ítem contiene respuestas ambiguas.	Modificar la redacción del ítem para hacerlo más comprensible.
Netiqueta	Item6	Opción múltiple	El enunciado es ambiguo.	Modificar la redacción del ítem para hacerlo más comprensible.
Netiqueta	Item14	Opción múltiple	El enunciado es ambiguo.	Ítem modificado
Netiqueta	Item21	Opción múltiple	Comentarios variados sobre este ítem. Parece que no hay una opinión común sobre esta cuestión.	Ítem eliminado del banco de ítems.
Netiqueta	Item22	Simulación	Parece que no hay una opinión común sobre este ítem, y no estábamos seguros de si incluirla o no.	Ítem sustituido por otro del banco de ítems.
Netiqueta	Item23	Simulación	El ítem contiene respuestas ambiguas.	Ítem sustituido por otro del banco de ítems.
Netiqueta	Item28	Basado en imagen	Tiene un error ortográfico y dos opciones son muy similares.	Ítem modificado
Netiqueta	Item32	Basado en imagen	No está muy clara la acción solicitada en el ítem. Sería recomendable utilizar otro formato de ítem.	Formato del ítem modificado para mejorar la funcionalidad.
Netiqueta	Item36	Simulación	Explica mejor el enunciado.	Reducido el número de intentos

			Demasiado fácil.	permitidos.
Netiqueta	Item42	Simulación	Comentarios variados sobre este ítem. Parece que no hay una opinión común.	Ítem sustituido por otro del banco de ítems.
CD1	Item1	Opción múltiple	Sería recomendable utilizar otro formato de ítem.	Formato del ítem modificado para mejorar la funcionalidad.
CD1	Item3	Opción múltiple	Explicar mejor el enunciado. Es un poco ambiguo.	Modificar la redacción del ítem para hacerlo más comprensible.
CD1	Item4	Simulación web	Sería recomendable aclarar mejor cómo introducir la solución (el formato)	Modificar la redacción del ítem para hacerlo más comprensible.
CD1	Item5	Basado en imagen	No está muy clara la acción solicitada en el enunciado. Sería recomendable utilizar otro formato de ítem. En este caso no es necesario utilizar una simulación. Añade una dificultad extra.	Formato del ítem modificado para mejorar la funcionalidad.
CD1	Item6	Simulación web	Sería recomendable aclarar mejor cómo introducir la solución (el formato)	Modificar la redacción del ítem para hacerlo más comprensible.
CD1	Item11	Simulación	Es posible resolver el ítem tomando un nuevo camino que no ha sido incluido en la simulación.	Simulación modificada incluyendo un nuevo camino válido para alcanzar la solución.
CD1	Item12	Basado en imagen	Comentarios relacionados con la calidad de la imagen de la pregunta (relevante teniendo en cuenta que es una pregunta basada en imágenes)	Mejora de la calidad de la imagen.
CD1	Item16	Simulación	Es posible resolver el ítem introduciendo un nuevo término válido que no ha sido incluido en la simulación.	Simulación modificada incluyendo un nuevo término válido para alcanzar la solución.
CD1	Item17	Basado en imagen	Sería recomendable utilizar otro formato de ítem. En este caso no es necesario utilizar una simulación. Añade una dificultad adicional.	Formato del ítem modificado para mejorar la funcionalidad.

CD1	Item18	Basado en imagen	Sería recomendable utilizar otro formato de ítem. En este caso no es necesario utilizar una simulación. Añade una dificultad adicional.	Formato del ítem modificado para mejorar la funcionalidad.
CD1	Item23	Simulación	La simulación tiene problemas técnicos.	Simulación modificada.
CD1	Item24	Interacción con el puesto	Problemas descargando el adjunto.	Ítem modificado
CD1	Item28	Simulación	Este ítem debería ser avanzado. Esta funcionalidad no es muy conocida por los usuarios.	Comentarios anotados para futura revisión.
CD1	Item31	Opción múltiple	Comentarios sobre la ambigüedad del ítem.	Ítem sustituido por otro del banco de ítems.
CD1	Item35	Simulación	Mejorar la usabilidad de la simulación.	Simulación modificada.
CD2	Item9	Simulación	Comentarios sobre la ambigüedad de dos posibles respuestas. Son demasiado similares.	Ítem sustituido por otro del banco de ítems.
CD2	Item10	Basado en imagen	Comentarios sobre la ambigüedad de dos posibles respuestas. Son demasiado similares.	Ítem modificado
CD2	Item11	Basado en imagen	Sugerencias para modificar el logotipo del periódico y poner uno genérico.	Mejora de la calidad de la imagen.
CD2	Item12	Simulación	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.
CD2	Item17	Basado en imagen	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.
CD2	Item20	Opción múltiple	Es demasiado similar a otro ítem.	Ítem sustituido por otro del banco de ítems.
CD2	Item22	Basado en imagen	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.
CD2	Item23	Basado en	Comentarios sobre la ambigüedad	Se ha mejorado la imagen del ítem.

		imagen	de las respuestas y sobre algunos aspectos de la imagen.	Se ha modificado la redacción del enunciado y las opciones para que el ítem sea más comprensible.
CD2	Item24	Basado en imagen	Comentarios sobre la ambigüedad de las respuestas y sobre algunos aspectos de la imagen.	Se ha mejorado la imagen del ítem. Se ha modificado la redacción del enunciado y las opciones para que el ítem sea más comprensible.
CD2	Item27	Opción múltiple	Comentarios sobre la ambigüedad de dos posibles respuestas. Son demasiado similares.	Ítem modificado
CD2	Item29	Basado en imagen	Comentarios sobre la ambigüedad de las respuestas y sobre algunos aspectos de la imagen.	Se ha mejorado la imagen del ítem. Se ha modificado la redacción del enunciado y las opciones para que el ítem sea más comprensible.
CD2	Item35	Basado en imagen	Comentarios sobre la ambigüedad de las respuestas y sobre algunos aspectos de la imagen.	Se ha mejorado la imagen del ítem. Se ha modificado la redacción del enunciado y las opciones para que el ítem sea más comprensible.
CD2	Item36	Simulación	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.
CD3	Item3	Opción múltiple	Comentarios sobre la ambigüedad de dos posibles respuestas.	Ítem modificado
CD3	Item4	Opción múltiple	Sugerencias para mejorar el enunciado.	Ítem modificado
CD3	Item5	Opción múltiple	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.
CD3	Item6	Opción múltiple	Comentarios sobre la ambigüedad de dos posibles respuestas.	Ítem modificado
CD3	Item7	Simulación	Comentarios sobre las acciones que deberían permitirse y otro camino hacia la solución que no está incluido en la simulación.	Simulación modificada incluyendo clics y dobles clics y añadiendo una nueva ruta a la solución.
CD3	Item8	Simulación	Sugerencias para incluir el doble clic.	Simulación modificada incluyendo la acción de doble clic.
CD3	Item9	Simulación	Sugerencias para incluir el botón derecho.	Simulación modificada incluyendo la acción de botón derecho.

CD3	Item11	Opción múltiple	Comentarios sobre la ambigüedad de las respuestas.	Ítem modificado
CD3	Item14	Opción múltiple	Este ítem debería ser avanzado. Esta funcionalidad no es muy conocida por los usuarios.	Comentarios anotados para futura revisión.
CD3	Item15	Opción múltiple	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.
CD3	Item17	Simulación	Comentarios para añadir otro camino a la solución.	Simulación modificada incluyendo un nuevo camino válido para alcanzar la solución.
CD3	Item21	Simulación	Sugerencias sobre la inclusión del doble clic, reducir el número de clics erróneos permitidos y reformular el enunciado.	Simulación modificada permitiendo el doble clic y disminuyendo el número de clics erróneos permitidos.
CD3	Item22	Opción múltiple	Comentarios sobre la ambigüedad de dos posibles respuestas.	Ítem modificado
CD3	Item25	Simulación	Sugerencias sobre la inclusión de algunas acciones y caminos permitidos que inicialmente no estaban incluidas.	Simulación modificada, incluyendo más acciones posibles y un nuevo camino.
CD3	Item28	Simulación	Sugerencias sobre la inclusión de algunas acciones que inicialmente no estaban incluidas.	Simulación modificada incluyendo más acciones posibles (doble clic y clic derecho).
CD3	Item29	Simulación	Sugerencias sobre la inclusión de algunas acciones y caminos permitidos que inicialmente no estaban incluidas.	Simulación modificada, incluyendo más acciones posibles y un nuevo camino.
CD3	Item30	Simulación	Diferentes tipos de comentarios y sugerencias.	Ítem sustituido por otro del banco de ítems.

Tabla 4.5. Ítems revisados y modificados después de la revisión.

A continuación, examinamos las sugerencias de incorporación de nuevos ítems, excluyendo las sugerencias orientadas a evaluar aplicaciones per se. Como resultado, desarrollamos varios ítems nuevos (ver tabla 4.6).

CD: Comentarios y sugerencias	Ítem	Formato	Descripción del ítem
Netiqueta: • Qué es apropiado publicar en redes sociales: críticas al jefe o empresa, revisiones inapropiadas de sitios visitados, etc.). • Basar los ítems en otras redes sociales muy usadas como LinkedIn, Facebook o Twitter. • Más específicas sobre acoso en redes sociales, especialmente orientado a denunciarlo a la policía. • Publicaciones con fotos de menores. • Más ítems simulando como responder a un troll, etc. • Más ítems sobre bullying en la red. Como actuar, etc.	Item1	Basado en imagen	Elija la respuesta correcta según la situación presentada en un hilo de un foro. (Los participantes tendrán que evaluar si el contenido y el formato son correctos).
	Item2	Basado en imagen	Elija la respuesta correcta según la situación presentada en un foro recién registrado (Los participantes tendrán que seleccionar la siguiente acción a realizar).
	Item3	Basado en imagen	Elija la respuesta correcta según la situación presentada en un hilo de un foro. (Los participantes tendrán que evaluar si el contenido y el formato son correctos).
	Item4	Opción múltiple	Selecciona las reglas de Netiqueta correctas a seguir en LinkedIn.
	Item5	Opción múltiple	Selecciona las reglas de Netiqueta correctas a seguir en redes sociales.
	Item6	Basado en imagen	Elija la respuesta correcta según el perfil de LinkedIn que se presenta en la imagen. (Los usuarios tendrán que evaluar si el contenido y el formato son correctos).
	Item7	Opción múltiple	Selecciona las reglas de Netiqueta correctas a seguir en Facebook.
	Item8	Opción múltiple	Selecciona las reglas de Netiqueta correctas a seguir en redes sociales.
	Item9	Opción múltiple	Selecciona las reglas de Netiqueta correctas a seguir en redes sociales.
	Item10	Basado en imagen	Elija la respuesta correcta según las situaciones que se muestran en la imagen (Ejemplos de textos citando fuentes).
	Item11	Basado en imagen	Elige la acción correcta de acuerdo con una imagen mostrando el ataque de un Troll en redes sociales.
CD1: • Búsquedas basadas en YouTube. • Búsqueda y filtrado de noticias.	Item1	Simulación	Acceder a un canal de YouTube, y filtrar los vídeos para guardar uno específico para verlo más tarde.
	Item2	Simulación	Filtrar los resultados de una búsqueda de vídeos para que sólo aparezcan los vídeos con una licencia específica.
	Item3	Simulación	Filtrar los resultados de una búsqueda de productos en

• Búsqueda y filtrado de productos en compras online. • Búsquedas basadas en Google Maps. • Búsquedas y filtros avanzados. • Más ítems basados en redes sociales: Facebook, YouTube, etc.			una tienda online, por la valoración media de los clientes y el precio.
	Item4	Opción múltiple	Seleccionar la afirmación correcta sobre el uso de los comandos de búsqueda avanzada.
	Item5	Basado en imagen	Seleccionar la afirmación correcta sobre posibles búsquedas en Facebook.
	Item6	Simulación	Filtrar los resultados de una búsqueda de noticias para que sólo se muestren las más recientes.
	Item7	Simulación	Realizar una búsqueda en Google Maps para saber cómo llegar a la estación de tren desde el trabajo en coche.
	Item8	Simulación	Realizar una búsqueda en Google Maps para localizar y saber cómo llegar a la farmacia abierta más cercana.
	Item9	Basado en imagen	De acuerdo con la situación mostrada, seleccionar la afirmación correcta sobre la búsqueda realizada y el filtro utilizado.
CD2: • Pago seguro en línea, sitios web seguros y diferentes métodos de pago. • Simulación de correos electrónicos sospechosos, phishing, etc. • Ítem sobre evaluar la información digital e identificar Fake News.	Item1	Opción múltiple	Indicar cuál de las siguientes opciones es una prueba clara de que nos encontramos ante en una tienda online potencialmente fraudulenta.
	Item2	Opción múltiple	Indicar cuáles de los siguientes puntos pueden utilizarse para identificar sitios de tiendas online fraudulentas.
	Item3	Basado en imagen	De acuerdo con la situación mostrada, ¿Es fiable y podemos comprar con seguridad?
	Item4	Basado en imagen	De acuerdo con la situación mostrada, ¿Es fiable y podemos comprar con seguridad?
	Item5	Basado en imagen	De acuerdo con la situación mostrada donde se muestra un correo electrónico recibido, identifica la situación (Phishing).
	Item6	Opción múltiple	Cómo proceder cuando recibimos un correo electrónico no solicitado de una persona desconocida.
	Item7	Opción múltiple	Cómo proceder para acceder a un banco en línea.
	Item8	Opción múltiple	Indicar si el enunciado sobre el phishing es correcto.

CD3: • Ítems basados en otras herramientas como Dropbox. • Más ítems basados en Google Drive como crear documentos o colaborar en red. • Sincronización de dispositivos utilizando la nube (Fotos, contactos, etc.).	Item1	Simulación	Una vez identificado en Google, crear un documento en la nube disponible desde diferentes dispositivos y ubicaciones.
	Item2	Opción múltiple	Indicar si el siguiente enunciado es correcto respecto a la sincronización de navegadores en distintos dispositivos.
	Item3	Opción múltiple	Indicar si el siguiente enunciado es correcto respecto al movimiento de información entre distintos dispositivos móviles.

Tabla 4.6. Comentarios y sugerencias recibidos, y nuevos ítems desarrollados.

Por último, también recibimos algunos comentarios y sugerencias generales relacionados con el entorno de las pruebas. Los comentarios sobre los problemas que tuvieron debido al uso de diferentes navegadores y resoluciones fueron descartados porque las pruebas no se realizaron en un entorno controlado y optimizado, utilizando un navegador específico y seguro como habíamos previsto inicialmente. También se nos sugirió posibilitar la comprobación de las respuestas correctas a las preguntas fallidas al final de la prueba. A pesar de estar fuera del alcance de este estudio, esta sugerencia fue debidamente anotada para futuras implementaciones. Como resultado de esta iteración, se modificaron las pruebas y se prepararon para la siguiente iteración.

Segunda iteración con ciudadanos

Los participantes fueron ciudadanos que completaron de forma anónima las pruebas de *IAD* (n=329) y *Netiqueta* (n=214). No seguimos ningún criterio de selección adicional, pero les pedimos que rellenaran además su sexo y rango de edad con fines demográficos. El tiempo total de realización de la prueba de *Netiqueta* fue de M = 916 s. (SD = 585 s.), es decir, más de 15 minutos, y de M = 2.007 s. (SD = 1.253 s.), es decir, casi 34 minutos para la prueba de *IAD*.

Además, eliminamos los resultados de 113 participantes de la prueba de *IAD* y 16 participantes de la prueba de *Netiqueta* debido a que sus datos tenían más de cinco ítems sin responder o sus intentos duraban menos de 5 minutos, y consideramos improbable que un examinando pudiera leer los ítems y responder en tan poco tiempo en condiciones. La distribución de los registros de los participantes y sus datos demográficos se resumen en la tabla 4.7. Hubo una ponderación demográfica a favor de los usuarios en el rango de edad de 25 a 54

años y un número ligeramente mayor de usuarios masculinos, especialmente en la prueba de *IAD*. La tabla 4.8 muestra las puntuaciones obtenidas en las pruebas.

Prueba	Género	Rango de edad
Netiqueta (n = 201)	Hombre 54.6% Mujer 46.4%	(16–24) 15.7%
		(25–54) 76.9%
		(55–74) 7.4%
IAD (n = 209)	Hombre 64.8% Mujer 35.2%	(16–24) 22.6%
		(25–54) 68.3%
		(55–74) 9.0%

Tabla 4.7. Distribución de registros y datos demográficos de los participantes.

Prueba	Media	Desviación Estándar
Netiqueta (n = 201)	24.70	8.70
IAD (n = 209)	42.69	11.25

Tabla 4.8. Resumen de los datos descriptivos de las pruebas.

En el desarrollo de un instrumento cuantitativo con fines de evaluación, es crucial medir su calidad (Bandalos, 2018; Mueller y Knapp, 2018), que consiste principalmente en medir su validez (si el instrumento de evaluación evalúa lo que debe medirse y se refiere a cómo se interpretan y utilizan las puntuaciones de las pruebas (AERA, APA, y NCME, 2014)) y fiabilidad (si el instrumento de evaluación produce resultados similares en condiciones equivalentes (Scholtes et al., 2011)). Para obtener evidencias de calidad, existen diferentes métodos, y los estudios que implican el desarrollo de un instrumento de evaluación deben incluir suficientes evidencias (Mueller y Knapp, 2018).

Para obtener pruebas de validez, consideramos la validez de contenido, la de constructo, así como la validez de los RP. En cuanto a la validez de contenido, en la primera iteración ya realizamos un proceso de validación basado en el juicio de expertos, concretamente con los facilitadores de KZgunea. Revisamos que el contenido de las pruebas representaba el constructo que se pretendía evaluar y que era adecuado para cumplir los objetivos de la prueba.

Además, algunos de los ítems incluidos en las pruebas ya habían sido examinados en nuestro trabajo anterior Bartolomé et al. (2020) y que se describe

detalladamente en el capítulo 5. En dicho estudio, analizamos los RP de los diferentes tipos de ítems incluidos en las pruebas, obteniendo información útil para comprender el rendimiento de los participantes e investigar si los criterios de evaluación de cada ítem estaban correctamente establecidos. Para el resto de los ítems incluidos en las pruebas, aplicamos los principios de diseño identificados en el estudio.

Cabe mencionar que este tipo de soluciones, donde se adopta un enfoque más pragmático, buscando la sencillez y que sea fácilmente entendible y adoptable, suele presentar una validez interna más débil (es decir, la evidencia de que el diseño refleja lo que se mide). Por lo tanto, tuvimos que equilibrar la validez interna y externa mediante diferentes decisiones metodológicas durante el diseño de las pruebas. La validez externa es la medida en que los resultados de nuestro estudio pueden generalizarse a otros contextos. Así mismo, participantes las relaciones entre las medidas obtenidas y otras posibles variables de confusión, como el género o el rango de edad de los participantes (Messick, 1995). Para obtener pruebas descriptivas básicas de validez en la construcción y validación de las pruebas, examinamos el parámetro de dificultad (valor p) de los ítems y los índices de discriminación como indicador de partida para justificar la elección del modelo. Así mismo, analizamos la validez dimensional y la fiabilidad.

En primer lugar, realizamos un análisis de ítems para examinar el parámetro de dificultad (valor p) de los ítems. Los ítems cuyo valor p es cercano a 0.00 (muy difícil) o cercano a 1,00 (muy fácil) deben ser eliminados. Además, es necesario investigar si los ítems tienen índices de discriminación similares como indicador de partida para justificar la elección del modelo (Hambleton et al., 1991). El modelo logístico de un parámetro (1 PLM) sólo tiene un parámetro libre (el parámetro de dificultad) y espera que todos los ítems tengan índices de discriminación similares de todos los ítems. En caso contrario, el 1 PLM no debería aplicarse. Por lo tanto, calculamos la distribución de las correlaciones punto-biserial, que es la correlación de Pearson entre cada ítem y la puntuación total de la prueba para cada examinando. Los ítems con un valor punto-biserial inferior a 0.15 deben ser eliminados (Varma y Simon, 2006). Sólo tuvimos que eliminar el ítem 5 de la prueba de *Netiqueta* porque su correlación era < 0.15 (véase la tabla 4.9).

Prueba IAD			Prueba Netiqueta		
Ítem	p-Valor	Correlaciones Point-Biserial	Ítem	p-Valor	Correlaciones Point-Biserial
Item1	0.87	0.567	Item2	0.72	0.541
Item2	0.83	0.435	Item3	0.55	0.257
Item3	0.53	0.527	Item4	0.37	0.458

Item4	0.82	0.468	Item5	0.21	0.113
Item5	0.25	0.200	Item6	0.61	0.311
Item6	0.68	0.553	Item7	0.56	0.431
Item7	0.82	0.627	Item8	0.31	0.398
Item8	0.46	0.386	Item9	0.67	0.474
Item9	0.41	0.269	Item10	0.59	0.422
Item10	0.66	0.444	Item11	0.29	0.256
Item11	0.39	0.467	Item13	0.67	0.413
Item12	0.62	0.450	Item14	0.75	0.396
Item13	0.47	0.439	Item15	0.67	0.615
Item14	0.51	0.450	Item16	0.64	0.478
Item15	0.63	0.342	Item17	0.44	0.183
Item16	0.73	0.500	Item19	0.64	0.568
Item17	0.76	0.549	Item20	0.64	0.466
Item18	0.71	0.431	Item22	0.60	0.441
Item19	0.45	0.299	Item23	0.43	0.233
Item20	0.67	0.383	Item24	0.41	0.377
Item21	0.82	0.317	Item25	0.69	0.604
Item22	0.89	0.560	Item28	0.70	0.455
Item23	0.78	0.504	Item29	0.66	0.659
Item25	0.87	0.541	Item30	0.66	0.374
Item26	0.84	0.405	Item31	0.29	0.164
Item27	0.84	0.453	Item32	0.60	0.329
Item28	0.60	0.261	Item33	0.62	0.571
Item29	0.71	0.414	Item34	0.66	0.534
Item30	0.67	0.291	Item35	0.50	0.554
Item31	0.81	0.626	Item36	0.60	0.694
Item32	0.80	0.445	Item37	0.41	0.243
Item33	0.88	0.506	Item38	0.67	0.534
Item34	0.59	0.282	Item39	0.76	0.501
Item35	0.76	0.575	Item40	0.37	0.173
Item36	0.80	0.517	Item41	0.81	0.622
Item37	0.87	0.460	Item42	0.56	0.373

Item38	0.65	0.404	Item43	0.55	0.537
Item39	0.70	0.350	Item44	0.34	0.229
Item40	0.76	0.529	Item45	0.61	0.545
Item41	0.78	0.415	Item46	0.74	0.568
Item42	0.89	0.580	Item47	0.60	0.433
Item43	0.69	0.344	Item48	0.51	0.251
Item44	0.80	0.433	Item49	0.50	0.431
Item45	0.71	0.552	Item50	0.79	0.191
Item46	0.52	0.358			
Item47	0.83	0.556			
Item48	0.90	0.443			
Item49	0.59	0.374			
Item50	0.74	0.410			
Item51	0.94	0.585			
Item52	0.91	0.528			
Item53	0.78	0.552			
Item54	0.46	0.219			
Item55	0.93	0.501			
Item56	0.74	0.435			
Item57	0.86	0.543			
Item58	0.66	0.512			
Item59	0.68	0.660			
Item60	0.57	0.342			

Tabla 4.9. Características de los ítems: p-valor y correlaciones punto-biserial. El ítem 5, en negrita, fue eliminado.

La evaluación de la validez interna de constructo y la dimensionalidad de una nueva medida es un elemento de prueba relevante para examinar si los efectos observados en nuestro estudio son causados por la manipulación de la variable independiente y no por otros factores (Messick, 1995). Como análisis preliminar, realizamos un *análisis factorial exploratorio* (AFE), que es una de las técnicas más frecuentemente aplicadas en los estudios de desarrollo y validación de pruebas para explorar el conjunto de variables latentes o factores comunes que explican las respuestas a los ítems de una prueba. Aplicamos el análisis factorial de componentes principales con una rotación varimax para examinar las cargas factoriales y la dimensionalidad de ambas pruebas. Antes de realizar el AFE,

calculamos la *prueba de esfericidad de Bartlett* para examinar la factorabilidad de los datos y la prueba de *Kaiser-Meyer-Olkin (KMO)* para evaluar la idoneidad del muestreo. Los resultados confirmaron un estadístico de prueba significativo para la prueba de esfericidad de Bartlett. Para la prueba de *IAD* se obtuvo un chi cuadrado de 339.326, $p < 0.001$ y un valor KMO de 0.717, y para la prueba de *Netiqueta* se obtuvo un chi cuadrado de 344.640, $p < 0.001$ y un valor KMO de 0.818, lo que significa que los datos tenían una adecuada detección de estructura. El AFE de los datos, mediante el método de extracción de componentes principales y una rotación varimax de todos los ítems, reveló un factor fuerte que explicaba el 80% de la varianza total para la prueba de *IAD* y el 70% para la prueba de *Netiqueta*. Estos resultados permiten concluir que existe un factor general fuerte con el que se relacionan todos los ítems de ambas pruebas y que puede interpretarse como la CD general de los participantes.

A continuación, investigamos qué modelo se ajustaba significativamente mejor a los datos, el modelo menos restringido (modelo Rasch multidimensional) o el modelo más simple (modelo Rasch unidimensional). Adoptamos dos enfoques diferentes para ambas pruebas. Para la prueba de *IAD*, consideramos las tres CD del AC como dimensiones independientes. Para la prueba de *Netiqueta*, consideramos las cuatro SCs seleccionadas como dimensiones independientes. Calculamos la diferencia de desviaciones en la estimación de los dos modelos diferentes, que se espera que siga una distribución chi-cuadrado, y los grados de libertad, que es la diferencia en el número de parámetros. Así, podemos calcular estadísticamente qué modelo se ajusta significativamente mejor a los datos.

Para la prueba de IAD, la diferencia entre las desviaciones de estos dos modelos sigue una distribución chi-cuadrado con cinco grados de libertad y una diferencia estimada de 75.9 en la desviación. Por lo tanto, el modelo tridimensional se ajusta mejor a los datos que el unidimensional (véase la tabla 4.10).

Modelo	Desviación	Número de parámetros
1-dim	11900.3	61
3-dim	11824.4	66

Tabla 4.10. Principales indicadores del modelo para la prueba de IAD.

Calculamos el estadístico de ajuste de la media ponderada para todos los ítems con el fin de comprobar la alineación de los ítems con el modelo multidimensional de Rasch. Este estadístico muestra la cantidad de inexactitud del sistema de medición, que debería estar cerca de la unidad (Varma y Simon, 2006). Sin embargo, los valores que se encuentran dentro del rango de 0.75 a 1.33 son ampliamente

aceptados (Wu y Adams, 2013). Sólo encontramos inaceptable el ítem 50, cuyo ajuste fue de 0.74 (véase la tabla 4.11).

Características y análisis de los ítems	Valor
Tamaño de la muestra	209
Número de ítems en la calibración	60
Ajuste ponderado MNSQ (0.75, 1.33) T sig.	1 (0.74)
Estimaciones de fiabilidad: fiabilidad EAP/PV	
CD 1.1	0.880
CD 1.2	0.840
CD 1.3	0.875

Tabla 4.11. Resultados del análisis de ítems (modelo multidimensional).

También examinamos la estimación de fiabilidad de la prueba que proporciona el software Conquest (Adams y Khoo, 1996), que es similar a otras estimaciones de fiabilidad como el alfa de Cronbach (Adams, 2005). Las correlaciones latentes estimadas entre las tres dimensiones fueron altas (véase la tabla 4.12), lo que implica que pueden estar evaluando el mismo rasgo, es decir, que hay un factor fuerte que subyace a todos los ítems, lo que puede interpretarse como la CD general. Encontramos resultados similares en la prueba de Netiqueta, como mostramos a continuación.

Dimensiones	CD 1.1	CD 1.2	CD 1.3
CD 1.1	-	-	-
CD 1.2	0.785	-	-
CD 1.3	0.929	0.847	-

Tabla 4.12. Correlaciones entre las tres dimensiones (basadas en las CD).

La figura 4.6 muestra la ubicación de los participantes y de los elementos en la misma escala utilizando un mapa de Wright, que es una herramienta gráfica potente pero sencilla. La "X" ilustra la ubicación de los examinados en cada dimensión. En la parte derecha del gráfico se muestran los ítems. Los ítems se representan a la derecha, aumentando su dificultad de abajo a arriba. Si el examinando y el ítem están alineados, la probabilidad de responder correctamente a ese ítem es prácticamente del 50%. Si la ubicación del examinando es mayor, la probabilidad de responder a ese ítem correctamente aumenta y viceversa.

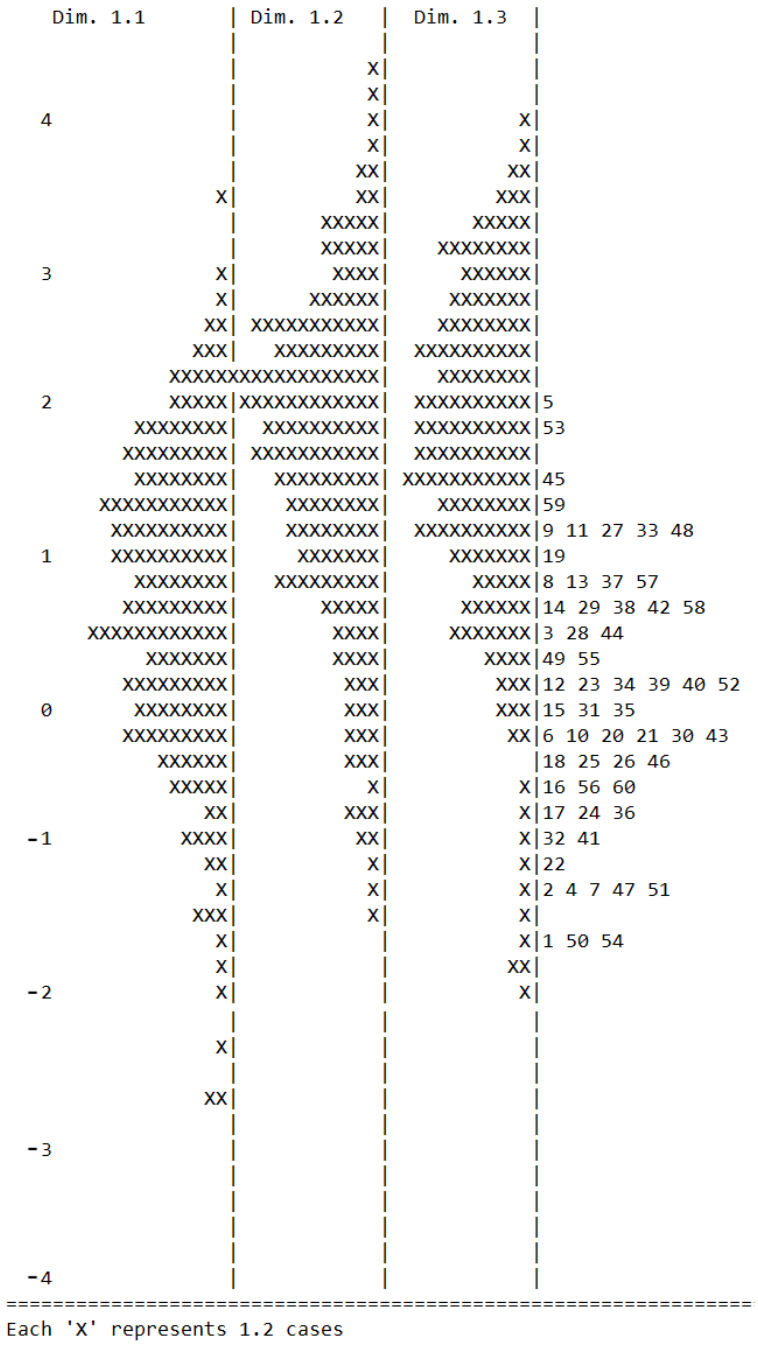


Figura 4.6. Representación de las dimensiones de la prueba de *IAD*.

Para la prueba de Netiqueta, observamos que la diferencia entre las desviaciones de estos dos modelos sigue una distribución chi-cuadrado con 9 grados de libertad y una diferencia estimada de 26.2 en la desviación. Por lo tanto, el modelo cuatridimensional se ajusta mejor a los datos que el unidimensional (véase la tabla 4.13).

Modelo	Desviación	Número de Parámetros
1-dim	10100.0	44
4-dim	10073.8	53

Tabla 4.13. Principales indicadores del modelo para la prueba de Netiqueta.

Calculamos el estadístico de ajuste de la media ponderada para todos los ítems para comprobar la alineación de los ítems con el modelo multidimensional de Rasch y no se encontró ningún ítem que estuviera fuera del rango aceptable (véase la tabla 4.14).

Características y análisis de los ítems	Valor
Tamaño de la muestra	201
Número de ítems en la calibración	43
Ajuste ponderado MNSQ (0.75, 1.33) T sig.	ninguno
Estimaciones de fiabilidad: fiabilidad EAP/PV	-
SC1	0.822
SC2	0.795
SC3	0.859
SC4	0.774

Tabla 4.14. Resultados del análisis de ítems (modelo multidimensional).

Las correlaciones latentes estimadas entre las cuatro dimensiones también fueron altas para la prueba de *Netiqueta* (véase la tabla 4.15).

Dimensiones	SC1	SC2	SC3	SC4
SC1	-	-	-	-
SC2	0.854	-	-	-
SC3	0.913	0.845	-	-
SC4	0.827	0.806	0.862	-

Tabla 4.15. Correlaciones entre las cuatro dimensiones correspondientes a las cuatro SCs.

La figura 4.7 muestra la ubicación de los examinados y los artículos en la misma base de escala utilizando un mapa de Wright.

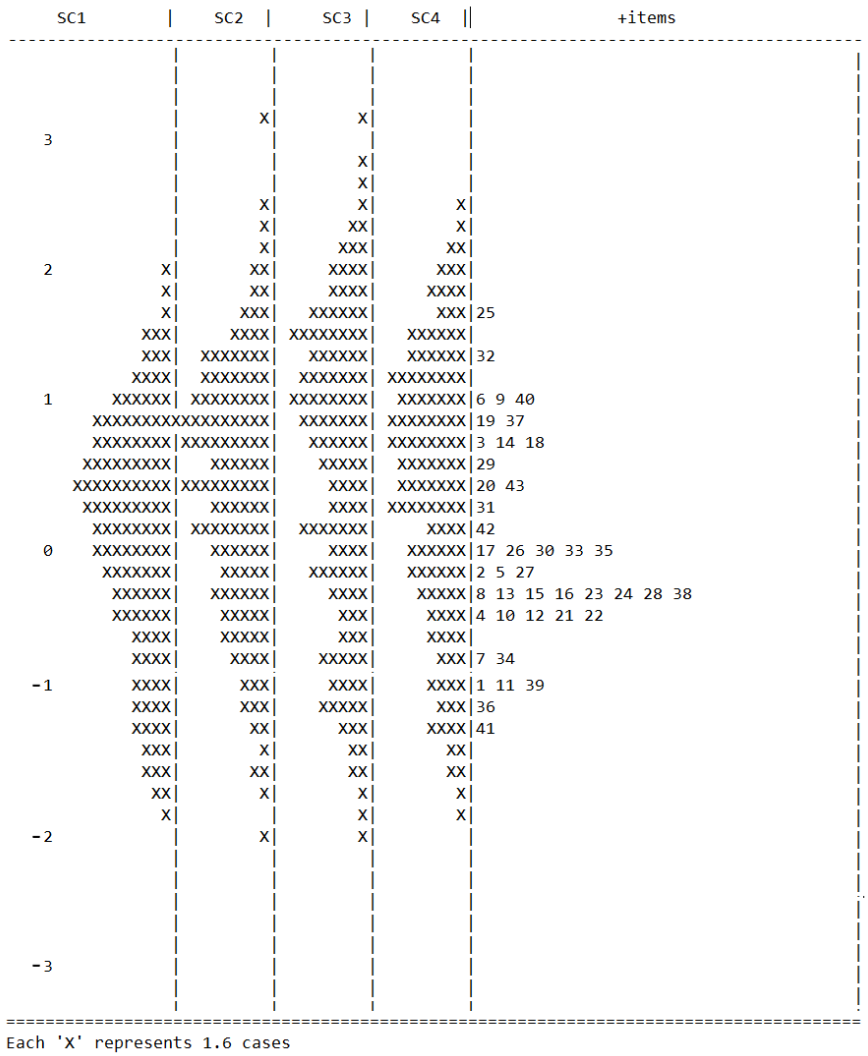


Figura 4.7. Representación de las dimensiones de la prueba de *Netiqueta*.

Por último, calculamos la estimación EAP/PV para investigar la consistencia interna de todas las dimensiones, con valores entre 0.78 y 0.88 (véanse las tablas 3_12 y 3_15), siendo todos los coeficientes superiores a 0.70 lo que indica una buena consistencia interna. El alfa de Cronbach para las pruebas globales fue de 0.93 (*IAD*) y 0.89 (*Netiqueta*).

4.4. Conclusiones y discusión

Para alcanzar los ODS de la UNESCO, es fundamental dotar a la ciudadanía de las capacidades adecuadas para utilizar la tecnología de forma significativa para participar en la sociedad actual, donde todos los ámbitos se ven afectados por los efectos de la tecnología digital en mayor o menor medida. En este contexto, el reconocimiento y acreditación de la CD es una de las principales líneas de actuación de la Comisión Europea en los últimos años junto con la promoción de un marco

común de referencia para la CD. Sin embargo, la mayoría de los sistemas de acreditación han sido generalmente insuficientes (Ferrari, 2013; Law et al., 2018; Santos y Serpa, 2017). Gracias al lanzamiento de DigComp y a las posibilidades ofrecidas por las TEA para desarrollar formatos de ítems innovadores como las simulaciones interactivas, se ha facilitado el desarrollo de implementaciones personalizadas de instrumentos para la evaluación de la CD, midiendo no sólo habilidades cognitivas de bajo orden. Además, la CD requiere una revisión constante debido a los constantes cambios en la forma en que los ciudadanos acceden y manejan la información a través de diferentes tipos de dispositivos. Con lo cual, llevar a cabo una implementación basada en un marco de referencia continuamente actualizado como es el caso del marco DigComp es crucial.

Los resultados presentados en este capítulo ayudan a entender los pasos llevados a cabo y las consideraciones y decisiones que se han tomado en cada paso y pueden servir de consideración para diseñar y validar una herramienta de evaluación de CD siguiendo enfoques similares. Revisamos las conclusiones asociadas a nuestros objetivos de investigación.

- PI_K1. *¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?*

  A lo lardo de este capítulo hemos mostrado los pasos dados en el estudio que llevamos a cabo haciendo uso de una metodología DBR para el diseño y validación de una herramienta de evaluación de CD orientada a la ciudadanía.

  La metodología seguida nos sirvió para describir los principios de diseño aplicados durante los diferentes pasos del desarrollo de las pruebas para las CD seleccionadas, con el fin de que puedan ser extendidos al resto de las CD incluidas en el marco de referencia DigComp, y para recabar las evidencias de validez y fiabilidad que nos ayuden a soportar la calidad de la herramienta desarrollada.

  El desarrollo se llevó a cabo en un proceso iterativo, que incluyó dos ciclos de pruebas y refinamiento, validando el contenido del constructo a evaluar y el diseño de los ítems, asegurando que se cubren los conocimientos y habilidades esperados, que los formatos de los ítems seleccionados son adecuados para ese objetivo, que la usabilidad de los ítems/pruebas es correcta y, por último, que los ítems están bien escritos y son fáciles de entender, lo cual es un factor clave en el desarrollo de un banco de ítems de CD.

  Nuestro objetivo era diseñar una herramienta de evaluación que incorporara diferentes formatos de ítems para evaluar también habilidades de orden superior, que normalmente son escasas. Por lo tanto, también diseñamos ítems

de acuerdo con los descriptores de los niveles intermedio y avanzado de DigComp. Elegir los formatos de ítems más adecuados, tales como las simulaciones interactivas desarrolladas a media, y diseñarlas correctamente para desencadenar los comportamientos previstos fue crucial en este punto. Esto es especialmente importante cuando se evalúan constructos cognitivos complejos como la CD, donde se requiere que los participantes pongan en acción sus conocimientos, lo que proporciona la imagen más precisa de su nivel de CD.

Uno de los puntos fuertes de nuestro estudio fue la selección de los casos de estudio. Hasta donde sabemos, la CD de Netiqueta no ha sido evaluada aún en profundidad como la que hemos llevado a cabo en nuestro estudio. En cuanto a la prueba de IAD, otros autores han diseñado pruebas de evaluación con objetivos similares, pero la mayoría de ellas han sido desarrolladas desde dos perspectivas, la bibliotecaria y la académica, y suelen ser de dominio específico, ya que muchas de ellas son herramientas de autoinforme o son herramientas basadas en preguntas de opción múltiple centradas en habilidades cognitivas de orden bajo (Catalano, 2016; Heer, 2012; Hollis, 2018).

Además, tuvimos que revisar en profundidad las CD para seleccionar los contenidos y SCs a evaluar en cada prueba, debido a su naturaleza de cambio constante. Durante la primera iteración con los facilitadores, parte de los esfuerzos se dedicaron a validar la propuesta de contenidos y SCs identificadas inicialmente a partir de la revisión bibliográfica. Así, seleccionamos dos casos de estudio siguiendo enfoques diferentes, una prueba basada en un AC (*IAD)*, y una prueba basada en una CD *(Netiqueta)*. Ambos enfoques son válidos, con sus singularidades, y han sido seleccionados por diferentes organismos para implementar sus herramientas de evaluación de CD, p.ej., basada en AC (Iglesias-Rodríguez et al., 2021) y basada en CD (BAIT).

Así pues, investigamos las peculiaridades de ambos enfoques y describimos los principios de diseño aplicados durante los distintos pasos del desarrollo de las pruebas.

- PI_K2. *¿Qué propiedades psicométricas tienen las pruebas? ¿Qué evidencias se pueden presentar que soporten las inferencias realizadas de las puntuaciones obtenidas?*

En el desarrollo de un instrumento cuantitativo con fines de evaluación, es crucial medir su calidad (Bandalos, 2018; Mueller y Knapp, 2018; Scholtes et al., 2011). Sin embargo, recientes revisiones de instrumentos para la evaluación de la CD concluyeron que las pruebas aportadas no son suficientes (Saltos-Rivas et al., 2021; Siddiq et al., 2016; Zhao et al., 2021). En vista de ello, diseñamos

nuestro estudio planificando varios estudios a lo largo de las diferentes fases para obtener suficientes evidencias que asegurasen la calidad de los instrumentos.

Comenzamos examinando las soluciones disponibles realizando una revisión de la literatura. Identificamos una falta de instrumentos adecuados para evaluar la CD de las personas, identificando las principales carencias: herramientas que sólo evalúan habilidades cognitivas de bajo orden, herramientas basadas en aplicaciones y dispositivos que no reflejan la realidad actual, y falta de suficientes evidencias de validez y fiabilidad que ayuden a soportar la calidad de las herramientas desarrolladas. Tras la revisión de los conocimientos y las prácticas actuales, definimos nuestros objetivos basándonos en los principios de diseño identificados en varios estudios clave, así como de la información intercambiada con expertos en la The DigComp Community of Practice.

A continuación, llevamos a cabo un estudio de validación basado en el juicio de expertos (facilitadores de los centros de CD de la red Kzgunea) para confirmar que el contenido de las pruebas representaba el constructo que se pretendía evaluar y que era adecuado para cumplir los objetivos de la prueba.

Además, llevamos a cabo un estudio para validar los RP de una selección de ítems incluidos en las pruebas, obteniendo información útil para comprender el rendimiento de los participantes e investigar si los criterios de evaluación de cada ítem estaban correctamente establecidos. Para el resto de los ítems incluidos en las pruebas, aplicamos los principios de diseño identificados en el estudio.

Por otra parte, aunque la multidimensionalidad del constructo de CD ha sido identificada en varios estudios, los estudios teóricos y empíricos han reportado resultados contradictorios (Reichert et al., 2020). Más concretamente, Vuorikari et al. [25] describieron teóricamente el constructo de CD en DigComp, pero las dimensiones identificadas no han sido confirmadas empíricamente o requieren más investigación. Para la prueba de *IAD*, consideramos las tres CD del AC como dimensiones independientes y mostramos como el modelo Rasch tridimensional se ajusta mejor a los datos que el modelo Rasch unidimensional. Para la prueba de *Netiqueta* consideramos las cuatro SC seleccionadas como dimensiones independientes, y mostramos como el modelo Rasch cuatridimensional se ajusta mejor a los datos que el modelo Rasch unidimensional. Sin embargo, teniendo en cuenta las altas correlaciones obtenidas, parece que todos los ítems se relacionan con un factor fuerte, que puede interpretarse como CD general. Estudios recientes como el llevado a cabo por Clifford et al. (2020), han apuntado en la misma dirección.

Además, calculamos la estimación EAP/PV para investigar la consistencia interna de todas las dimensiones, con valores entre 0.84 y 0.88 para la prueba de IAD y valores entre 0.78 y 0.86 para la prueba de Netiqueta, siendo todos los coeficientes superiores a 0.70 lo que indica una buena consistencia interna. El alfa de Cronbach para las pruebas globales fue de 0.93 para la prueba de IAD y 0.89 para la prueba de Netiqueta.

Ambas pruebas mostraron propiedades psicométricas sólidas que los convierten en instrumentos fiables y válidos para medir la CD. Cabe mencionar que los dos enfoques, basados en AC y basados en CD, difieren en el grado de profundidad en la evaluación de las CD cubiertas por sus respectivas pruebas. Es decir, en el caso de la prueba de *Netiqueta*, el número de preguntas por DC era mucho mayor que en el caso de las tres CD incluidas en el test de *IAD*.

Teniendo en cuenta los resultados obtenidos durante las pruebas en el segundo ciclo iterativo con usuarios finales, son interesantes los resultados obtenidos en la prueba de *Netiqueta*, en la que las mujeres obtuvieron puntuaciones medias más altas que los hombres. Además, los participantes de entre 55 y 74 años obtuvieron mejores resultados. El número de participantes de esta franja de edad fue muy bajo (7,4%). Sería interesante confirmar estos resultados administrando la prueba a un mayor número de participantes, incluyendo más personas de esta franja de edad.

Sin embargo, cabe señalar algunas limitaciones a la hora de interpretar los resultados de este estudio. La muestra de participantes en el segundo ciclo iterativo fue relativamente pequeña teniendo en cuenta el grupo de interés, especialmente en algunos rangos de edad. Aunque el principal rango de edad de interés de nuestro estudio es el de 25-54 años, sería interesante ver si los resultados se replican con una muestra más grande y variada.

Otro punto de interés es que, aunque la herramienta ha sido diseñada para la ciudadanía en general, en el proceso de validación con los facilitadores de KZgunea los criterios de validación que se aplicaron podrían no haberse aplicado en otras regiones, p.ej., los facilitadores de otras redes de telecentros podrían considerar otras SC de mayor interés. Por lo tanto, puede haber cierta variabilidad entre las distintas regiones que puede requerir adaptaciones en las herramientas.

Así mismo, también queremos mencionar algunos puntos débiles. Desarrollar simulaciones interactivas e integrarlas en las herramientas de evaluación requiere conocimientos técnicos altos y un gran esfuerzo. Además, es necesario diseñar las simulaciones interactivas a partir de aplicaciones que sean lo más neutras posibles, para que los participantes que no las hayan utilizado puedan realizar las tareas de forma lógica y que al menos esas aplicaciones tengan comportamientos lo más extendidos y aceptados posibles (p.ej., el modo de desplegarse los menús y forma

de distribuir las distintas opciones, despliegue de menús contextuales desde el botón derecho del ratón, etc.). Por desgracia, la mayoría de las aplicaciones se actualizan constantemente, introduciendo muchas de ellas importantes cambios de diseño. Este hecho aconseja revisar constantemente las simulaciones, examinando si los cambios son lo suficientemente importantes como para considerar su rediseño. En este sentido, nos encontramos en un mundo en rápida transformación debido a los constantes avances tecnológicos con la continua aparición de nuevas tecnologías y aplicaciones, los cuales generan cambios constantes en el comportamiento y los hábitos de las personas. Por lo tanto, la CD es compleja y se encuentra en constante evolución, y debe ser revisada periódicamente. Del mismo modo, los criterios de evaluación también deben revisarse constantemente, ya que algunas SC pueden quedar obsoletas y pueden surgir otras nuevas. Las revisiones y actualizaciones deberán ser periódicas y constantes.

Las organizaciones interesadas en desarrollar sus propias implementaciones podrían encontrar este estudio de interés para decidir qué principios de diseño podrían ser de interés para sus propias implementaciones de acuerdo con sus necesidades.

Respecto a los distintos formatos de ítems a utilizar, DigComp abarca 21 CD diferentes, cada una de ellas con sus propias peculiaridades. Dependiendo de los descriptores de cada CD, puede ser conveniente utilizar un formato de ítems u otro. En futuros estudios se podría considerar la incorporación de nuevos formatos de ítems, que desencadenen RP más complejos como, p.ej., basados en juegos desarrollados a medida.

Las palabras son vida. Si tus ojos pudieras hablar, ¿qué dirían?

Max Ben Schnetzer

5.

Análisis de los procesos de respuesta al ítem

La evaluación de la validez de una prueba requiere tener en cuenta no sólo el contenido de los ítems de evaluación, sino también si los participantes responden a los ítems desplegando RP relevantes para el constructo. El análisis de los datos generados a partir de los RP durante las evaluaciones puede ayudar a validar el diseño de los ítems de evaluación, presentando evidencias de que los ítems desencadenaron los conocimientos y habilidades esperados. Sin embargo, examinar si los participantes responden a las tareas realizando RP relevantes para el constructo, no suele tenerse en cuenta a la hora de examinar la validez de una prueba. Recientes revisiones sobre la fiabilidad y la validez de las herramientas para la evaluación de la CD realizadas por Saltos-Rivas et al. (2021), Siddiq et al. (2016) y Zhao et al. (2021) han llegado a la conclusión de que se dispone de muy poca información sobre las formas de garantizar la validez y la fiabilidad del instrumento utilizado. Más aún, en lo referente a los diferentes tipos posibles de evaluación de la calidad, las evidencias basadas en RP son escasas o inexistentes.

Los métodos de ET amplían aún más la información basada en RP, proporcionando observaciones del examinando a un nivel más detallado (Oranje et al., 2017), que no suelen ser captados por otras fuentes de datos de procesos como los protocolos de pensamiento en voz alta o los registros generados por ordenador. Sin embargo, hay relativamente pocos estudios que abarquen ET en el contexto de la evaluación educativa y menos todavía en el área de la evaluación de la CD (Bartolomé et al., 2020). Los datos de la mirada proporcionan información detallada sobre las estrategias de respuesta de los participantes a los ítems de las pruebas, lo que permite perfilar el compromiso del examinando y los RP de un rendimiento satisfactorio.

Por este motivo, durante el desarrollo de ETCD realizamos este estudio exploratorio usando datos de ET dentro del proceso de validación de las pruebas. Como parte del argumento de validez, mostramos cómo los RP apoyados con datos de ET pueden revelar información importante sobre las diferencias que no suelen ser captadas por medios tradicionales.

La sugerencia de Cronbach (1980) a las pruebas de validez es relevante en este contexto, ya que no aportan pruebas directas de que el proceso cognitivo implicado en los RP sea el mismo utilizado en una situación real que en resolver el ítem, pero pueden contribuir a demostrar que no es equivalente (Kane y Mislevy, 2017). Por ejemplo, si los datos del ET demuestran que los participantes resolvieron las tareas presentadas en las simulaciones sin examinar detenidamente el enunciado y las instrucciones, se desvirtuaría que estos ítems requerían el mismo proceso cognitivo que se requiere en una tarea real como la planteada en el ítem.

Además, los datos del ET podrían mostrar cómo participantes con diferentes niveles de CD podrían mostrar diferentes RP. Es decir, los datos de ET no proporcionan una confirmación directa del proceso cognitivo, pero las pruebas proporcionadas pueden utilizarse para obtener inferencias útiles sobre el proceso (Fitts, Jones y Milton, 1950).

El objetivo es validar las inferencias realizadas entre las afirmaciones y el comportamiento observado, evaluando interpretaciones alternativas sobre cómo los participantes procesaron la información de los ítems en función de las AOIs examinadas y del orden de AOIs seguido. Los resultados presentados en las siguientes secciones ayudan a confirmar que los RP de los participantes que respondieron correctamente no representan una alternativa a la interpretación de los resultados que pusiera en evidencia los criterios de evaluación definidos.

5.1. Objetivos de investigación

Con el objetivo de dar respuesta a las siguientes preguntas de investigación, llevamos a cabo el siguiente estudio exploratorio mediante la tecnología de ET, la cual propusimos a un grupo aleatorio de empleados de Tecnalia y analizamos sus registros:

- PI_K3. *¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos generar y probar inferencias sobre el constructo de interés?*
- PI_K4. *¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos utilizar esta información para mejorar los diseños de los ítems?*

5.2. Metodología

Llevamos a cabo un estudio exploratorio utilizando los MOs de los participantes para proporcionar información sobre su rendimiento en las pruebas de evaluación.

En la primera parte del estudio, recogimos las interacciones de los participantes durante las pruebas para determinar si los ítems de evaluación diseñados desencadenan los conocimientos y habilidades previstos y sirven para diferenciar participantes con distintos niveles de pericia según los criterios de evaluación definidos para cada ítem. Analizamos principalmente las variaciones entre los participantes con diferentes niveles de pericia utilizando métricas basadas en la fijación, apoyadas por técnicas de visualización.

En la segunda parte del estudio, evaluamos la credibilidad de una interpretación alternativa de los resultados de las pruebas, en la cual los participantes utilizan una estrategia que sólo puede estar relacionada de forma imprecisa con la CD que se pretende evaluar. Con este objetivo, examinamos las trayectorias de exploración de los RP desplegados por los participantes durante la resolución de los ítems basados en una imagen o simulación. Para ello, creamos las rutas de exploración en función de los elementos visuales visitados, y exploramos cómo el análisis de estas rutas puede contribuir al argumento de validez de las inferencias basadas en las puntuaciones.

En esta segunda parte del estudio, nos centramos en las preguntas basadas en una imagen o simulación, ya que en las simulaciones interactivas podemos identificar la ruta de exploración seguida por los participantes gracias al registro de clics. Además, este tipo de preguntas puede ser adecuado para evaluar las habilidades de orden superior, en las cuales los participantes deben evaluar críticamente las situaciones presentadas en la imagen o simulación y elegir la opción correcta de las posibles. Más aun, para este tipo de ítems basados los datos de la mirada son la única fuente disponible para describir la estrategia de los participantes y analizar las variaciones entre ellos. Así pues, utilizamos los datos del ET para validar las inferencias realizadas entre las afirmaciones y el comportamiento observado, centrándonos en dos posibles vías, evaluando interpretaciones alternativas de cómo los participantes procesaron la información de los ítems en términos de:

- AOI examinados.
- AOI examinados y orden seguido examinando los diferentes AOI.

Para la primera vía, comenzamos llevando a cabo un análisis de la trayectoria de exploración, examinando si los participantes con éxito tienen trayectorias de exploración más largas en comparación con los participantes sin éxito, y si había alguna diferencia entre las puntuaciones y el tipo de ítem (si el ítem requiere un

enfoque sistemático para resolverlo o no). Además, analizamos la influencia de cada una de las AOI con un doble propósito. En primer lugar, queríamos identificar las áreas más relevantes para evaluar correctamente cada pregunta. Para ello, realizamos un análisis de correlación de características basado en la información mutua entre cada una de las AOIs visitadas y el resultado de la interacción. En segundo lugar, queríamos detectar patrones de comportamiento que anticiparan el resultado de la interacción. Para ello, creamos un clasificador para predecir la respuesta de los participantes en función del AOI visitado.

Para la segunda vía, evaluamos interpretaciones alternativas basadas en las AOI examinadas y el orden de procesamiento de las AOI. Para ser más prácticos, dejamos de lado el ítem 21 porque era muy similar al ítem 4. En primer lugar, creamos las rutas de exploración en función de los elementos visuales de las imágenes incluidas. Por ejemplo, si un participante fijaba los elementos A, B y C respectivamente (es decir, sus fijaciones caían en los elementos A, B y C respectivamente), su trayectoria de exploración se generaba como ABC manteniendo las duraciones de fijación. También decidimos eliminar los datos del participante 28 en el elemento 24 y del participante 5 y 31 en el elemento 5, ya que tuvimos problemas para registrar sus interacciones y su información era inconsistente. A continuación, utilizamos el algoritmo String-edit, ampliamente en la investigación de ET (Tai et al., 2006), para investigar la varianza dentro de los grupos, apoyado con el uso de la herramienta ScanGraph (Dolezalova y Popelka, 2016). Luego, modelamos el procesamiento de los elementos por parte de los participantes con éxito, aplicando un modelo ponderado basado en la posición, para identificar la trayectoria de exploración común que representa a todo el grupo. De este modo, podemos validar de forma muy sencilla que la trayectoria de exploración común no representa una alternativa a la interpretación de los resultados que pondría en evidencia los criterios de evaluación definidos. Además, estimamos la dificultad de los ítems analizando las duraciones de las fijaciones y los lugares donde se producen las fijaciones más largas. Por último, examinamos los lugares que atraen las fijaciones más largas para comprobar si el comportamiento mostrado por los participantes es el esperado según los criterios de evaluación definidos para cada ítem, con el fin de rechazar una interpretación alternativa de las puntuaciones de la prueba.

5.2.1. Participantes

Un total de 30 participantes (15 hombres y 15 mujeres) con diferentes niveles de CD y trabajadores de Tecnalia Research & Innovation[53], fueron seleccionados para el estudio. Cabe destacar que el número de participantes necesario para analizar las secuencias de MOs no ha sido estudiado en profundidad en la literatura, tal y como mencionan King et al. (2019). A pesar de este hecho, algunos estudios relevantes destinados a examinar las diferencias entre diferentes grupos de participantes

[53] https://www.tecnalia.com/

también se han llevado a cabo con tamaños de muestra similares, p. ej., (Andrzejewska y Stolińska, 2016; Eraslan et al., 2019; Kucharský et al., 2020; Inal, 2016; León et al., 2019). El tamaño limitado de la muestra de este estudio exploratorio se debió en parte a la profundidad del análisis que queríamos realizar y al esfuerzo necesario en llevarlo a cabo.

En la selección de los participantes no se tuvo en cuenta el número de hombres y mujeres, ya que no teníamos previsto analizar la información por género. El único criterio de selección fue estar en el rango de edad de 25 a 54 años, que es el rango de edad que representa más del 90% de los destinatarios del servicio BAIT. Asimismo, a pesar de que no todos los participantes tenían una visión perfecta, esto no fue una fuente potencial de variabilidad entre los participantes ya que el dispositivo de ET utilizado no fue intrusivo.

En las tablas 5.1, 5.2, 5.3 y 5.4 se muestran una descripción de la autoevaluación que hizo cada participante sobre su nivel en cada CD, así como los resultados que obtuvieron en las pruebas.

Exam. ID	Autoeval. SC1	Prueba SC1 (%)	Autoeval. SC2	Prueba SC2 (%)	Autoeval. SC3	Prueba SC3 (%)	Autoeval. SC4	Prueba SC4 (%)	Total (%)
1	Avanzado	50%	Avanzado	100%	Básico	100%	Básico	50%	70%
2	Básico	0%	Básico	100%	Básico	33,3%	Básico	0%	20%
3	Intermedio	50%	Intermedio	0%	Intermedio	66,7%	Intermedio	100%	60%
4	Básico	50%	Básico	100%	Básico	66,7%	Básico	100%	70%
5	Intermedio	75%	Intermedio	100%	Intermedio	100%	Intermedio	50%	80%
6	Intermedio	50%	Intermedio	100%	Básico	66,7%	Intermedio	50%	60%
7	Intermedio	50%	Intermedio	0%	Intermedio	66,7%	Básico	50%	50%
8	Básico	75%	Básico	100%	Básico	66,7%	Básico	50%	70%
9	Avanzado	75%	Intermedio	100%	Intermedio	100%	Intermedio	50%	80%
10	Avanzado	50%	Intermedio	0%	Avanzado	33,3%	Intermedio	50%	40%
11	Intermedio	50%	Intermedio	100%	Intermedio	33,3%	Intermedio	100%	60%
12	Básico	50%	Intermedio	0%	Intermedio	33,3%	Intermedio	50%	40%
13	Básico	25%	Básico	0%	Básico	100%	Intermedio	0%	40%
14	Avanzado	25%	Avanzado	100%	Intermedio	100%	Intermedio	50%	60%
15	Intermedio	25%	Avanzado	100%	Avanzado	66,7%	Avanzado	0%	40%
16	Intermedio	75%	Intermedio	100%	Intermedio	100%	Intermedio	100%	90%

17	Intermedio	25%	Intermedio	0%	Intermedio	66,7%	Intermedio	50%	40%
18	Intermedio	50%	Intermedio	0%	Intermedio	33,3%	Intermedio	50%	40%
19	Avanzado	75%	Avanzado	100%	Avanzado	100%	Intermedio	100%	90%
20	Avanzado	75%	Avanzado	100%	Avanzado	66,7%	Avanzado	50%	70%
21	Básico	50%	Básico	0%	Básico	33,3%	Básico	0%	30%
22	Intermedio	25%	Intermedio	0%	Intermedio	100%	Intermedio	50%	50%
23	Intermedio	75%	Intermedio	100%	Intermedio	100%	Intermedio	100%	90%
24	Básico	75%	Básico	100%	Básico	100%	Básico	100%	90%
25	Intermedio	0%	Intermedio	0%	Intermedio	66,7%	Intermedio	50%	30%
26	Básico	50%	Básico	100%	Básico	66,7%	Básico	50%	60%
27	Básico	25%	Básico	100%	Básico	66,7%	Intermedio	50%	70%
28	Intermedio	25%	Intermedio	100%	Intermedio	66,7%	Básico	50%	50%
29	Básico	75%	Intermedio	100%	Básico	66,7%	Básico	100%	80%
30	Básico	25%	Intermedio	0%	Intermedio	100%	Intermedio	100%	60%

Tabla 5.1. Netiqueta, resultados de la autoevaluación y los obtenidos en la prueba.

Exam. ID	Autoeval. SC1	Prueba SC1 (%)	Autoeval. SC2	Prueba SC2 (%)	Total (%)
1	Avanzado	77,8%	Avanzado	100%	85,7%
2	Intermedio	11,1%	Intermedio	20%	14,3%
3	Intermedio	33,3%	Intermedio	100%	57,1%
4	Avanzado	77,8%	Básico	80%	78,6%
5	Avanzado	77,8%	Avanzado	80%	78,6%
6	Intermedio	55,6%	Intermedio	100%	71,4%
7	Avanzado	55,6%	Intermedio	100%	71,4%
8	Avanzado	77,8%	Avanzado	100%	85,7%
9	Avanzado	33,3%	Básico	60%	42,9%
10	Avanzado	55,6%	Intermedio	60%	57,1%
11	Avanzado	77,8%	Intermedio	80%	78,6%
12	Avanzado	44,4%	Intermedio	60%	50,0%
13	Avanzado	77,8%	Intermedio	100%	85,7%
14	Avanzado	100%	Avanzado	100%	100%

15	Avanzado	66,7%	Intermedio	100%	78,6%
16	Intermedio	77,8%	Intermedio	100%	85,7%
17	Intermedio	55,6%	Intermedio	60%	57,1%
18	Avanzado	88,9%	Avanzado	100%	78,6%
19	Avanzado	77,8%	Avanzado	100%	85,7%
20	Avanzado	77,8%	Avanzado	100%	85,7%
21	Básico	44,4%	Básico	20%	35,7%
22	Avanzado	55,6%	Intermedio	80%	64,3%
23	Avanzado	88,9%	Intermedio	100%	92,9%
24	Intermedio	77,8%	Intermedio	80%	78,6%
25	Intermedio	55,6%	Intermedio	60%	57,1%
26	Intermedio	55,6%	Básico	60%	57,1%
27	Intermedio	77,8%	Básico	40%	64,3%
28	Avanzado	77,8%	Intermedio	100%	85,7%
29	Intermedio	55,6%	Básico	80%	64,3%
30	Intermedio	44,4%	Intermedio	60%	50%

Tabla 5.2. CD1, resultados de la autoevaluación y los obtenidos en la prueba.

Exam. ID	Autoeval. SC3	Prueba SC3 (%)	Autoeval. SC4	Prueba SC4 (%)	Total (%)
1	Avanzado	50%	Avanzado	80%	61,5%
2	Intermedio	25%	Intermedio	40%	30,8%
3	Intermedio	87,5%	Intermedio	40%	69,2%
4	Intermedio	25%	Básico	0%	15,4%
5	Avanzado	37,5%	Avanzado	60%	46,2%
6	Intermedio	12,5%	Intermedio	40%	38,5%
7	Avanzado	75%	Intermedio	40%	61,5%
8	Avanzado	50%	Avanzado	60%	53,8%
9	Intermedio	50%	Intermedio	40%	46,2%
10	Intermedio	75%	Intermedio	60%	69,2%
11	Intermedio	12,5%	Intermedio	80%	61,5%
12	Intermedio	62,5%	Intermedio	40%	53,8%
13	Básico	75%	Intermedio	80%	76,9%
14	Avanzado	75%	Avanzado	100%	84,6%
15	Avanzado	87,5%	Avanzado	60%	76,9%
16	Intermedio	62,5%	Intermedio	100%	76,9%
17	Básico	62,5%	Intermedio	60%	61,5%

18	Intermedio	62,5%	Intermedio	60%	61,5%
19	Intermedio	100%	Intermedio	60%	84,6%
20	Avanzado	75%	Avanzado	60%	69,2%
21	Básico	62,5%	Básico	80%	69,2%
22	Intermedio	62,5%	Intermedio	80%	69,2%
23	Intermedio	62,5%	Intermedio	80%	69,2%
24	Intermedio	50%	Intermedio	80%	61,5%
25	Intermedio	75%	Intermedio	100%	84,6%
26	Básico	12,5%	Intermedio	20%	15,4%
27	Básico	75%	Básico	40%	61,5%
28	Intermedio	75%	Intermedio	60%	69,2%
29	Intermedio	62,5%	Básico	60%	61,5%
30	Básico	62,5%	Básico	80%	69,2%

Tabla 5.3. CD2, resultados de la autoevaluación y los obtenidos en la prueba.

Exam. ID	Autoeval. SC5	Prueba SC5 (%)	Total (%)
1	Avanzado	90%	90%
2	Básico	50%	50%
3	Básico	80%	80%
4	Intermedio	90%	90%
5	Intermedio	60%	60%
6	Intermedio	20%	20%
7	Avanzado	50%	50%
8	Básico	80%	80%
9	Intermedio	70%	70%
10	Intermedio	90%	90%
11	Intermedio	50%	50%
12	Intermedio	80%	80%
13	Básico	100%	100%
14	Avanzado	90%	90%
15	Avanzado	90%	90%
16	Intermedio	100%	100%
17	Intermedio	40%	40%
18	Intermedio	60%	60%

19	Básico	80%	80%
20	Básico	80%	80%
21	Básico	30%	30%
22	Avanzado	100%	100%
23	Avanzado	90%	90%
24	Intermedio	90%	90%
25	Intermedio	80%	80%
26	Básico	60%	60%
27	Básico	80%	80%
28	Intermedio	80%	80%
29	Básico	70%	70%
30	Básico	80%	80%

Tabla 5.4. CD3, resultados de la autoevaluación y los obtenidos en la prueba.

En la tabla 5.5 se muestran los descriptores generales de cada nivel de CD.

Nivel de competencia	Descriptores generales de nivel
Básico	Tengo conocimientos básicos y, o no tengo habilidades, o son muy pobres. Tareas sencillas y normalmente con ayuda.
Medio	Entiendo los conceptos fundamentales y tengo algunas habilidades para operar por mi cuenta. Tareas bien definidas.
Avanzado	Tengo conocimientos avanzados y sé cómo aplicarlos para realizar las tareas más adecuadas. Incluso puedo guiar a otras personas.

Tabla 5.5. Descriptores generales de cada nivel de CD.

La mayoría de los participantes consideraron sus niveles *avanzados* o *intermedios*. Sólo en *Netiqueta* hubo una disminución del número de niveles *avanzados* y un aumento del número de niveles *básicos*, probablemente debido al desconocimiento de esa CD.

5.2.2. Materiales

Los participantes completaron individualmente los seis estudios diseñados en Tobii Pro Lab (TPL)[54] configurados con ítems de ETCD, pertenecientes a la CD de *Netiqueta* y al AC de *IAD*. Los diferentes formatos de ítems seleccionados fueron:

- Ítems basados en una imagen o simulación, en las que los participantes tienen que examinar una imagen o simulación y elegir la opción correcta. Este formato es adecuado para evaluar habilidades de orden cognitivo superior, que requieren que los participantes evalúen críticamente las situaciones y apliquen sus conocimientos en la resolución del ítem.
- Simulaciones interactivas, en las que se simulan diferentes situaciones, y los participantes tienen que interactuar para resolver la tarea demandada (p. ej., acceder a los marcadores del navegador o responder a un correo electrónico). Además, las diseñamos para contemplar los diferentes caminos posibles para resolver las tareas, y establecemos un número máximo de clics erróneos en función del grado de dificultad que intuimos. De este modo, los participantes pueden explorar hasta cierto punto en los programas y situaciones presentadas, y no tienen que saber de memoria dónde tiene que estar cada opción, favoreciendo el juicio y la toma de decisiones.
- Ítems de opción múltiple, ampliamente utilizados en la evaluación educativa.
- Tareas abiertas, en las que los participantes tienen que interactuar con la estación de trabajo y sus programas para resolver la tarea requerida, p. ej., descargar un archivo comprimido, extraer el contenido y localizar un dato.

Los detalles de los ítems seleccionados en cada estudio diseñado en TPL se muestran en la tabla 5.6.

TPL (CD)	Ítem	SC	(Formato) Enunciado	Comportamiento esperado y criterios de evaluación
Web (Net)	4	SC1	(Simulación) Indicar si cumple las normas de netiqueta o si algún campo las incumple.	Deben examinar los campos de un correo electrónico y responder a la pregunta. En concreto, el asunto es demasiado corto, sin dar suficiente información.
	10	SC1	(Simulación) Indicar cumple las normas de netiqueta o si algún campo las incumple.	Deben examinar los campos de un correo electrónico y responder a la pregunta. En concreto, el correo electrónico es correcto.
	16	SC1	(Simulación) Indicar cumple las normas de netiqueta o si algún campo las incumple.	Deben examinar los campos de un correo electrónico y responder a la pregunta. En concreto, el cuerpo no da suficiente información de contacto del usuario

[54] https://www.tobii.com/products/software/data-analysis-tools/tobii-pro-lab

				como número de teléfono, dirección, etc.
	21	SC1	(Simulación) Indicar cumple las normas de netiqueta o si algún campo las incumple.	Deben examinar los campos de un correo electrónico y responder a la pregunta. En concreto, el asunto tiene un error ortográfico.
	34	SC2	(Simulación) Las situaciones mostradas (mensajes de diferentes tipos de comunicaciones) son:	Deben evaluar las 4 conversaciones. Tres de ellas incluyen errores ortográficos, pero la de Twitter con hashtags es correcta.
	38	SC3	(Basada en imagen) Filtrar los resultados de la búsqueda mostrando imágenes que se puedan utilizar, compartir y modificar libremente, incluso con fines comerciales.	Deben evaluar la situación y centrar su atención en el menú horizontal con las opciones de filtrado (Hay dos posibles caminos correctos).
	39	SC3	(Simulación) Una vez registrado e identificado en un foro por primera vez, llevar a cabo la primera acción necesaria antes de publicar.	Deben examinar la situación, comprobando la cabecera y observar que el usuario ya está identificado y preparado para publicar un mensaje. A continuación, deben examinar la lista de mensajes y accede al primero que incluye las normas a leer antes de publicar un mensaje.
	45	SC3	(Simulación) Comprobar la bandeja de entrada de correo electrónico y marcar como SPAM el mensaje correspondiente.	Deben comprobar el asunto de los correos electrónicos de la bandeja de entrada. Comprobando el asunto puede ser suficiente para identificarlo, pero también es posible acceder a los detalles de cada mensaje. Deben marcar como SPAM el mensaje.
	61	SC4	(Simulación) Bloquear la cuenta mostrada a la que sigues porque consideras su comportamiento no adecuado.	Deben evaluar la situación y acceder a "Siguiendo". Interesante ver si los participantes que normalmente no trabajan con Twitter tuvieron algún problema.
	65	SC4	(Simulación) Las situaciones mostradas son ejemplos de: lenguaje inapropiado, falta de respeto, formas incorrectas de escribir, abuso de poder.	Deben evaluar dos noticias de dos periódicos distintos. El título de ambas noticias está mal escrito y el significado no es el deseado.
Web (CD1)	1	SC5	(Opción múltiple) Identificar el filtro adecuado para obtener los números entre dos rangos (entre 5 y 10).	Deben evaluar las instrucciones, declaración y opciones, así como la imagen decorativa, que no debería llamar mucho su atención.
	4	SC5	(Simulación) Filtrar los resultados de	Deben evaluar la situación, para localizar el menú de

			la búsqueda, mostrando imágenes de tamaño medio y etiquetadas para su reutilización con modificaciones.	filtros y seleccionar la opción "Imágenes".
	9	SC5	(Opción múltiple) ¿Qué es un buscador?	Deben evaluar las instrucciones, enunciado y opciones. así como la imagen decorativa, que no debe llamar mucho su atención.
	12	SC5	(Simulación) Buscar información sobre 'cita previa' mediante una búsqueda rápida que sólo muestre información de la web de Osakidetza.	Deben evaluar la situación, introducir los términos de búsqueda y hacer clic en el botón "buscar en Google".
	26	SC5	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben evaluar las 5 instrucciones, enunciado, pestaña "imagen/opciones", opciones e imagen). Interesante comprobar la usabilidad de la pregunta, así como las AOIs de la imagen examinadas.
	44	SC6	(Simulación) Lleva a cabo una búsqueda inversa de imágenes seleccionando la imagen del perro de tu dispositivo.	Deben evaluar la situación, y hacer clic en 'Imágenes' del menú derecho. A continuación, deberán seleccionar la imagen de búsqueda.
	50	SC6	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben evaluar las 5 instrucciones, enunciado, pestaña "imagen/opciones", opciones e imagen). Interesante comprobar la usabilidad de la pregunta, así como las AOIs de la imagen examinadas.
	51	SC6	(Simulación) Acceder a la web de Tecnalia almacenada en tus favoritos.	Deben evaluar las instrucciones, enunciado y opciones del menú del navegador, así como el menú del buscador Bing que podría llegar a confundirles si no tuvieran claros los conceptos y el contexto.
	57	SC6	(Simulación) Acceder a las contraseñas guardadas en el navegador, mostrando la primera de la lista.	Deben evaluar las instrucciones, enunciado, número de errores y opciones del menú del navegador. Tienen que hacer clic en el menú principal del navegador para acceder a las contraseñas almacenadas.
	62	SC6	(Simulación) Crear una alerta que notifique la presencia en Internet de "Alem Tellaeche", notificando cuando se produzcan.	Deben evaluar las instrucciones, enunciado, área de creación de alerta y área de alerta de presencia en Internet, para identificar la alerta previamente creada y editarla.
Web	4	SC7	(Basada en imagen) De acuerdo con las situaciones mostradas en la	Deben examinar las dos noticias, y se espera que examinen las principales AOIs identificadas (la

(CD2)			imagen, seleccionar la afirmación correcta.	cabecera de la noticia, logotipo, imágenes, evidencias, etc.),
	6	SC7	(Basada en imagen) En cuanto a la fiabilidad de esta cuenta, selecciona el enunciado correcto.	Deben un canal de YouTube, prestando atención a el logotipo del canal, el número de suscriptores y el icono de "check" (elemento clave).
	16	SC7	(Basada en imagen) De los resultados de la búsqueda mostrados, selecciona la web oficial de Euskaltel.	Deben examinar los resultados de una búsqueda con enlaces a diferentes sitios web, para identificar el sitio oficial, en base a la URL y su dominio.
	18	SC7	(Basada en imagen) De acuerdo a la web mostrada, ¿en qué situación nos encontramos?	Deben examinar un diario que muestra diferentes noticias e imágenes, prestando atención por lo menos al logotipo, la fecha y el autor de la noticia principal.
	24	SC7	(Basada en imagen) De acuerdo con la web mostrada, ¿en qué situación nos encontramos?	Deben examinar la página de identificación de una cuenta bancaria. La URL es sospechosa y el sitio debe ser marcado como no confiable.
	30	SC8	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben examinar una revista que muestra los detalles de una noticia, prestando atención a el logotipo, así como especialmente al autor y la fecha de la noticia.
	32	SC8	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben examinar un periódico que muestra los detalles de una noticia. Deben prestar atención a la url, el logotipo, así como especialmente al autor y la fecha de la noticia.
	37	SC8	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben examinar un texto y darse cuenta de que carece de información de anclaje.
	42	SC8	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben examinar un periódico que muestra una noticia. Deben prestar atención a la cabecera de la noticia, así como especialmente al autor y su fecha.
	44	SC8	(Basada en imagen) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deben examinar un periódico que muestra una noticia. Deben prestar atención al logotipo, así como especialmente al encabezamiento de la noticia, el autor y la fecha.
Web	4	SC3	(Simulación) Utilizar el Explorador de Windows para crear una carpeta	Deben acceder al escritorio de Windows partiendo de un explorador de Windows (hay tres caminos

(CD3)			en el escritorio.	posibles).
	5	SC3	(Simulación) Eliminar el marcador de la web de Tecnalia de los marcadores.	Deben evaluar las instrucciones, enunciado, botón de "tres puntos" cerca del campo url, icono de "catálogo" e icono de "tres barras horizontales" para acceder al menú del navegador. Si pierden el tiempo en el menú de Bing es que no tienen los conceptos claros.
	6	SC3	(Simulación) Crear una carpeta en Google Drive con el nombre "*Facturas*" y subir el archivo "*Gas_Factura_Enero*".	Deben evaluar la página principal de Google Drive desde que tienen que crear la carpeta haciendo clic en el botón "plus".
	16	SC3	(Simulación) Subir el archivo "Factura_Gas_Enero.pdf" de tu dispositivo al directorio Facturas2019 (ya creado) en tu Dropbox.	Deben evaluar la página principal de Dropbox donde tienen que empezar a subir el archivo haciendo clic en el botón "Subir archivos". Interés en ver si tienen problemas para subir el archivo sin seleccionar primero la carpeta de destino.
	18	SC3	(Simulación) Utilizar la app InfoJobs de tu carpeta 'Empleo', y marca la oferta de Administrativo en Sopela como de interés.	Deben evaluar la entrada principal de una aplicación, desde la que podrán acceder al detalle de la oferta de empleo o marcarla directamente como "favorita" sin acceder al detalle de la oferta.
	23	SC3	(Simulación) Responde al correo recibido adjuntando el archivo comprimido del CV de tu carpeta "CV" en tu Google Drive.	Deben evaluar la bandeja de entrada de Google Gmail donde es visible el correo electrónico mencionado en el enunciado, al que tienen que responder adjuntando el CV desde Google Drive.
	24	SC3	(Simulación) Activar la sincronización del navegador para acceder a los marcadores del navegador con tu cuenta de Google.	Deben prestar atención especialmente al menú horizontal del navegador y hacer clic en el icono "usuario" para acceder a la opción de sincronización.
	28	SC3	(Simulación) Eliminar de los marcadores, de la carpeta Deportes, la web Marca.com.	Deben prestar atención especialmente al menú del navegador y hacer clic en el botón de los tres puntos para acceder al menú del navegador, teniendo en cuenta que el menú del buscador Bing puede confundir a los que no tengan claros los conceptos y el contexto.
	41	SC3	(Simulación) Guardar el sitio web actual utilizando la extensión Pocket.	Deben prestar atención especialmente al área donde suelen aparecer las extensiones en los navegadores, para encontrar la extensión de Pocket.
	49	SC3	(Simulación) Responde al correo	Deben prestar atención a la funcionalidad de

			electrónico recibido adjuntando el CV comprimido almacenado en tu dispositivo móvil.	adjuntar archivos de Google Gmail.
Rec. (CD1)	16	SC5	(Interactuar puesto) Introducir el nombre del país que ha vendido más unidades (información dentro de una hoja de cálculo proporcionada).	Tendrán que descargar en el puesto de trabajo el archivo mencionado en el enunciado, localizarlo y abrirlo con el programa adecuado (hoja de cálculo). Entonces tendrán que identificar la columna clave y ordenarla para encontrar los datos requeridos.
	17	SC5	(Interactuar puesto) Introducir el nombre del documento que contiene la palabra "Florencia" (localizar en una serie de documentos proporcionados).	Tendrán que descargar al puesto de trabajo el archivo mencionado en el enunciado, localizarlo y descomprimirlo. Una vez extraídos los documentos, tendrán que utilizar la casilla "Buscar" para buscar el término en todos los documentos del directorio.
	18	SC5	(Interactuar puesto) Introducir el nombre del país que tuvo la tercera tasa más alta en 2014 (localizar la información en una hoja de cálculo proporcionada).	Tendrán que descargar en el puesto de trabajo el archivo mencionado en el enunciado, localizarlo y abrirlo con el programa adecuado (hoja de cálculo). Tendrá que identificar la columna clave y ordenarla para encontrar los datos requeridos.
	73	SC5	(Web simulada) Solicitar la oferta de profesor en Sopela e introducir la referencia de la solicitud (Interactuar con la web de empleo simulada).	Tendrán que abrir la web mencionada en una nueva pestaña, para posteriormente interactuar con los filtros de búsqueda para encontrar la oferta pedida, y así poder registrarse y copiar el código de la aplicación.
Rec. (CD1)	5	SC7	(Interactuar simulación) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deberán abrir la simulación mencionada en una nueva pestaña, para interactuar con ella para encontrar las evidencias que le permitan responder al ítem.
	11	SC7	(Interactuar simulación) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deberán abrir la simulación mencionada en una nueva pestaña, para interactuar con ella para encontrar las evidencias que le permitan responder al ítem.
	26	SC7	(Interactuar simulación) De acuerdo con las situaciones mostradas en la imagen, seleccionar la afirmación correcta.	Deberá abrir la simulación en una nueva pestaña, interactuar con ella para encontrar evidencias que le permitan responder al ítem. ¿Prestaron especial atención a la marca de confianza para sitios web de comercio electrónico?

Tabla 5.6. Ítems seleccionados en cada estudio en TPL (En verde marcado los ítems seleccionados para el análisis en profundidad de los datos del ET) (*) En el campo SC coloreada la columna según el nivel inicialmente asignado: verde=básico, amarillo=intermedio, y rojo=avanzado.

Los ítems se mostraron únicamente en español y tenían que ser contestados dentro de ETCD, en un entorno controlado, sin tener que salir de la interfaz principal de la prueba (en la figura 5.1 se muestra un ejemplo de la interfaz de la prueba). Los ítems fueron presentados a los participantes en el mismo orden, y no se les permitía posponer la resolución o cambiar su orden. Las interacciones y los intentos de los participantes quedaban registrados en la plataforma y los resultados se calculaban automáticamente. Asimismo, se registraba el tiempo empleado en responder a cada pregunta, así como el tiempo total empleado por prueba.

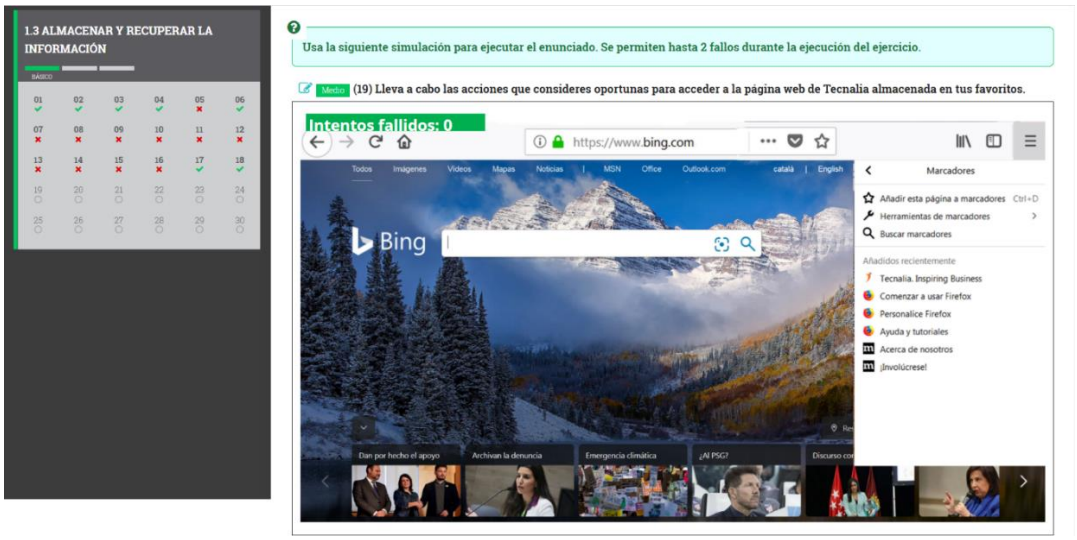


Figura 5.1. Ejemplo de la interfaz de una de las pruebas.

Tras recoger los resultados de los 30 participantes en los estudios, decidimos realizar un filtrado de los datos seleccionando los ítems más representativos para examinar las métricas del ET. Seleccionamos los ítems basados en una imagen/simulación, en los cuales los participantes tienen que examinar y evaluar una imagen/simulación, para posteriormente elegir la opción correcta de una lista de opciones posibles. Este formato de ítem puede ser muy interesante para evaluar las destrezas de orden cognitivo alto para los niveles de CD intermedio y avanzado, que requieren que los participantes evalúen críticamente las situaciones o apliquen sus conocimientos realizando determinadas tareas. Los ítems finalmente seleccionados y los criterios de evaluación definidos para cada ítem, se muestran en la tabla 5.7.

Item	SC	Criterios de evaluación
24	SC7	Examinar una página de identificación de una cuenta bancaria y observar que la URL es sospechosa y, por tanto, el sitio no es fiable. La URL es un elemento clave en la tarea.

32	SC8	Examinar las partes clave de una noticia publicada en un sitio web para aceptarla como fiable (URL, logotipo, autor, fecha, etc.).
4	SC1	Examinar los campos de un correo electrónico e identificar el campo incorrecto o hacer clic en el botón correcto. En este caso, el asunto es demasiado corto y sin suficiente información. Además, es necesario comprobar todos los campos para elegir el campo incorrecto.
21	SC1	Examinar los campos de un correo electrónico e identificar el campo incorrecto. El asunto tiene un error ortográfico, pero es necesario comprobar todos los campos.

Tabla 5.7. Ítems seleccionados y criterios de evaluación.

Los gráficos estaban siempre situados en la mitad derecha de la pantalla. Las instrucciones y el enunciado de la pregunta aparecían siempre en la parte superior derecha de la pantalla, mientras que la respuesta con respuestas alternativas aparecía en la parte inferior derecha. El botón "*responder*" para guardar los resultados y avanzar a la siguiente pregunta aparecía en la esquina inferior izquierda de cada elemento. Pretendimos minimizar los MO extraños con una disposición coherente de la pantalla.

5.2.3. Materiales

La información se recogió en un portátil un *MSI GS75 stealth i7* situado en una sala de reuniones de Tecnalia durante una semana de enero del 2020. Situamos el portátil lo suficientemente lejos para evitar distracciones causadas por las acciones realizadas por el investigador. Utilizamos el software *Tobii Pro Lab* (TPL) versión 1.130.24185 y el navegador integrado en el programa, para visualizar las pruebas en un monitor *DELL e2310* de 23 pulgadas conectado al portátil. El MO de los participantes se monitorizó con el rastreador ocular *Tobii X2-30*, que es un rastreador ocular autónomo y no intrusivo. El rastreador ocular capturó datos de los MOs a 30 Hz, y lo colocamos en la parte inferior de otro monitor de 17 pulgadas que usamos para seguir la posición de los ojos y la cabeza de los participantes. Además, también registramos la ubicación y el tiempo de los clics del ratón.

Diseñamos seis estudios diferentes en TPL para llevar a cabo el estudio (ver tabla 5.6). Cada proyecto comenzaba con una autoevaluación en la que los participantes indicaban el nivel de competencia que creían tener. Cada estímulo presentado en el TPL estaba vinculado a un ítem específico previamente seleccionado de las pruebas. Los 6 proyectos y sus estímulos se mostraron en el mismo orden. Utilizamos dos tipos de estímulos:

- Estímulo web, para mostrar páginas web a los participantes durante una grabación. El TPL abre la URL de la página web en el navegador incorporado, y este registra automáticamente todos los clics del ratón, las pulsaciones del teclado y las páginas web a las que se accede durante el estudio.

- Estímulo de grabación de pantalla, registra todos los clics del ratón, las pulsaciones del teclado, los programas y las páginas web a las que se accede durante el estudio.

En concreto, los ítems seleccionados para estudiar en profundidad se registraron utilizando un estímulo web. Para cada estímulo, definimos diferentes AOI y TOI. El AOI es un concepto utilizado en TPL que permite al investigador calcular medidas cuantitativas del MO. Para ello, trazamos un límite alrededor de todos los elementos del estímulo que consideramos de interés para nuestro estudio. Algunos AOI eran relevantes para los ítems, mientras que otras AOI no lo eran. Creamos AOI en todas las áreas de los ítems que creímos que podían atraer su atención. Las AOIs definidas para los ítems 24, 32 y 4 se muestran en las figuras 5.2, 5.3 y 5.4.

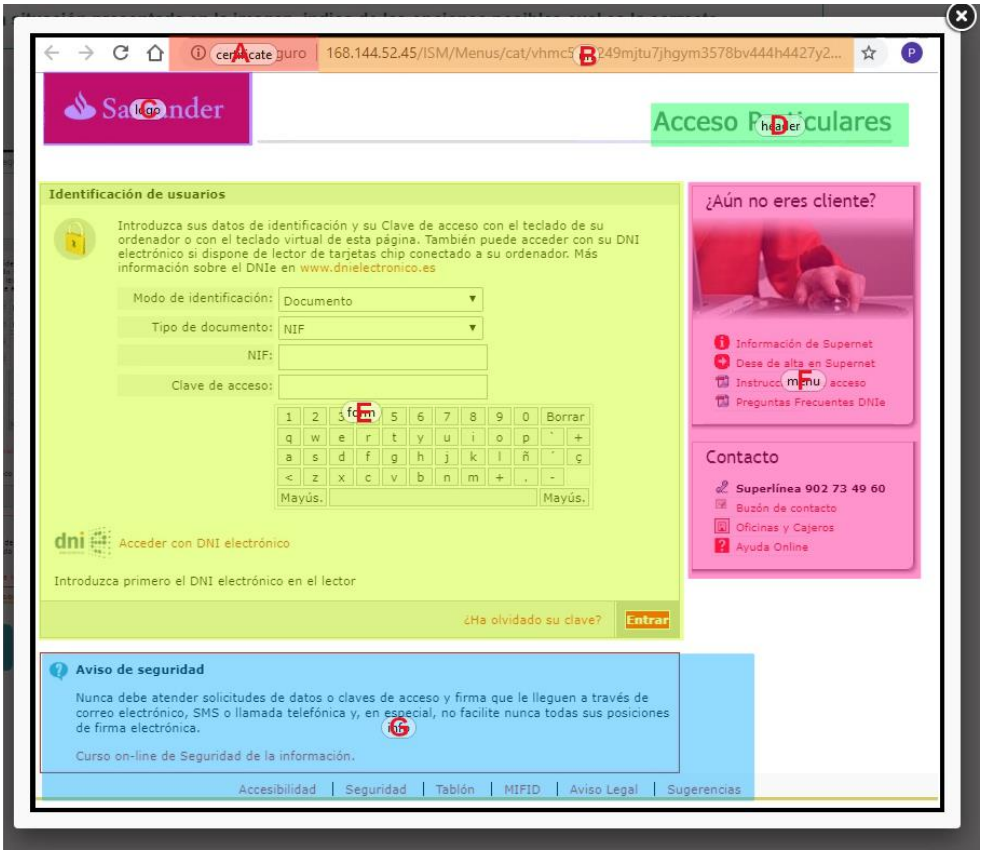


Figura 5.2. AOIs definidas para el ítem 24

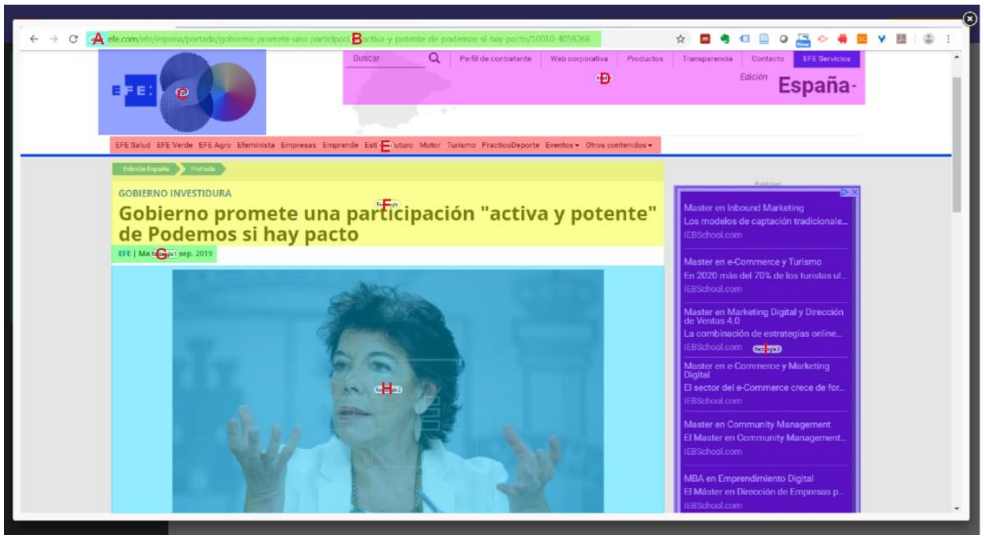


Figura 5.3. AOIs definidas para el ítem 32

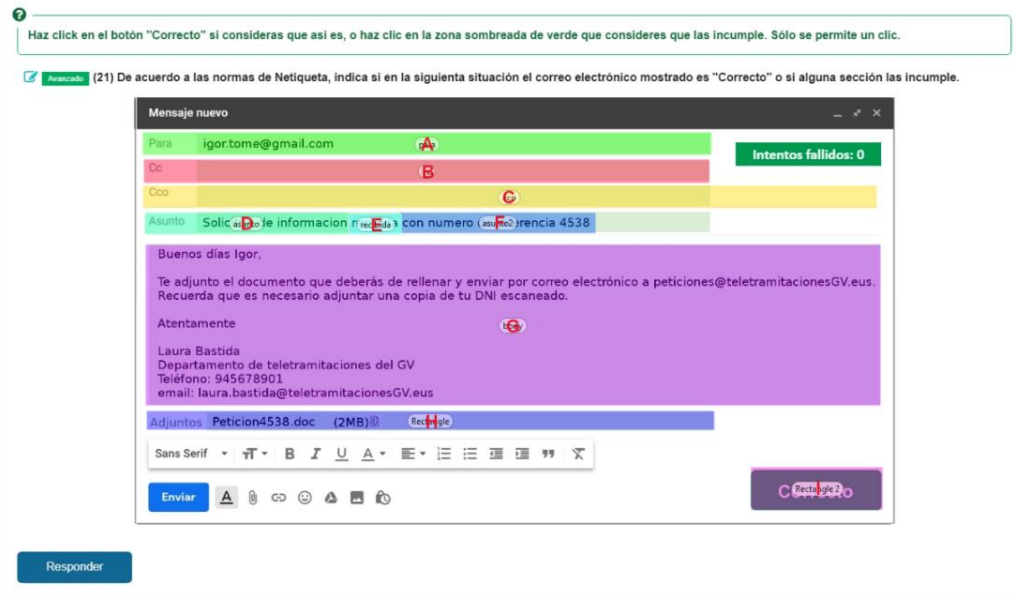


Figura 5.4. AOIs definidas para el ítem 4

El TOI es otro concepto de TPL que proporciona cierto grado de flexibilidad analítica, permitiendo a los investigadores organizar los datos de registro según los intervalos de tiempo durante los cuales se producen comportamientos y eventos significativos. En los ítems basados en imágenes o simulaciones, elegimos los intervalos en los que los participantes estaban examinando la imagen. A partir de la definición de las AOI y los TOI. TPL calcula las distintas métricas dentro del AOI durante el TOI.

5.2.4. Procedimiento

Para participar en el estudio, los participantes tuvieron que leer y firmar el formulario de consentimiento informado, que explicaba el objetivo y las condiciones del estudio. A continuación, los participantes tomaban asiento frente al monitor y calibrábamos individualmente el sistema en un proceso que duraba aproximadamente tres minutos. Una vez calibrado el sistema, fuimos mostrando los ítems a los participantes en el mismo orden. El rastreador ocular recogía datos de la mirada y flujos de clics, que se complementaban con la información registrada por el TEA. Para finalizar, seleccionamos el filtro I-VT (Fijación) para exportar los datos del TPL con el umbral de velocidad por defecto (30 grados/segundo).

5.3. Resultados

5.3.1. Análisis de los comportamientos mostrados por participantes con diferentes niveles de pericia

Empezamos analizando dos métricas de uso frecuente en la validación de pruebas: las puntuaciones y los tiempos de respuesta latentes. A pesar de su importancia, estas métricas sólo proporcionan una medida bruta del rendimiento de los participantes y no aportan ninguna información sobre las estrategias que adoptaron durante la resolución de los ítems. El tiempo medio empleado por ítem y prueba se tomó como indicador del compromiso de los participantes. Cada pregunta contó como un punto. En la tabla 5.8 se muestran la media de puntuación obtenida por estudio, así como el tiempo medio empleado. En la tabla 5.9 se muestra para cada ítem el tiempo medio empleado agrupado por participantes que respondieron bien a los ítems y los que respondieron mal.

Estudio	N.º Ítems	Puntuación media	Tiempo medio
1.1	10	6.95 (SD = 2.00)	650.32s. (SD = 174.45)
1.2	10	6.94 (SD = 1.73)	517.00s. (SD = 189.61)
1.3	10	7.65 (SD = 1.88)	524.21s. (SD = 127.60)
2.5	10	6.12 (SD = 2.08)	567.03s. (SD = 192.42)
1.1 (*)	4	3.41 (SD = 0.60)	539.07s. (SD = 105.50)
1.2 (*)	3	1.82 (SD = 0.95)	244.50s. (SD = 70.64)

Tabla 5.8. Puntuación media y tiempo medio empleado por estudio (*) tareas abiertas incluidas.

Ítem	Aciertos/Fallos	Tiempo medio (acierto)	Tiempo medio (fallo)
1	22/8	24.09s. (SD=12.86)	30.38s. (SD=6.46)
44	23/7	61.92s. (SD=29.64)	109.25s. (SD=52.23)
51	26/4	53.63s. (SD=19.82)	82.75s. (SD=30.32)
24	25/5	40.13s. (SD=17.50)	55.50s. (SD=19.50)
32	25/5	44.87s. (SD=22.41)	62.50s. (SD=31.50)
4	12/18	46.29s. (SD=16.82)	61.80s. (SD=20.96)
21	21/9	37.33s. (SD=20.13)	62.20s. (SD=33.54)
45	30/0	52.81s. (SD=16.16)	

Tabla 5.9. Comparación del tiempo medio empleado en cada ítem por participantes que lo respondieron correctamente y los que fallaron.

Cuando se examinan los resultados a nivel de ítem, el tiempo medio para completar los ítems seleccionados varía significativamente en función de si han tenido éxito o no. Además, el tiempo medio que se tarda en completar los ítems de las tareas abiertas es significativamente mayor, como cabía esperar. A continuación, mostramos los resultados del rastreador ocular agrupados por tipo de ítem.

Ítem de opción múltiple: Los participantes que lo respondieron correctamente necesitaron menos fijaciones y emplearon menos tiempo en el TOI que los que lo fallaron (ver tablas 5.10 y 5.11).

Media de fijaciones	Media de fijaciones en TOI	Media de duración del TOI
16.38s. (SD=7.95)	25.19 (SD= 12.36)	24.40s. (SD=9.32)

Tabla 5.10. Ítem 1 (CD1), datos de resultados y fijaciones (acertados n=22)

Media de fijaciones	Media de fijaciones en TOI	Media de duración del TOI
22.13s. (SD=9.80)	36.00 (SD=12.58)	31.92s. (SD=6.38)

Tabla 5.11. Ítem 1 (CD1), datos de resultados y fijaciones (fallidos n=8)

Ítems basados en una imagen o simulación: Al comparar las fijaciones de los participantes que respondieron correctamente al ítem con los que lo fallaron, se observan notables diferencias en el número de fijaciones, su densidad y su agrupación, tal y como se muestra en la tabla 5.12. Los participantes que

respondieron incorrectamente realizaron más fijaciones y dedicaron más tiempo a las fijaciones que los participantes que respondieron correctamente.

Ítem	Acierto / Fallo	Media de fijaciones	Media de fijaciones en TOI	Media de duración del TOI
24	Aciertos (n=24)	11.63 (SD=6.47)	20.54 (SD=8.78)	12.00s. (SD=5.91)
24	Fallos (n=6)	20.33 (SD=13.65)	28.17 (SD=15.59)	20.45s. (SD=11.95)
32	Aciertos (n=24)	23.30 (SD=11.73)	33.65 (SD=10.94)	17.94s. (SD=6.87)
32	Fallos (n=6)	45.80 (SD=26.27)	69.80 (SD=31.04)	39.87s. (SD=23.10)
4	Aciertos (n=12)	22.08 (SD=9.43)	61.75 (SD=32.34)	48.38s. (SD=28.39)
4	Fallos (n=18)	32.11 (SD=16.75)	88.06 (SD=42.08)	65.34s. (SD=26.83)
21	Aciertos (n=21)	22.48 (SD=15.27)	38.45 (SD=18.72)	32.99s. (SD=15.85)
21	Aciertos (n=9)	32.50 (SD=13.75)	45.33 (SD=21.43)	47.22s. (SD=13.39)

Tabla 5.12. Resultados y datos de fijaciones para los ítems basados en una imagen o simulación

Visualizando la información mediante *mapas de calor* nos ayudaron a comprender mejor el comportamiento de los participantes con diferentes niveles de CD y a validar los criterios de evaluación definidos para los ítems. Por ejemplo, en el ítem 21, el participante que respondió correctamente centró las fijaciones en el error ortográfico localizado ("*recivida*" en lugar de "*recibida*"), sabiendo que era un claro indicador de incumplimiento de las normas de netiqueta. En cambio, el participante que no acertó el ítem distribuyó las fijaciones por todos los campos, realizando más fijaciones y más densas (ver figura 5.5).

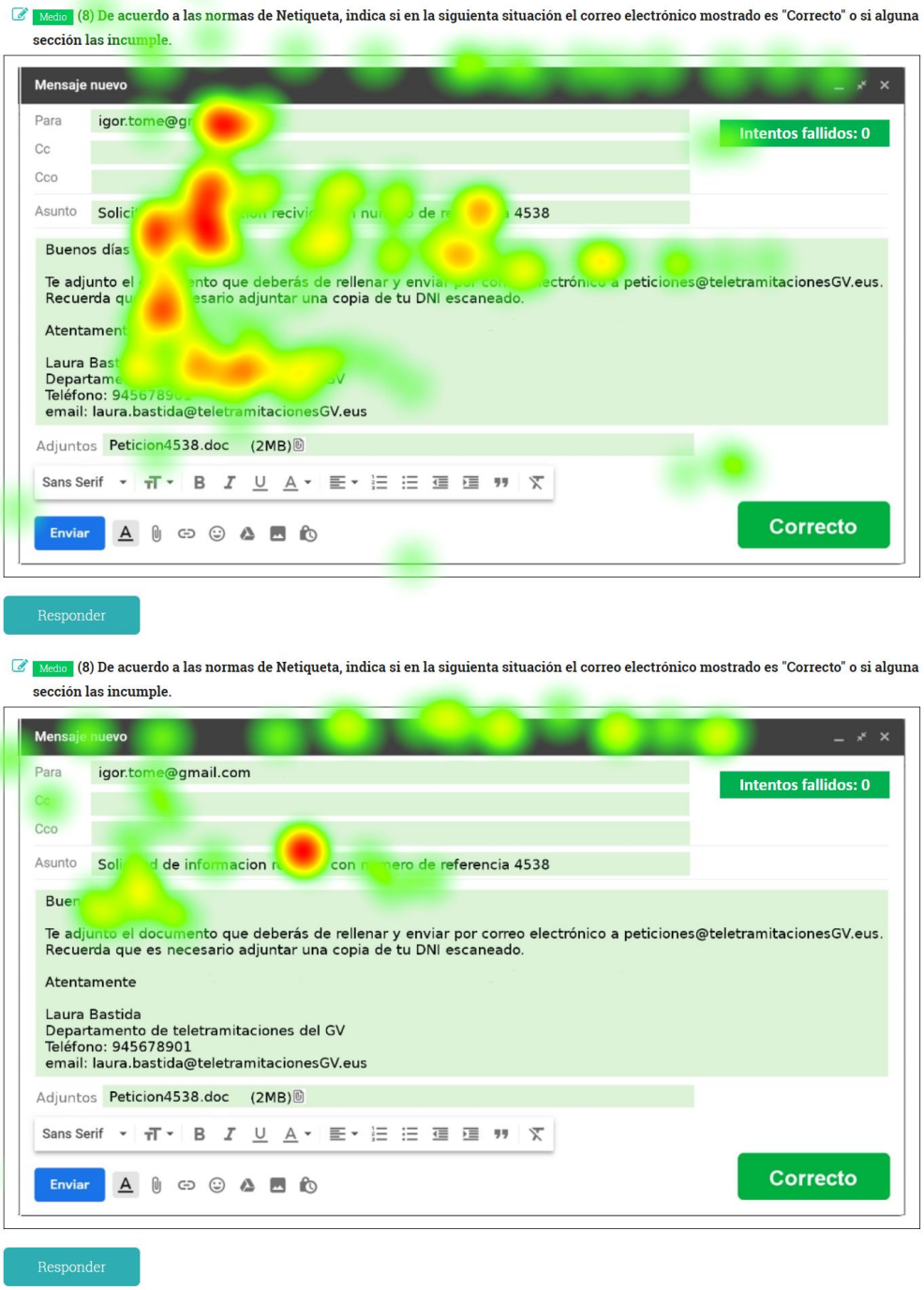


Figura 5.5. Ítem 21, en la parte superior un participante que falló el ítem y en la parte inferior un examinando que lo respondió correctamente.

En el ítem 24, el participante que lo acertó centró las fijaciones en el área de la URL, sabiendo que el certificado web era un componente clave y que la URL no era correcta. En cambio, el participante que falló distribuyó las fijaciones en diferentes áreas, realizando más fijaciones y más densas, pero sin prestar atención al área de la URL, mostrando desconocimiento en la materia (ver figura 5.6).

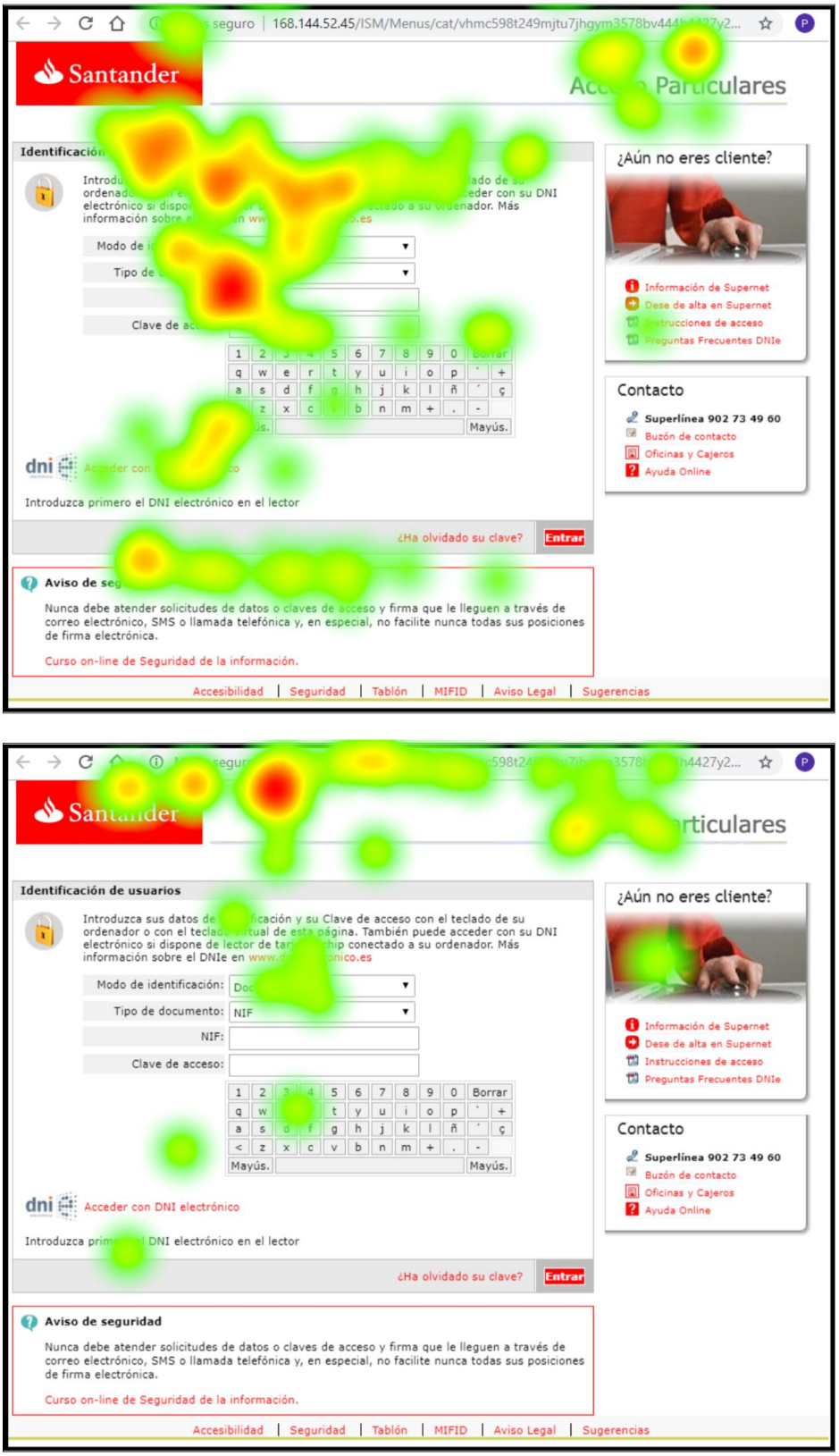


Figura 5.6. Ítem 24, en la parte superior un participante que falló y en la parte inferior un examinando que lo respondió correctamente.

Del mismo modo, los datos recogidos en el ítem 32 contribuyeron a confirmar que los participantes que lo acertaron, examinaron al menos las áreas clave necesarias para evaluar la fiabilidad de la fuente: *candado*, *URL*, *logotipo* y *rectángulo1*. Los que lo acertaron realizaron más fijaciones en las áreas clave con respecto al número total de fijaciones (56,79%) que los que fallaron (31,92%).

En el ítem 4, los resultados apuntaron en la misma dirección. El asunto era el área clave. Era demasiado corto y carecía de detalles, incumpliendo las normas de netiqueta. Los participantes que lo aceraron realizaron más fijaciones al área clave con respecto al número total de fijaciones (12,83%) que los que fallaron (6,05%).

Simulaciones interactivas: También hay diferencias notables en el número de fijaciones, su densidad y su agrupación entre las fijaciones de los participantes que la acertaron con los que la fallaron (ver tabla 5.13). Los participantes que la fallaron realizaron más fijaciones y dedicaron más tiempo a las fijaciones que los participantes que fallaron. En el ítem 44, también examinamos el tiempo consumido antes de que los participantes realizaran el primer clic, lo que reveló datos interesantes. La primera área clave para hacer clic fue la de "*Imágenes*" (necesaria para realizar una búsqueda basada en imágenes). Los participantes que la acertaron emplearon menos tiempo en hacer clic en esta zona (AV: 32,06 s.; SD: 9,88 s.) que los que la fallaron (AV: 66,53 s.; SD: 48,30 s.).

Ítem	Acierto / Fallo	Media de fijaciones	Media de fijaciones en TOI	Media de duración del TOI
44	Aciertos (n=23)	20.13 (SD=12.67)	73.35 (SD=31.56)	48.21s. (SD=20.93)
44	Fallos (n=7)	36.00 (SD=16.65)	118.57 (SD=52.09)	90.62s. (SD=38.57)
51	Aciertos (n=26)	15.58 (SD=8.85)	49.27 (SD=28.58)	40.65s. (SD=19.47)
51	Fallos (n=4)	24.75 (SD=8.85)	70.00 (SD=28.58)	77.94s. (SD=30.61)
45	Aciertos (n=30)	34.04 (SD=19.94)	56.13 (SD=33.38)	32.55s. (SD=11.56)
45	Fallos (n=0)			

Tabla 5.13. Resultados y datos de las fijaciones para los ítems basados en simulaciones

En el ítem 51, los datos recogidos contribuyeron a confirmar que los participantes habían acertado comprobando al menos las zonas clave: "*url*", "*headerstar*", "*Rectángulo*" y "*Rectángulo 1*", que son los lugares más comunes para analizar en la mayoría de los navegadores. Los encuestados que acertaron realizaron ligeramente más fijaciones a las áreas clave con respecto al número total de fijaciones (10,59%) que los que lo fallaron (10,10%).

En cuanto al ítem 45, la pregunta se diseñó esperando que los participantes examinaran el contenido de cada mensaje. Pero el comportamiento no fue el esperado, ya que los participantes revisaron secuencialmente todos los remitentes y asuntos de arriba a abajo y cuando llegaron al cuarto mensaje, lo marcaron como incorrecto sin acceder al contenido de los mensajes. El último mensaje fue claramente el menos examinado (98, 70, 83, 85 y 29 fijaciones respectivamente), probablemente porque los participantes ya percibieron la respuesta. Quizá la respuesta era demasiado obvia, lo cual hizo que todos lo respondieran bien.

Otro aspecto clave que queríamos analizar era si el número máximo de intentos definido para cada ítem era correcto obteniendo el nivel de dificultad esperado. Para ello, comprobamos los clics realizados en las simulaciones para identificar los pasos que requerían un análisis en profundidad. Por ejemplo, los dos primeros pasos del ítem 44 centralizaban el mayor número de clics erróneos y de abandonos (ver tabla 5.14). Este ítem tenía 6 clics erróneos como límite. Esperábamos que los usuarios exploraran las opciones siguiendo un criterio razonable. Sin embargo, después de analizar las miradas de los participantes que acertaron que hicieron más de 2 clics, detectamos que su comportamiento no mostraba un criterio razonable, pero el elevado número de intentos disponible les permitía descubrir la solución. Por ello, decidimos establecer 2 clics erróneos como límite. Con este cambio, habríamos obtenido 19 respuestas acertadas, con un nivel de dificultad más adecuado a nuestras necesidades. En el ítem 51, tres pasos centralizaron el mayor número de clics erróneos y de abandonos, y especialmente uno de ellos (ver tabla 5.15). Establecimos 4 clics erróneos como límite, y 26 participantes respondieron con éxito. Observando los mapas de calor, vimos que varios participantes tenían problemas para identificar la sección "*Catálogo*" como el lugar donde podrían estar los marcadores. Aceptamos estos resultados como correctos, pero también analizamos diferentes posibilidades para ajustar el nivel de dificultad. Fijando 2 clics erróneos como límite habríamos obtenido 16 aciertos.

Paso1	Paso2	Paso 3	Paso 4	Paso5	Paso6
17(*)	21(*)	8	0	4	0

Tabla 5.14. Ítem 44, pasos y errores (*) Incluidos 2 abandonos

Paso1	Paso2	Paso3	Paso final	Pasoa	Pasob	Pasoc
9	8	0	0	23 (*)	2	0

Tabla 5.15. Ítem 51, pasos y errores (*) Incluidos 2 abandonos

Resumiendo, los resultados mostrados revelan diferencias significativas en los RP entre los participantes que acertaron los ítems y los que lo fallaron, siendo los

primeros los que presentan tiempos medios de respuesta, cantidad de fijaciones en las AOI y duraciones del TOI más bajos. Encontramos diferencias notables en el número de fijaciones, su densidad y su agrupación. Los participantes que fallaron realizaron más fijaciones y dedicaron más tiempo en las fijaciones que los participantes que acertaron. Además, los participantes que acertaron realizaron más fijaciones en las áreas clave con respecto al número total de fijaciones que los que fallaron.

5.3.2. Evaluación de interpretaciones alternativas basadas en las AOI examinadas

Comenzamos examinando las longitudes de los recorridos de exploración de los participantes agrupados por resultado (acertadas o no), y la relación entre las puntuaciones y las variables seleccionadas. Examinamos si los participantes que acertaron tienen recorridos de exploración más largos en comparación con los que fallaron. Por ello, calculamos la longitud de los recorridos de exploración individuales de los participantes de ambos grupos para cada elemento (es decir, el número total de elementos visuales en sus recorridos de exploración, incluidas las repeticiones) (véase la tabla 5.16).

Item	Media de longitudes (aciertos)	Media de longitudes (fallos)
24	25.1 (21.4)	44.3 (30.6)
32	42.6 (24.7)	81.3 (42.9)
4	56.2 (22.8)	68.3 (40.9)

Tabla 5.16. Media (y desviación estándar) de las longitudes de las rutas de exploración para los ítems basados en una imagen o simulación.

Los participantes que fallaron miraron más puntos y, por lo tanto, tuvieron recorridos de exploración más largos en comparación con los participantes que acertaron. Las longitudes de los recorridos de exploración individuales pueden considerarse como un valor indicativo del grado de confianza del participante al realizar la tarea.

Además, analizamos la relación entre las puntuaciones, el tipo de ítem (si requería un enfoque sistemático para resolver el ítem o no), el tiempo para resolver el ítem, la longitud del recorrido de exploración y las AOI a comprobar del número total de AOIs definidos. Para ello, hemos calculado el coeficiente de Pearson (véase la tabla 5.17). Los resultados muestran que existe una correlación significativa entre las puntuaciones y el tipo de pregunta $r(87) = -0{,}389$, $p < 0{,}01$, el tiempo de respuesta $r(87) = -0{,}313$, $p < 0{,}01$, la longitud del recorrido de exploración $r(87) = -0{,}370$, $p <$

0,01 y el número de AOIs a comprobar del total de AOIs definidos y la longitud del recorrido de exploración r (87) = 0,941, p < 0,01. Es decir, el ítem que no requiere un enfoque sistemático, en el que hay que examinar todos las AOIs, es más difícil que los que sí lo requieren. Los participantes que fallaron también necesitaron recorridos de exploración y tiempos de respuesta más largos.

	Tipo de ítem	Tasa de AOIs	Tiempo de respuesta	Resultado	Longitud de la ruta
Tipo de ítem	1	,941**	,219*	-,389**	,317**
(Sig)		,000	,042	,000	,003
Tasa de AOIs	,941**	1	,170	-,378**	,211*
(Sig)	,000		,116	,000	,050
Tiempo de respuesta	,219*	,170	1	-,313**	,699**
(Sig)	,042	,116		,003	,000
Resultado	-,389**	-,378**	-,313**	1	-,370**
(Sig)	,000	,000	,003		,000
Longitud de la ruta	,317**	,211*	,699**	-,370**	1
(Sig)	,003	,050	,000	,000	

Tabla 5.17. Relación entre las puntuaciones con el tipo de ítem, el tiempo de respuesta, la longitud de la ruta de exploración y la tasa de AOIs. Correlaciones de Pearsons (* P < 0,05 (bilateral); ** P < 0,01 (bilateral); R (87))

Como parte del proceso de análisis, comprobamos cuáles de las AOI relacionados con cada ítem podían contener información crítica para la resolución satisfactoria del mismo. Para ello, combinamos la información de los diferentes participantes que visitaron cada una de las áreas para los diferentes ítems y sus resultados en ese ítem en concreto. En primer lugar, realizamos un estudio mínimo de la varianza de los datos para descartar los datos invariables que resultarían poco prácticos para el procedimiento analítico. La tabla 5.18 muestra la proporción de participantes que visitaron cada una de las AOI identificadas en los ítems objetivo. Los valores en negrita indican las áreas invariantes (visitadas por ninguno o por todos los participantes) que no ofrecen ninguna información en el proceso analítico. Un guion indica que el AOI no existe en esa tarea concreta.

Item	A	B	C	D	E	F	G	H	I
24	0,37	0,44	0,68	0,65	1,00	0,72	0,51	-	-
32	0,00	0,63	0,83	0,90	0,80	1,00	0,30	0,76	0,76
4	0,92	0,82	0,85	1,00	1,00	0,60	0,42	0,57	-

Tabla 5.18. Tasa de visitas de la AOI dentro de los ítems objetivo.

Una vez descartados los rasgos invariables, realizamos un análisis de correlación de rasgos entre cada uno de las AOI restantes y el resultado obtenido en el ítem. Teniendo en cuenta la naturaleza categórica tanto de las características como del objetivo, utilizamos la información mutua estimada para variables discretas como métrica de correlación. La figura 5.7 muestra los resultados para cada uno de los elementos del objetivo, señalando un fuerte desequilibrio en los valores de correlación, lo que conduce a la identificación de las AOI más críticos para cada elemento.

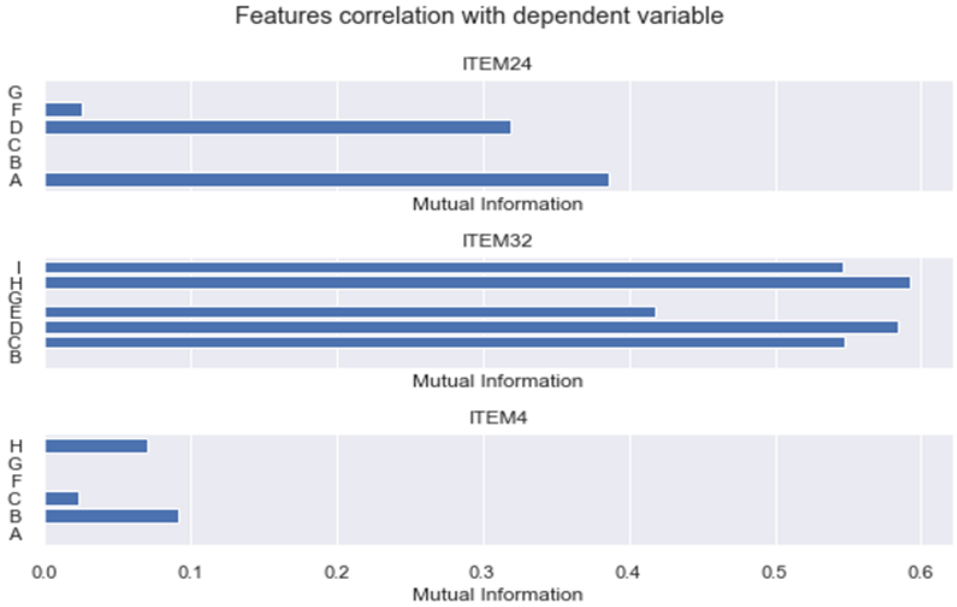


Figura 5.7. Correlación de características de las respuestas correctas para cada uno de los ítems.

Podemos descartar cualquier factor que tenga un impacto significativo en los resultados y que pueda socavar que los RP asociados no se basen en el mismo proceso cognitivo. En el ítem 24, se examinaron todos las AOIs con diferentes tasas de visita, pero no pudimos encontrar ningún AOI clave con una tasa de visita muy baja que hubiera planteado dudas. En el ítem 32, B, C, E y F eran críticos y todos estos AOIs tenían tasas de visita relevantes. Si hubiéramos encontrado un índice de visitas muy bajo para el AOI B, la situación habría sido muy difícil de explicar, ya que no era posible responder correctamente al ítem sin darse cuenta de que la URL no era válida. En el ítem 4, las AOI D y E eran críticos (la clave para resolver este ítem estaba ahí, en el campo del asunto del correo electrónico, incluyendo un error

ortográfico) y todos los participantes prestaron atención a estas áreas. En cambio, si hubiéramos visto que estas áreas apenas se habían tenido en cuenta en la resolución de la tarea, la cuestión habría suscitado dudas.

Además del análisis puramente analítico de los resultados, decidimos realizar una clasificación categórica del comportamiento de los participantes con un doble propósito. En primer lugar, este tipo de análisis permite identificar la flexibilidad de un ítem, proporcionando una visión directa de si el ítem está de alguna manera guiado, con la mayoría de los participantes siguiendo una única ruta de resolución, o si se trata de un ítem muy libre, con un patrón no evidente para resolverlo. En segundo lugar, es posible relacionar esos posibles patrones con los resultados obtenidos para comprobar si alguno de ellos es más eficaz que otros para resolver el ítem en cuestión.

Utilizamos una clasificación de clustering de k-means para identificar los diferentes patrones de comportamiento durante el estudio, utilizando las visitas de cada examinando a cada AOI como características de clasificación. Después de eliminar los datos invariables siguiendo los resultados de la tabla 5.18, empleamos el algoritmo OPTICS (Ankerst et al., 1999) para determinar un clustering adecuado respecto al comportamiento de los participantes. Este algoritmo ofrece dos ventajas significativas con respecto a otros algoritmos tradicionales, como el de las k-means. En primer lugar, es compatible con las distancias booleanas amigables, como la distancia de Hamming. Teniendo en cuenta que la naturaleza que todas las características utilizadas durante la clasificación son de naturaleza booleana, este es un factor clave para seleccionar este enfoque. En segundo lugar, a diferencia de k-means, este algoritmo detecta la cantidad óptima de clusters basándose en la propia distribución de los datos. El uso de este enfoque permite evitar la predefinición del número de clusters a descubrir durante el proceso. Las tablas 5.19 y 5.20 muestran los resultados de la evaluación del clustering.

Item	Elementos totales	Elementos clasificados	Elementos no clasificados	Número de clusters
24	29	15	14	5
32	30	16	14	5
4	28	23	5	6

Tabla 5.19. Descripción de los resultados de la agrupación para cada ítem.

Item	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5	Cluster #6
24	5	4	2	2	2	-
32	5	4	3	2	2	-
4	7	5	4	3	2	2

Tabla 5.20. Distribución de elementos a lo largo de los clusters identificados para cada ítem.

Los registros de los participantes muestran una gran variedad de comportamientos en sus respuestas. La tabla 5.19 muestra la existencia de varios elementos no clasificados, especialmente para los ítems 24 y 32. Además, la tabla 5.20 muestra que, incluso los elementos clasificados, se funden en clusters de pequeño tamaño muy fragmentados. Estas conclusiones son coherentes con la hipótesis de que el estudio no está guiado y, por tanto, los participantes eligen el camino de análisis que desean.

Para comprobar si alguno de los patrones de comportamiento identificados es más eficiente para resolver un ítem, asignamos a cada uno de los participantes al cluster correspondiente y, a continuación, realizamos una prueba ANOVA para comprobar si existen diferencias significativas entre los valores medios tanto del resultado (la respuesta correcta para el ítem específico) como del rendimiento global (la puntuación global considerando la tarea relacionada con la misma CD del ítem en cuestión). Sin embargo, como se muestra en la tabla 5.21, los resultados están lejos de ser significativos y, por tanto, no es posible concluir que ninguno de los patrones de comportamiento sea más eficiente que el resto.

Item	Puntación parcial	Puntuación total en la CD
Item24	0.250	0.753
Item32	0.499	0.758
Item4	0.661	0.593

Tabla 5.21. P-value de la prueba ANOVA considerando el resultado parcial y el rendimiento global

Aunque el análisis de agrupación muestra una capacidad significativa para predecir el resultado de cada ítem, la figura 5.7 muestra que varios AOI están mucho más correlacionados con los resultados de los ítems que otros. Para evaluar la capacidad de predicción de los ítems en función del comportamiento de los participantes, diseñamos un clasificador de Árbol de Decisión predictivo y lo entrenamos sobre 15 divisiones estratificadas aleatorias. La figura 5.8 muestra los resultados de este proceso, junto con la existencia de diferencias significativas entre la predictibilidad de los ítems. Como se ha señalado en el resultado anterior, el comportamiento por sí mismo no es suficiente para predecir los resultados de la interacción con cada

ítem, a pesar de que algunos ítems son más sensibles a los meros datos de comportamiento que otros.

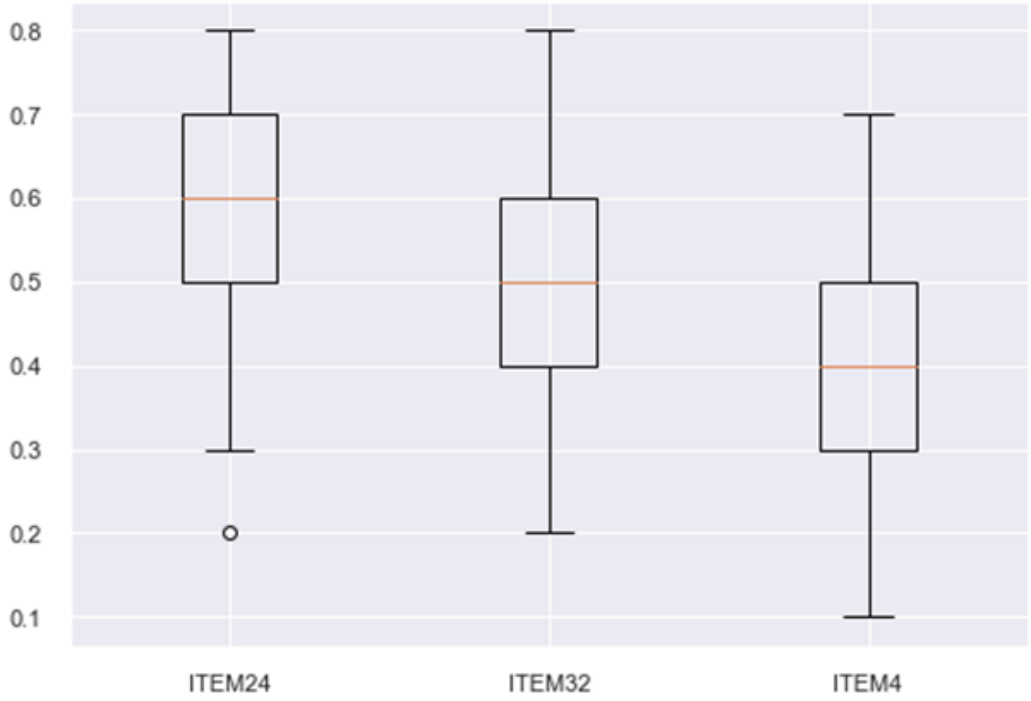


Figura 5.8. Puntuación del clasificador del Árbol de decisión a lo largo de 15 divisiones.

5.3.3. Evaluación de interpretaciones alternativas basadas en el orden de las AOI examinadas

Empezamos analizando la varianza dentro de los grupos para llegar a algunas conclusiones relacionadas con el tipo de resolución que requiere cada ítem (sistemáticamente o no). Utilizamos el algoritmo *String-edit*, que calcula la distancia entre dos rutas de exploración transformando una ruta de exploración en otra con el mínimo número de operaciones de edición (sustitución, eliminación y adición). La similitud entre las dos rutas de exploración puede calcularse como un porcentaje basado en la distancia *String-edit*. Para dos rutas de exploración idénticas no hay varianza, y para dos rutas de exploración completamente diferentes la varianza está en el nivel máximo. Por ejemplo, el algoritmo String-edit calcula la distancia entre ABCD y ABCA como uno, ya que basta una sola operación de sustitución entre D y A para transformar uno de ellos en el otro.

Como se muestra en la tabla 5.22, la media de los cambios necesarios basados en el método de Levenshtein son menores para el grupo que respondió correctamente que para el grupo que falló. Es decir, la similitud media dentro del grupo que acertó es mayor que la del grupo de falló. Luego, podemos sugerir que la varianza es mayor en el grupo que falló en comparación con el grupo que acertó.

Item	Aciertos	Fallos
24	9.5 (4.8) (n=24)	10.3 (3.8) (n=6)
32	15.2 (7.7) (n=24)	30.3 (13.3) (n=6)
4	16.9 (7.1) (n=12)	18.6 (8.9) (n=18)

Tabla 5.22. Media (y desviación estándar) de los cambios necesarios según el método Levenshtein para los ítems basados en una imagen o simulación.

Para ilustrar claramente la alta varianza dentro del grupo de participantes sin éxito, también utilizamos la herramienta *ScanGraph*[55], la cual está disponible públicamente y puede utilizarse para generar un gráfico visual basado en la similitud *String-edit* en el que pueden observarse como las rutas de exploración similares están conectadas entre sí, mostrándose como agrupaciones en este gráfico. Las figuras 5.9, 5.10 y 5.11 ilustran el gráfico aconsejado (un gráfico con el 5% de las aristas posibles) para los participantes del grupo que respondieron correctamente y los participantes del grupo que fallaron para los diferentes ítems seleccionados.

Para los ítems 24 y 32, este gráfico muestra cómo las rutas de exploración del grupo que acertaron están más conectadas entre sí, ya que son más similares entre sí en comparación con el grupo que no acertaron. En el caso del ítem 4, las rutas de exploración del grupo que no acertaron están más conectadas. Esto puede deberse a las características del ítem, ya que era necesario examinar todos los campos del correo electrónico. Los participantes que respondieron correctamente el ítem tuvieron rutas de exploración más cortas, lo que puede entenderse como que identificaron antes el campo erróneo y centraron sus observaciones, mientras que los que lo fallaron tuvieron una revisión más larga de todos los campos, realizando una estrategia más similar. En los ítems 24 y 32, los participantes tuvieron que examinar la fiabilidad de la página de identificación de un banco y de una noticia publicada en una página web, respectivamente. En ambos ítems se requería comprobar sólo algunas partes clave de forma sistemática. Los participantes no tuvieron que examinar todos los elementos posibles, lo que dio lugar a mayores similitudes en las rutas de exploración de los participantes que respondieron correctamente a estos ítems.

[55] http://eyetracking.upol.cz/scangraph/

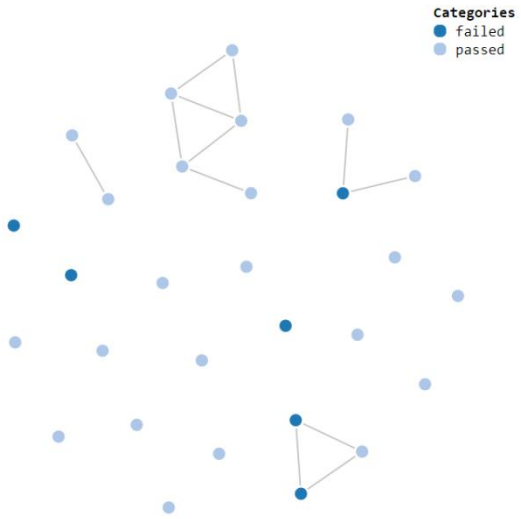


Figura 5.9. Ítem 24, gráficos de similitud de rutas de exploración entre los participantes que acertaron y los que fallaron, elaborados con la herramienta ScanGraph.

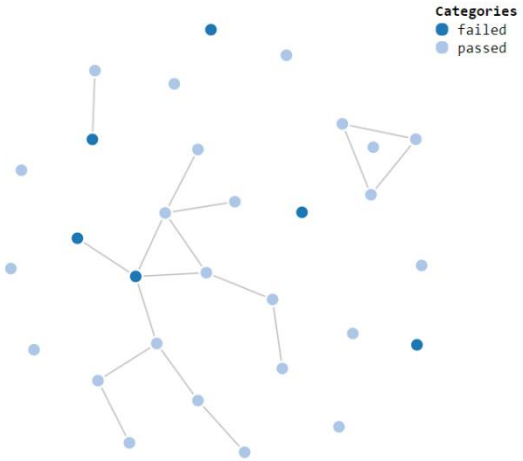


Figura 5.10. Ítem 32, gráficos de similitud de rutas de exploración entre los participantes que acertaron y los que fallaron, elaborados con la herramienta ScanGraph.

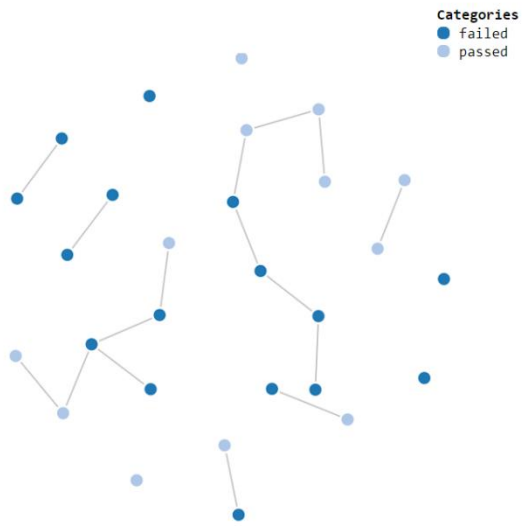


Figura 5.11. Ítem 4, gráficos de similitud de rutas de exploración entre los participantes que acertaron y los que fallaron, elaborados con la herramienta ScanGraph.

También examinamos las rutas de exploración de los participantes que acertaron con el objetivo de identificar una ruta de exploración que representara a todo el grupo, lo que se conoce como la *ruta de exploración común*. Queríamos saber cómo se comportaban los participantes que acertaron a resolver estos ítems, es decir, en qué AOI realizaban las primeras fijaciones y el orden de las fijaciones posteriores. Esta representación visual puede ser muy útil para validar de forma sencilla que la ruta de exploración común no representa una alternativa a la interpretación de los resultados que socavaría los criterios de evaluación definidos para un ítem concreto. A continuación, aplicamos un modelo ponderado basado en la posición, tal y como Holsanova et al. (2006) aplicaron para analizar las trayectorias de lectura y las prioridades de lectura en las páginas de los periódicos.

En primer lugar, dividimos las imágenes presentadas en el estímulo en sus AOI y, a continuación, las clasificamos en función de cuáles fueron las primeras AOI visitadas por los participantes. Por ejemplo, en el ítem 4 definimos seis AOI asociados a los diferentes elementos visuales. Por lo tanto, cuando se aplica el modelo ponderado basado en la posición a las rutas de exploración dando 6 puntos a los elementos visuales visitados por primera vez y ningún punto a los elementos visuales no visitados, la secuencia de los elementos visuales para todas las rutas de exploración se identifica como sigue: DABCHG (véanse las tablas 5.23, 5.24 y 5.25). Antes de aplicar el modelo, decidimos simplificar las AOI originales uniendo algunos AOI muy cercanos y estrechamente relacionados. En el ítem4 unimos DEF en un único AOI D que representa todo el campo del asunto del correo electrónico. En el ítem 24 unimos AB en un único AOI A que representa toda la sección URL del navegador. Y finalmente, en el ítem 32 unimos AB en un único AOI A que representa toda la sección URL del navegador, y FG en un único AOI F que

representa la sección de titulares de la noticia. Aunque el mismo AOI podía ser visitado varias veces por los participantes, las repeticiones no se tuvieron en cuenta en este enfoque.

Ruta de Exploración	A	B	C	D	G	H
DABCHG	52	43	40	55	12	29

Tabla 5.23. Ítem 4, resultados de la aplicación del modelo ponderado basado en la posición.

Ruta de Exploración	A	C	D	E	F	G
ECGDA	55	67	63	106	0	66

Tabla 5.24. Ítem 24, resultados de la aplicación del modelo ponderado basado en la posición.

Ruta de Exploración	A	C	D	E	F	H	I
FCDAEHI	83	99	99	74	141	63	56

Tabla 5.25. Ítem 32, resultados de la aplicación del modelo ponderado basado en la posición.

En las figuras 5.12, 5.13 y 5.14 se muestran las *rutas de exploración comunes* identificadas para cada ítem.

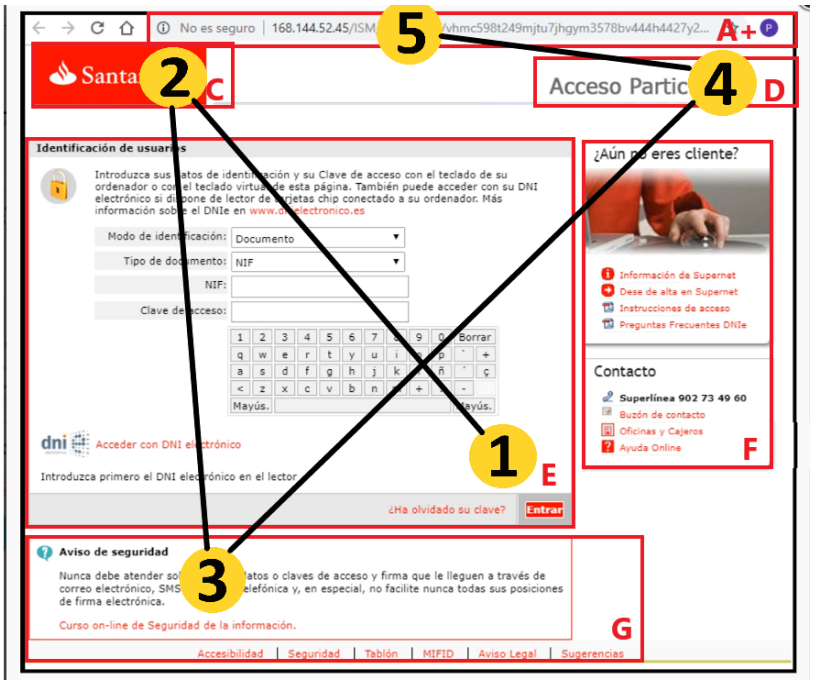


Figura 5.12. Ruta de exploración común identificada para el ítem 24.

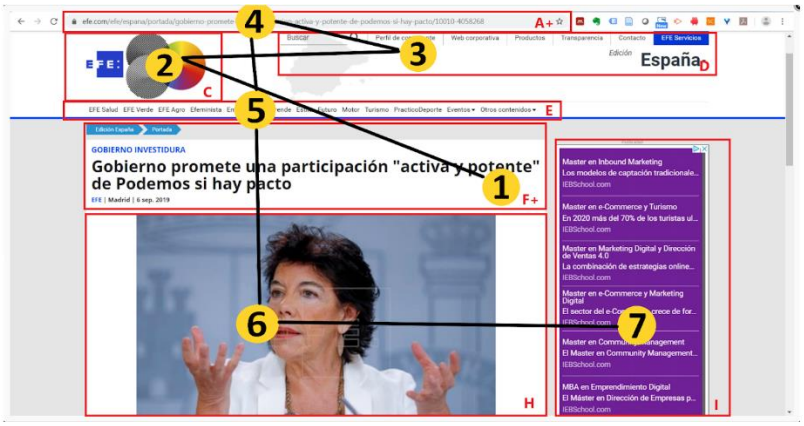


Figura 5.13. Ruta de exploración común identificada para el ítem 32.

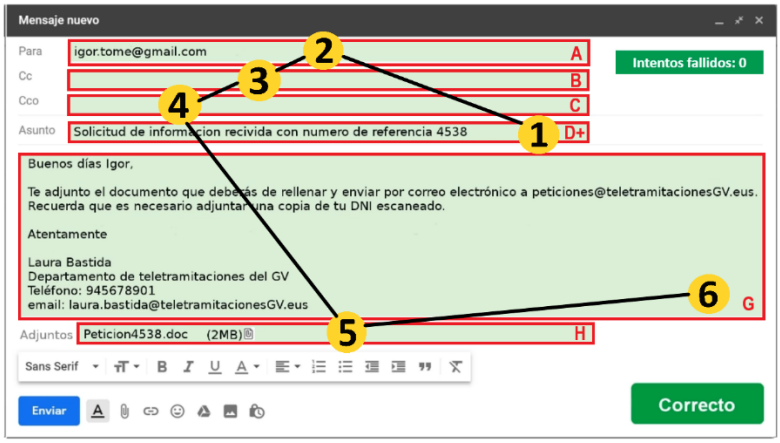


Figura 5.14. Ruta de exploración común identificada para el ítem 4.

En el ítem 4 los participantes hicieron su primera fijación en el campo del asunto del correo electrónico. Parece que este AOI fue clave para ellos para contextualizar el correo electrónico. Además, este es el campo que contiene la falta de ortografía, factor clave para no cumplir las normas de netiqueta. A continuación, los participantes continúan revisando el resto de los campos adoptando un enfoque descendente, excepto el AOI relacionado con el contenido del mensaje, que se dejó para el último lugar. Es notable ver cómo los participantes dejaron este AOI para el final, dando poca importancia al contenido del mensaje a la hora de aplicar las reglas de netiqueta.

En el ítem 24 los participantes realizaron su primera fijación en la zona principal de la página web donde se muestra un formulario de identificación del usuario. A continuación, los participantes siguieron examinando el logotipo de la página web, el AOI que muestra una advertencia de seguridad y la parte superior derecha de la cabecera que indica que la página web corresponde con un acceso privado a una cuenta bancaria. Parece que estos AOIs fueron claves para que contextualizar el

sitio web. A continuación, los participantes siguieron examinando la AOI relacionada con la URL del sitio web. Esta AOI es clave para resolver con éxito el ítem y los participantes deben comprobar si el sitio web tiene un certificado web válido y el dominio es válido de acuerdo con el banco accedido.

En el ítem 32 los participantes realizaron su primera fijación en la cabecera de la noticia. A continuación, siguieron examinando el logotipo del sitio web y la parte superior derecha de la cabecera, incluido el menú. Parece que estos AOI fueron clave para que contextualizar la página web. El comportamiento fue bastante similar al mostrado por los participantes en el ítem24. A continuación, siguieron examinando el AOI relacionado con la URL del sitio web para comprobar si el sitio web tiene un certificado web válido, y si el dominio es válido según el periódico accedido. Por último, los participantes continuaron examinando el resto de los campos adoptando un enfoque descendente.

A pesar de las limitaciones indicadas por Eraslan et al. (2016), decidimos aplicar este modelo para obtener los primeros datos sobre el rendimiento de los participantes que respondieron correctamente debido a la simplicidad del método y a la utilidad de los resultados obtenidos. Se trata de una forma muy sencilla de poder comprobar que los participantes que respondieron a la pregunta correctamente no mostraron un comportamiento alternativo que invalidara los criterios de evaluación del ítem. En futuros estudios se podría utilizar otro modelo, p. ej., considerando las duraciones de las fijaciones y las posiciones de los elementos en los estímulos visuales (Eraslan et al., 2016).

Por último, analizamos la duración de las fijaciones como un estimador de la dificultad de los ítems. Según las duraciones de las fijaciones y los lugares donde se produjeron las fijaciones más largas que se muestran en la tabla 5.26, el ítem 32 fue el menos difícil y el ítem 24 fue el más difícil. Todos los ítems fueron seleccionados inicialmente como niveles avanzados, pero tras examinar las duraciones de fijación de cada ítem, podemos confirmar que el nivel de dificultad y el esfuerzo requerido para resolver las tareas, no son los mismos.

Item	Duración de las fijaciones
24	265.7 (195)
32	234.3 (155.7)
4	246.4 (177.1)

Tabla 5.26. Media (y desviación estándar) de las duraciones de fijación de los ítems basados en una imagen o simulación

También examinamos qué lugares son los que atraen más fijaciones para comprobar si el comportamiento mostrado por los participantes es el esperado según los criterios de evaluación definidos para cada ítem (véase la tabla 5.27).

AOI	Item 24	Item 32	Item 4
A	345.4 (288.9)	292.0 (0.0)	263.0 (176.5)
B	295.9 (210.9)	282.7 (184.8)	292.0 (218.6)
C	233.0 (125.0)	264.3 (181.7)	274.0 (218.7)
D	203.3 (99.8)	229.0 (133.9)	278.3 (256.0)
E	272.5 (208.7)	258.6 (203.8)	232.5 (156.6)
F	230.6 (130.0)	211.0 (141.4)	212.0 (122.7)
G	264.1 (173.9)	328.6 (158.6)	246.6 (96.8)
H	----------	265.6 (172.5)	287.3 (183.1)
I	----------	209.0 (109.6)	----------

Tabla 5.27. Media (y desviación estándar) de las duraciones de fijación para cada AOI.

En el ítem 24, los participantes realizaron las fijaciones más largas en las AOI correspondientes al certificado web y a la URL del sitio. Las AOI correspondientes al formulario de entrada y al área con información de seguridad también recibieron fijaciones largas. En el ítem 32, los participantes realizaron las fijaciones más largas en el AOI relacionado con el autor y la fecha de publicación. Las AOI correspondientes al certificado web y a la URL del sitio también recibieron fijaciones largas. En el ítem 4, las medias de las duraciones de fijación fueron más similares. Esto se debe a que los participantes necesitaban examinar todos las AOI para llegar a una conclusión. Las AOI correspondientes a los campos CC y adjunto recibieron fijaciones ligeramente mayores y las AOI correspondientes al asunto donde se muestra el número de referencia y el campo con el contenido del correo electrónico, recibieron fijaciones ligeramente más breves. Para resolver los ítems 24 e ítem 32, los participantes siguieron un enfoque sistemático. Sin embargo, el ítem 4 requirió un enfoque diferente en el que los participantes tuvieron que examinar todas las AOI. La distribución de las fijaciones y sus duraciones contribuyeron para confirmar que los criterios de evaluación definidos para los ítems estaban bien establecidos y que los participantes habían mostrado el comportamiento esperado. Es decir, no pudimos identificar un comportamiento alternativo que hubiese puesto en entredicho los criterios de evaluación establecidos.

5.4. Conclusiones y discusión

Los resultados presentados en este capítulo ayudan a entender el comportamiento de los participantes con diferentes niveles de CD durante la realización de las pruebas. En concreto, utilizamos las observaciones del rastreador ocular para llenar el "vacío explicativo" al proporcionar datos sobre la variación de los RP de los participantes al resolver los ítems que no se recogen en los TEA tradicionales.

Por ello, nos centramos en evaluar una interpretación alternativa de las puntuaciones en las pruebas de acuerdo con la sugerencia de Cronbach (1980). Con este objetivo, buscamos invalidar la hipótesis de que los ítems desencadenaban los conocimientos y habilidades esperados para resolver los ítems seleccionados, poniendo a prueba una interpretación alternativa: que los participantes no prestan atención a las áreas clave (PI_K3) o siguen un orden de fijaciones sin sentido para responder correctamente a los ítems (PI_K4). Es decir, que resuelven las tareas de forma incoherente con lo esperado según los criterios de evaluación previamente establecidos para cada ítem. En concreto, respecto a nuestros objetivos de investigación:

- PI_K3. *¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos generar y probar inferencias sobre el constructo de interés?*

  Para evaluar esta interpretación alternativa de las puntuaciones en las pruebas, examinamos las rutas de exploración de los ítems basados en una imagen o simulación. Investigamos cómo los participantes procesaban las diferentes AOI de las imágenes incluidas en los ítems, para evaluar la situación y elegir la respuesta correcta.

  Intentamos responder a dos preguntas. En primer lugar, qué AOIs dentro de los ítems fueron examinadas e investigar si los patrones específicos de AOIs visitados podrían poner en evidencia que los ítems requieran el mismo proceso cognitivo que se requiere en las tareas del mundo real según los criterios de evaluación previamente definidos. Analizando los datos, no pudimos identificar patrones alternativos en términos de tasas de visitas de las AOI que pudieran invalidar los criterios de evaluación establecidos. Todos las AOI fueron participantes con diferentes tasas de visita, pero no pudimos encontrar ningún patrón de respuesta con una tasa de visita inesperada y difícil de explicar que hubiera planteado dudas sobre si el ítem estaba generando los RP esperados. En segundo lugar, llevamos a cabo una agrupación de las respuestas en términos de AOI visitadas, aplicando un algoritmo de clasificación k-means no supervisado para examinar si patrones específicos de AOI visitadas suponían mayores tasas de éxito en el rendimiento global (la puntuación global considerando la tarea relacionada con la misma CD de la tarea en cuestión) y de

manera inconsistente con las expectativas para cada ítem. Pudimos identificar clusters con mayores tasas de éxito, sin embargo, los resultados estaban lejos de ser significativos y, por lo tanto, no fue posible concluir que alguno de los patrones de comportamiento fuera más eficiente que el resto. Sería muy interesante volver a realizar este análisis con un mayor número de participantes.

- PI_K4. *¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos utilizar esta información para mejorar los diseños de los ítems?*

  Examinamos las rutas de exploración de los distintos elementos con mayor profundidad considerando el orden de procesamiento de los diferentes AOI. Como valor indicativo del grado de confianza del examinando al responder el ítem, encontramos que los participantes que fallaban tenían rutas de exploración más largas que los que acertaban. Además, utilizamos el método de *Levenshtein* con el apoyo de la herramienta *ScanGraph* para descubrir que la varianza es mayor en el grupo de participantes que fallaron. Así mismo, calculamos las rutas de exploración comunes de los participantes que acertaron, para profundizar en las estrategias de resolución que habían seguido.

  A pesar de las limitaciones del modelo que se aplicó, obtuvimos interesantes conocimientos sobre el rendimiento de los participantes que respondieron correctamente a los ítems que no podrían haberse obtenido en un TEA tradicional. Como resultado, obtuvimos una representación visual que nos permite comprobar de forma sencilla que los participantes que respondieron correctamente al ítem no mostraron un comportamiento imprevisto que invalidara los criterios de evaluación del ítem.

  Además, esta información nos parece útil para los participantes que no han respondido correctamente este tipo de preguntas. Si quisieran revisar el ítem para conocer la respuesta correcta, podríamos mostrarles la opción correcta y la imagen con la ruta de exploración común, como se muestra en las figuras 5.12, 5.13 y 5.14, añadiendo un valor extra al proceso de revisión. Sin embargo, en el metaanálisis llevado a cabo por Xie et al. (2021), mostró que el uso de ejemplos de modelado de MOs era beneficioso para el rendimiento de los alumnos cuando se utilizaban tareas no procedimentales en lugar de tareas procedimentales. Por lo tanto, habría que seguir investigando hasta qué punto el uso de ejemplos de modelado en el proceso de revisión de una prueba de CD podría ser apropiado dependiendo del tipo de pregunta.

  Por último, también examinamos la dificultad de los ítems analizando las duraciones de las fijaciones y los lugares donde se producían fijaciones más largas. Pudimos confirmar que los participantes que respondieron

correctamente realizaron las fijaciones más largas en las zonas esperadas, y no pudimos identificar una interpretación alternativa que socavara los criterios de evaluación definidos. Además, basándonos en la distribución de las duraciones de las fijaciones, pudimos validar el diseño de los ítems identificando dos enfoques diferentes, sistemático y no sistemático. Mientras que en el enfoque sistemático sólo algunas AOI eran más fijadas, en el enfoque no sistemático la distribución de las duraciones de las fijaciones era homogénea, lo que sugería que todas las AOIs se examinaban de forma similar.

Revisando la literatura científica, hemos podido ver como en estudios anteriores en campos relacionados con las CD seleccionadas ya se utilizaron datos provenientes de rastreadores oculares. Por ejemplo, los patrones de fijación son relevantes en el análisis y reconocimiento rápido de la información (Ashraf et al., 2018; Brunyé et al., 2019; Manning et al., 2006) y son útiles para entender el comportamiento de los usuarios en los motores de búsqueda (Lewandowski y Kammerer, 2021), o para mostrar las diferencias en las estrategias para responder a las preguntas de opción múltiple en las tareas de comprensión y recuerdo de códigos (Sharafi et al., 2012). Así mismo, sólo Yaneva et al. (2021) utilizaron los datos de los rastreadores oculares con fines de validación, para examinar cómo las distribuciones de las opciones en los ítems de opción múltiple influían en la forma en que los participantes respondían a los ítems.

Sin embargo, hasta donde sabemos, ningún estudio anterior se ha centrado en la evaluación de la CD, y menos aún, con el objetivo de validar los RP. Hemos mostrado diferentes formas de utilizar los datos recogidos sobre las áreas de fijación y la densidad que podrían utilizarse para respaldar la validación y el desarrollo de pruebas de evaluación de CD. Además, también hemos mostrado como el análisis de los datos de los rastreadores oculares puede formar parte del argumento de validez, específicamente para formatos de ítems como el elegido en el estudio, en el que los datos de la mirada son la única fuente disponible para describir la estrategia de los participantes y servir de ayuda para ajustar mejor los ítems y su nivel de dificultad. De hecho, tal y como Engelhardt et al. (2017) afirmaron, los cambios en las características del ítem pueden influir en su dificultad sin modificar el constructo representado.

Por otra parte, a la hora de interpretar los resultados del presente estudio, hay que tener en cuenta algunas limitaciones, como la muestra relativamente pequeña de participantes en el estudio. Sería interesante ver si estos resultados se replican con una muestra más amplia que incluyera también a personas con un nivel muy básico de CD.

Además, sería recomendable realizar más investigaciones que profundicen sobre la eficacia de mostrar en las revisiones de las pruebas de evaluación la *ruta de*

exploración común de los participantes que respondieron correctamente, teniendo en cuenta que la varianza en cuanto a la ruta de exploración seguida es mayor en el grupo que no acertó en comparación con el grupo que sí lo ha tenido.

Por último, en futuros estudios sería interesante considerar ítems con diferentes formatos de respuesta, que podrían implicar RP más complejos como, p. ej., haciendo uso de juegos. Pensamos que los resultados presentados son útiles y pueden suponer una interesante contribución al proceso de construcción de un argumento de validez en la evaluación de CD. Pero cabe destacar que son limitados debido a que los estudios usando rastreadores oculares consumen mucho tiempo y, por esta razón, las muestras de participantes en los estudios suelen ser limitadas. De igual forma, hay más métricas disponibles para aplicar a los datos de la mirada que podrían investigarse y que podrían contribuir a una mayor comprensión de los comportamientos de los participantes en la resolución de los ítems. Los resultados de esta investigación podrían tenerse en cuenta a la hora de pensar en cómo utilizar los datos de un rastreador ocular para validar el diseño de los ítems cuando se miden constructos cognitivos complejos como el de la CD, donde es necesario utilizar diferentes formatos de ítems.

Lo bueno si breve, dos veces bueno.

Baltasar Gracián

6.

Test Adaptativo Informatizado evaluatucompetenciadigital.com

Tras revisar y comentar el creciente número y variedad de herramientas existentes para evaluar las CD y cómo se ha implementado la evaluación de distintas formas, consideramos que era importante realizar un estudio de viabilidad de una implementación adaptativa de la herramienta ETCD, destacando los principales puntos a considerar. De todas las herramientas existentes de evaluación de CD, sólo la desarrollada por Vie et al. (2017) estaba basada en un TAI de DC y la información disponible es escasa.

El objetivo de este estudio busca evaluar los puntos clave para convertir ETCD en un TAI, en la que los ítems se seleccionan gradualmente para el participante, de acuerdo con su competencia. El valor de la competencia de cada participante se actualiza después de cada respuesta, basándose en el modelo de la TRI (Aybek y Demirtasli, 2017). En este caso, para un participante determinado los ítems muy fáciles o difíciles no se le presentarán, lo que reduce la duración y el tiempo de ejecución de la prueba, mejorando su eficiencia y objetividad, y produciendo un resultado inmediato (Weiss, 1982). El TAI de CD proporcionará claras ventajas psicométricas y económicas sobre las formas clásicas de test y será una alternativa atractiva y beneficiosa especialmente en un contexto de certificación, donde su uso previsto sería evaluar a un gran número de personas en muy poco tiempo.

Una de las principales ventajas de las TAI es que pueden reducir considerablemente la duración de la prueba sin afectar demasiado a la calidad de las capacidades estimadas. Para justificar la elección de un TAI, el objetivo de este estudio es comparar las pruebas lineales para ambas pruebas (de clasificación) y sus correspondientes TAIs, haciendo uso del mismo conjunto de datos simulados. Más

concretamente, se compararán las clasificaciones de los participantes con respecto a un umbral especificado de "aprobado-no aprobado", tanto con el banco de ítems 1PL completo de la prueba lineal como en el diseño TAI.

El uso de simulaciones, como las que hemos realizado en este estudio haciendo uso del paquete R "catR" (Magis y Barrada, 2017), permite la compilación inmediata de resultados de las pruebas, no genera inconvenientes en los participantes y reduce el tiempo y esfuerzo que conllevaría llevar a cabo las pruebas con usuarios reales (p. ej., Barnard, 2018; Magis y Raîche, 2012). Más aún, permite abarcar una mayor población y más variada, dado que el usuario final al que va destinado es muy amplio y variado. De hecho, cuando la población no está distribuida en toda la amplitud del colectivo que la prueba pretende medir, los parámetros estimados en las partes del colectivo con muy pocos participantes en el estudio de calibración podrían desviarse considerablemente, cambiando los valores de los parámetros.

En futuras líneas de investigación, sería conveniente aplicar el instrumento a una submuestra de participantes para comprobar el funcionamiento de los TAIs y así contrastar los resultados obtenidos con las simulaciones.

6.1. Objetivos de investigación

Con el objetivo de dar respuesta a la siguiente pregunta de investigación, llevamos a cabo el siguiente estudio basado en simulaciones:

- PI-K5. *¿Qué consideraciones deben tenerse en cuenta para pasar de un diseño lineal a un diseño adaptativo en este tipo de sistemas?*

6.2. Metodología

El programa objeto de este estudio consiste en dos TAI, uno para evaluar la CD de *Netiqueta* y el otro para evaluar el AC de *IAD*. Ambas pruebas incluyen un banco de ítems relativos a las CD objetivo, los cuales se han diseñado tomando como usuario final la ciudadanía. Para justificar la elección de los TAI en ambas pruebas, lo que hemos hecho ha sido comparar las pruebas lineales para ambas pruebas y sus correspondientes TAI, haciendo uso del mismo conjunto de datos simulados. A continuación, se detallan los principales aspectos considerados en las simulaciones de ambos tipos de pruebas.

6.2.1. Bancos de ítems

Los ítems seleccionados para la implementación de los TAI fueron los incluidos en ETCD basados en el proceso descrito en el capítulo 4, utilizando el modelo de Rasch para llevar a cabo su calibración. Un banco de ítems calibrado desempeña un

papel fundamental en el TAI (Wise y Kingsbury, 2000), siendo una colección de ítems psicométricos bien organizados en los que los ítems contienen datos como el objetivo a evaluar, el nivel de dificultad del ítem y otros rasgos psicométricos del ítem. Además, si tenemos dos puntuaciones medidas a partir de dos conjuntos diferentes de ítems de la prueba, estas pueden compararse ya que todos los ítems proceden del mismo banco de ítems calibrado.

Las características de los ítems se revisaron en función de la validez empírica, la fiabilidad de la prueba y el índice de dificultad de los ítems. La validez de los ítems se determinó a partir de la adecuación entre el ítem y el modelo de puntuación. Los detalles de los ítems, sus formatos, y las SC a las que pertenecen pueden ser consultados en el capítulo 4. El banco inicial de ítems quedó establecido con la siguiente distribución para ambas pruebas:

- La prueba de *IAD* incluye ítems de las 3 CD que la engloban con las siguientes proporciones: CD1 (20 ítems); CD3 (20 ítems); y CD4 (20 ítems).
- La prueba de *Netiqueta* incluye 44 ítems.

Ambos bancos de ítems son dicotómicos, es decir, con sólo dos opciones de respuesta, correcto o incorrecto, para todos los distintos formatos de ítems incluidos (de opción múltiple, ítems basados en imágenes con información, simulaciones, etc.). Las características de los ítems se revisaron en función de la validez empírica, la fiabilidad de las puntuaciones y el índice de dificultad de los ítems. En la literatura revisada es ampliamente aceptado que un ítem es bueno si tiene un valor de $-2 \leq b \leq 2$ (Hambleton et al., 1991), tal y como fue el caso de los índices de dificultad de ambos bancos de ítems. Luego, el índice de dificultad de los ítems fue aceptado como bueno.

La estructura de los bancos de ítems para ambas pruebas, así como la distribución por cada sub-competencia se muestra en la tabla 6.1.

Prueba	N.º de ítems	SC
Netiqueta	10	SC1
	11	SC2
	16	SC3
	7	SC4
IAD	10	SC5
	10	SC6
	10	SC7

	10	SC8
	20	SC9

Tabla 6.1. Distribución del número de ítems para cada sub-competencia en cada prueba.

La teoría de la respuesta al ítem (TRI) fue introducida por primera vez por Lord (1952). De los modelos de la TRI, el modelo de Rasch es el más utilizado y su base teórica está basada en la relación entre el nivel de habilidad de una persona y de la dificultad del ítem. En nuestro contexto, la capacidad de la persona se entiende como la CD del examinando, concretamente en *Netiqueta* o en *IAD*, según la prueba realizada. El análisis de los datos recogidos se llevó a cabo con el software *jMetrik* versión 4.1.1, basado en el modelo de valoración de Rasch.

6.2.2. Diseño de las simulaciones para las pruebas lineales

El conjunto de datos generado sigue la estructura del banco de ítems basada en el modelo de Rasch. Los niveles de capacidad reales los configuramos para que fuesen tomados de una secuencia regular de -3 a 3 en pasos de 0,2 (por lo que hay 31 valores diferentes en total). Además, generamos 1.000 patrones de respuesta para cada nivel de capacidad real, con lo que obtuvimos 31.000 examinados simulados. De esta forma, buscamos poder comparar las eficiencias de ambas pruebas, el lineal y TAI, en la clasificación de una variedad detallada de niveles de habilidad y con suficientes participantes por cada nivel de habilidad.

Para la generación de datos utilizamos la función `genPattern()` de la librería *catR*. A continuación usamos la función `thEst()` para extraer, para un patrón de respuesta dado del conjunto de datos, los límites inferior y superior del intervalo de confianza del 95% estimando con el método de máxima verosimilitud (ML) la capacidad y todo el banco de ítems de la prueba (*Netiqueta* e *IAD*), utilizando las funciones `thetaEst()` y `semTheta()`. Para extraer los límites inferior y superior de cada patrón de respuesta generado utilizamos el siguiente código: `res.linear <- t(apply(data, 1, thEst))`. Un esquema del código utilizado para diseñar las simulaciones de las pruebas lineales puede verse en la figura 6.1. Los resultados obtenidos de la evaluación lineal simulada para ambas pruebas son los que usamos posteriormente para compararlo con los resultados obtenidos en las simulaciones de los TAI.

```
1 # Uso de la librería carR
2 > library(catR)
3
4 #Generación de los patrones de respuesta donde it.1PL
5 corresponde a cada uno de los dos bancos de ítems #previamente
6 calibrados con las respuestas reales mediante el modelo de
7 Rasch.
8 s <- seq(-3, 3, 0.2)
9 th <- rep(s, each = 1000)
10 data <- genPattern(th, it.1PL, seed = 1)
11
12 #Extracción de los límites inferior y superior del intervalo
13 de confianza del 95% estimando con el método de # máxima
14 verosimilitud 13 (ML)
15 thEst <- function(x) {
16 + pr <- thetaEst(it.1PL, x, method = "ML")
17 + se <- semTheta(pr, it.1PL, x, method = "ML")
18 + res<-c(pr+qnorm(.025)*se, pr+qnorm(.975)*se)
19 + return(res)}
20
```

Figura 6.1. Esquema del código utilizado para diseñar las simulaciones de las pruebas lineales

6.2.3. Diseño de las simulaciones para los TAI

Para diseñar los TAI nos basamos en un esquema para sus componentes esenciales como se muestra en la figura 6.2.

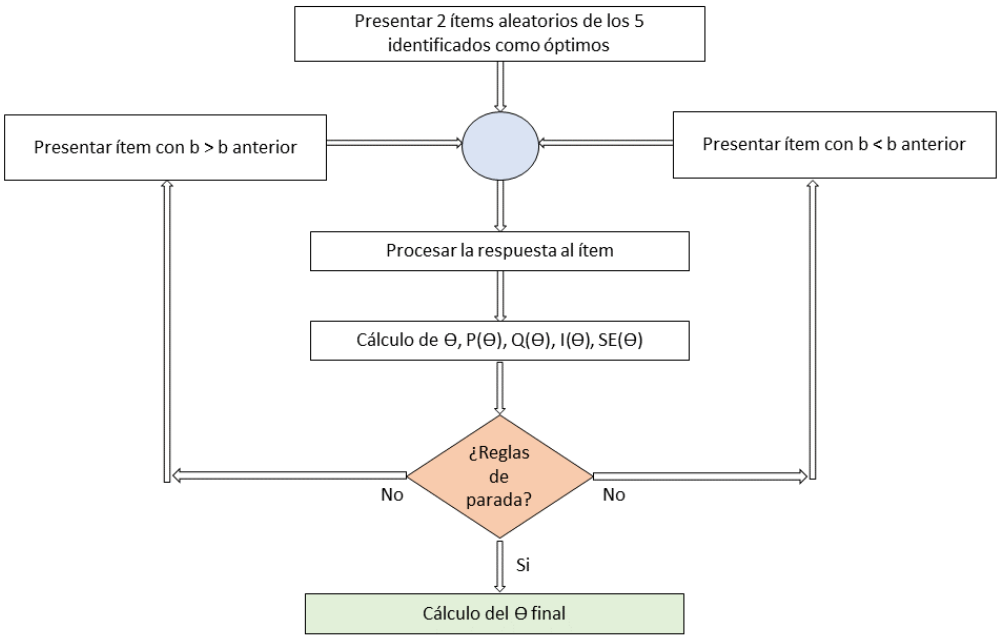


Figura 6.2. Diagrama de flujo de los TAI de ETCD.

Primero comenzamos realizando simulaciones con el conjunto de datos para comparar los resultados con las mismas respuestas a los ítems de los participantes. Además, seleccionamos el umbral de clasificación de -0,1 para la prueba de *Netiqueta* y de 0 para la prueba de *IAD*, ya que corresponden a los niveles de capacidad que maximizan la función de información del banco (ver figuras 6.3 y 6.4).

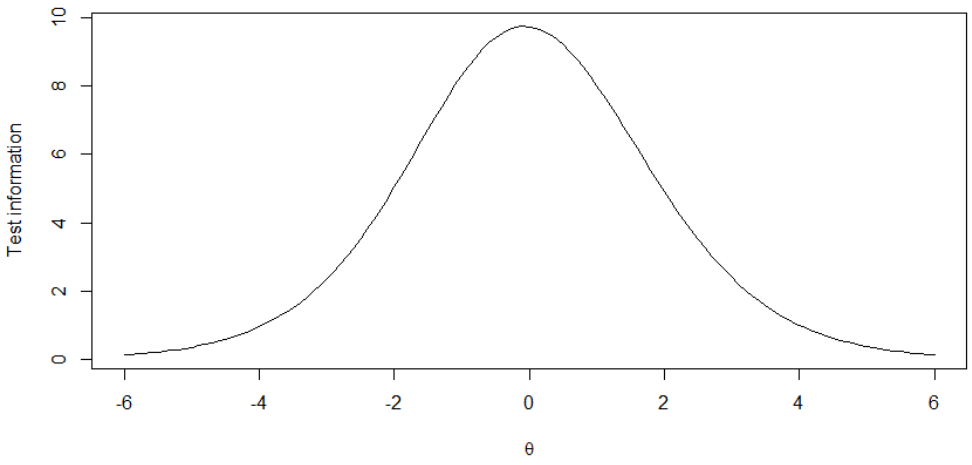


Figura 6.3. Información de la prueba de Netiqueta.

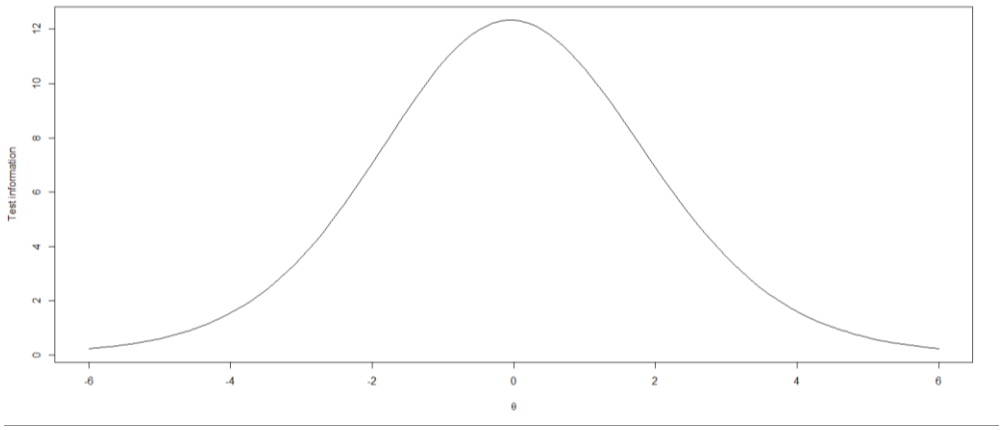


Figura 6.4. Información de la prueba de Información y alfabetización digital (IAD).

A continuación, establecimos el diseño del TAI con las siguientes características. El TAI comienza mostrando dos elementos elegidos como más informativos en torno a los valores de capacidad de -1 y 1 (de esta manera están uniformemente repartidos en torno al umbral de clasificación de ambas pruebas (-0,1 y 0 respectivamente), llevando a cabo una selección aleatoria (*randomesque*) entre los cinco elementos óptimos por valor de capacidad inicial. A continuación, el TAI elegirá el siguiente elemento utilizando la información de *Kullback-Leibler* (KL) propuesta por Chang y Ying (1996). Decidimos no utilizar el criterio tradicional de selección de ítems basado en la maximización de la función de información de Fisher, debido a que, si la estimación actual de θ está lejos de su valor real, lo cual es muy probable en las etapas iniciales de un TAI, este criterio puede ser inapropiado. Como una alternativa, Chang y Ying (1996) propusieron un procedimiento de selección de ítems basado en la información global promedio, llamado regla de selección de ítems de Kullback-Leibler (KL). Este procedimiento se basa en la distancia entre la capacidad verdadera θ y la estimación de la capacidad posterior esperada de θ (*EAP*). Cuanto mayor sea el valor de esta información, mayor será la discrepancia entre las dos funciones. La estimación provisional y final de la capacidad la realiza mediante el método ML (como en las pruebas lineales), con un ajuste variable del tamaño de los pasos al inicio del TAI en caso de que el patrón sea constante. Finalmente, el criterio de parada de la prueba define el momento en que se deja de administrar ítems al examinando. Los dos criterios de parada comúnmente más utilizados se basen en definir un número fijo de ítems a administrar en la prueba, o definir una precisión mínima para la estimación del rasgo latente, es decir, un valor mínimo predeterminado para su error estándar. En ambas pruebas, decidimos que el TAI dejará de administrar ítems y se detuviese cuando el examinando pudiera ser clasificado con respecto a sus respectivos umbrales con un 95% de confianza, o cuando la prueba alcanzase 33 o 45 ítems respectivamente (que suponen 3/4 del tamaño del banco de ítems total). Por cada TAI que se administre obtendremos la longitud total de la prueba (con un máximo de 33 y 45 ítems respectivamente, un 25% menos de ítems que la prueba lineal), y los límites inferior y superior del intervalo de confianza final del 95%. Un esquema

del código utilizado para diseñar las simulaciones de las pruebas lineales puede verse en la figura 6.5.

```
#Uso de las mismas respuestas para poder comparar los resultados del enfoque lineal y del adaptativo
> res.cat <- matrix(NA, nrow(data), 3)

# Para fijar la semilla de la selección aleatoria, en la definición de las listas de inicio y de pruebas se lleva a cabo basado en un bucle sobre todos los examinados. El resultado se almacena en una matriz con una fila por examinando y tres columnas (longitud de la prueba, límite inferior y límite superior, respectivamente).

# Además es necesario proporcionar algún valor de entrada de las respuestas para la compatibilidad, luego ponemos a cero trueTheta aunque no se utiliza más en el   14 proceso.
> for (i in 1: nrow(data)) {
start <- list(theta = c(-1, 1), randomesque = 5, random.seed = i)
test <- list(method = "ML", constantPatt = "var", itemSelect = "KL", randomesque = 10, random.seed = i)

#Para la prueba de Netiqueta 33 items y -0.1 como umbral
 stop <- list(rule = c("length", "classification"), thr = c(33, -0.1))

#Para la prueba de IAD 45 items y 0 como umbral
 stop <- list(rule = c("length", "classification"), thr = c(33, -0.1))
final <- list(method = "ML")
pr <- randomCAT(trueTheta = 0, itemBank = it.1PL, responses = data[i,], start = start, test = test, stop = stop, final = final)

res.cat[i,] <- c(length(pr$pattern), pr$ciFinal)
}
```

Figura 6.5. Esquema del código utilizado para diseñar las simulaciones de las pruebas lineales.

En este estudio nos hemos decantado por definir dos criterios de parada, pero la opción de administrar un número fijo de ítems podría ser otra opción coherente con el actual sistema de BAIT, donde los participantes cuentan con un máximo de dos horas para terminar la prueba. De esta forma, nos evitaríamos los problemas que podrían surgir de que no todos los usuarios respondan la misma cantidad de ítems y les genere cierto malestar. Lo examinaremos en profundidad en futuros estudios.

6.3. Resultados

A continuación, se muestran los resultados obtenidos en las simulaciones para ambas pruebas, así como la comparación entre la versión lineal y la versión adaptativa.

6.3.1. Netiqueta

A continuación, examinamos las dos matrices de salida resultado de ejecutar ambas simulaciones (es decir, la versión lineal y la versión adaptativa), junto con el vector de niveles de habilidad verdaderos utilizado para generar los datos. Para ambos versiones de la prueba, calculamos por cada nivel de habilidad, el porcentaje de clasificaciones correctas (es decir, porcentajes de participantes cuyo intervalo de confianza se encuentra en el lado correcto del umbral de clasificación) (ver figura 6.6), el porcentaje de clasificaciones incorrectas (es decir, porcentaje de participantes cuyo intervalo de confianza se encuentra en el lado incorrecto del umbral de clasificación) (ver figura 6.7) y el porcentaje de clasificaciones indeterminadas (es decir, porcentaje de participantes cuyo intervalo de confianza se solapa con el umbral de clasificación) (ver figura 6.8).

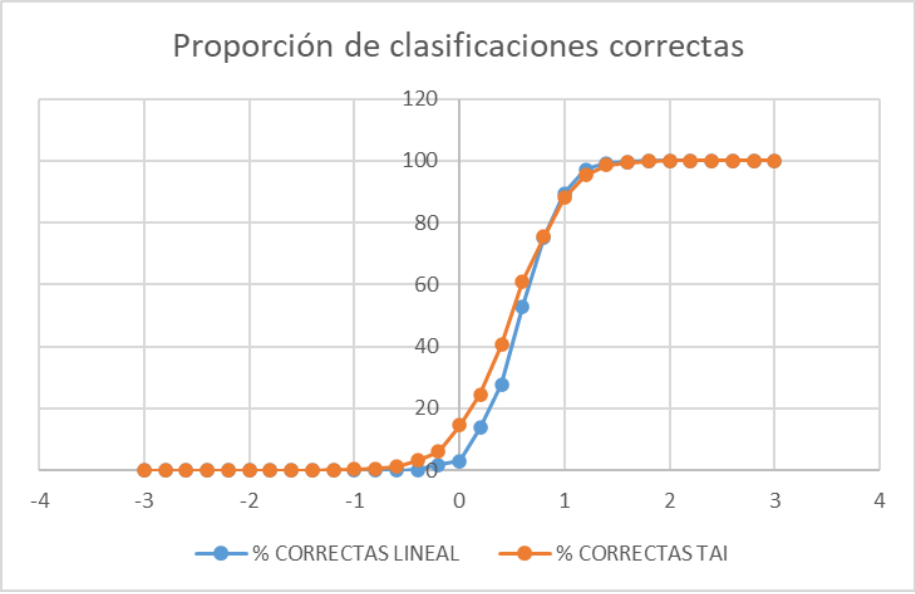


Figura 6.6. Porcentajes de clasificaciones correctas para la prueba de Netiqueta.

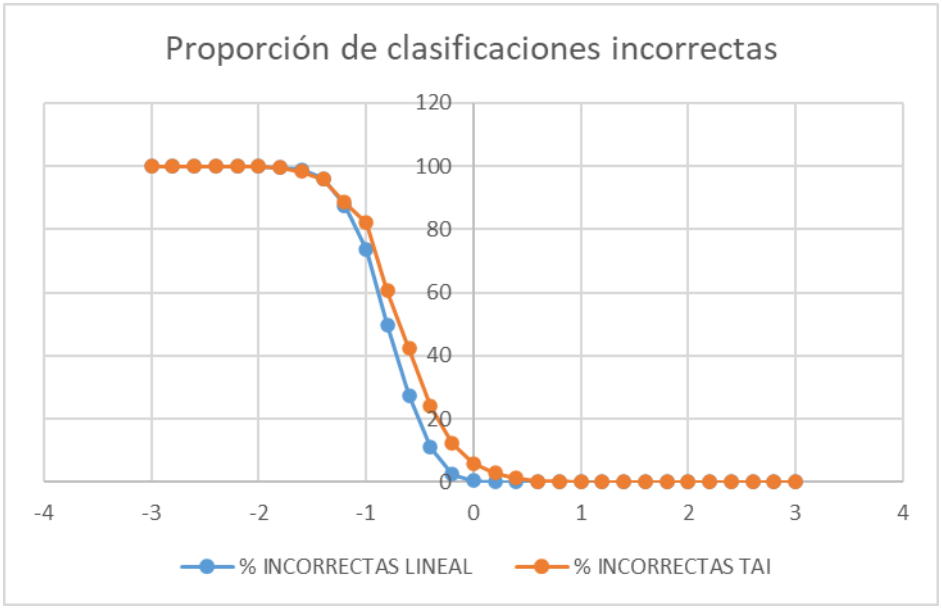


Figura 6.7. Porcentajes de clasificaciones incorrectas para la prueba de Netiqueta.

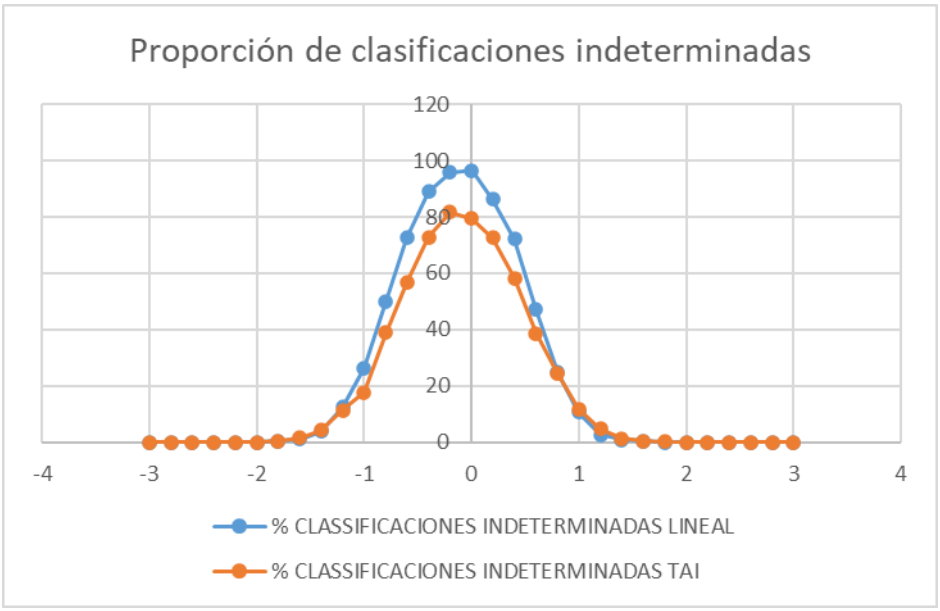


Figura 6.8. Porcentajes de clasificaciones indeterminadas para la prueba de Netiqueta.

Además, en la figura 6.9 se muestra el gráfico de cajas de las distribuciones de las longitudes de las pruebas TAI por nivel de capacidad.

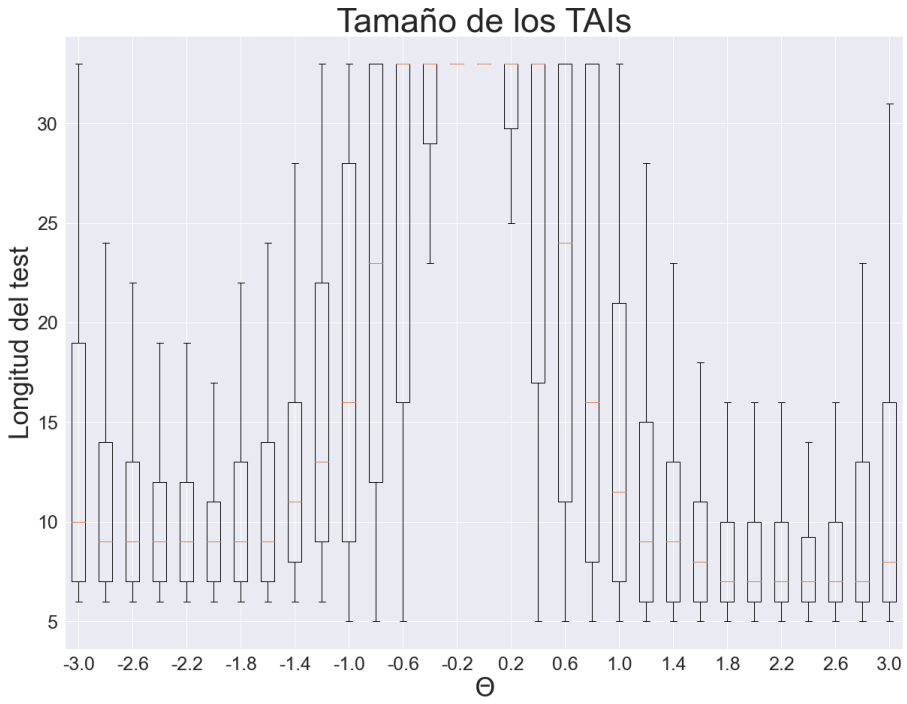


Figura 6.9. Longitudes de las pruebas adaptativas para cada nivel de capacidad real ϴ para la prueba de Netiqueta

En primer lugar, se puede apreciar como las diferencias entre las evaluaciones lineal y la adaptativa se producen principalmente en torno al umbral de clasificación. En el límite superior de esta escala de capacidad, el porcentaje de clasificaciones correctas son casi máximas y consecuentemente, los otros dos tipos de porcentajes se aproximan a cero.

A medida que el nivel de capacidad se acerca al umbral de clasificación, nos encontramos las tendencias esperadas: la proporción de clasificaciones correctas disminuye drásticamente, mientras que la proporción de clasificaciones indeterminadas aumenta de forma casi compensatoria, pero en menor medida. Esto se debe al hecho de que discriminar los niveles de habilidad verdaderos alrededor del umbral de clasificación es mucho más difícil que para niveles de habilidad alejados del umbral, lo que indica el mayor riesgo de cometer un error de clasificación cuando uno se encuentra cerca del umbral de decisión.

Finalmente, cabe destacar las diferencias entre ambos tipos de evaluación en torno al umbral de clasificación. En primer lugar, la proporción de clasificaciones correctas es mayor para el diseño adaptativo (14,6%) que para el diseño lineal (3% en el mismo nivel). En segundo lugar, la proporción de clasificaciones incorrectas es menor para el diseño adaptativo que para las pruebas lineales: 0,5% frente a 6% en torno al umbral de clasificación. Por último, el porcentaje de clasificaciones indeterminadas es mayor para el ensayo lineal (valor máximo de 96,5%) que para el diseño adaptativo (porcentaje correspondiente de 79,4%).

En conjunto, estos resultados indican que, a pesar de administrar menos ítems que en las pruebas lineales, los diseños adaptativos devuelven con más frecuencia alguna clasificación final (es decir, menos conclusiones indeterminadas), lo que conduce a un aumento de los porcentajes tanto de las clasificaciones correctas como de las incorrectas (ver tabla 2). Sin embargo, entre los casos indeterminados con pruebas lineales que arrojan una clasificación con el diseño adaptativo, hay en general más clasificaciones correctas que incorrectas.

Tipo de enfoque	Clasificaciones indeterminadas	Clasificaciones correctas	Clasificaciones incorrectas
Lineal	9.927	12.602	11.472
Adaptativo	5.772	13.088	12.141

Tabla 6.2. Diferencias de clasificaciones totales entre ambos enfoques.

En resumen, el diseño adaptativo mejoró o mantuvo la clasificación de la mayoría de los examinados con una reducción de la longitud de la prueba de al menos un 25%, incluso más cuando el nivel de capacidad real se encuentra muy lejos del umbral de clasificación, tal y como puede apreciarse en la figura 6.7. El TAI devuelve las clasificaciones finales con más frecuencia que las pruebas lineales, a costa de un aumento de la tasa de clasificación incorrecta alrededor del umbral. Pero la ganancia parece ser globalmente significativa.

6.3.2. Información y alfabetización digital (IAD)

A continuación, examinamos las dos matrices de salida resultado de ejecutar ambas simulaciones (es decir, la versión lineal y la versión adaptativa), junto con el vector de niveles de habilidad verdaderos utilizado para generar los datos. Para ambos versiones de la prueba, calculamos por cada nivel de habilidad, el porcentaje de clasificaciones correctas (es decir, porcentajes de participantes cuyo intervalo de confianza se encuentra en el lado correcto del umbral de clasificación) (ver figura 6.10), el porcentaje de clasificaciones incorrectas (es decir, porcentaje de participantes cuyo intervalo de confianza se encuentra en el lado incorrecto del umbral de clasificación) (ver figura 6.11) y el porcentaje de clasificaciones indeterminadas (es decir, porcentaje de participantes cuyo intervalo de confianza se solapa con el umbral de clasificación) (ver figura 6.12).

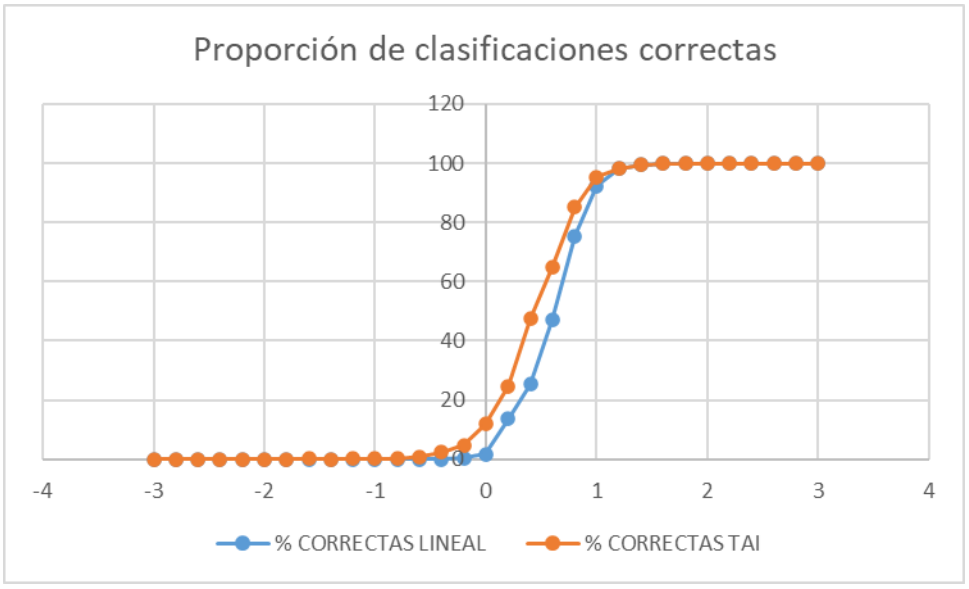


Figura 6.10. Porcentajes de clasificaciones correctas para la prueba de IAD.

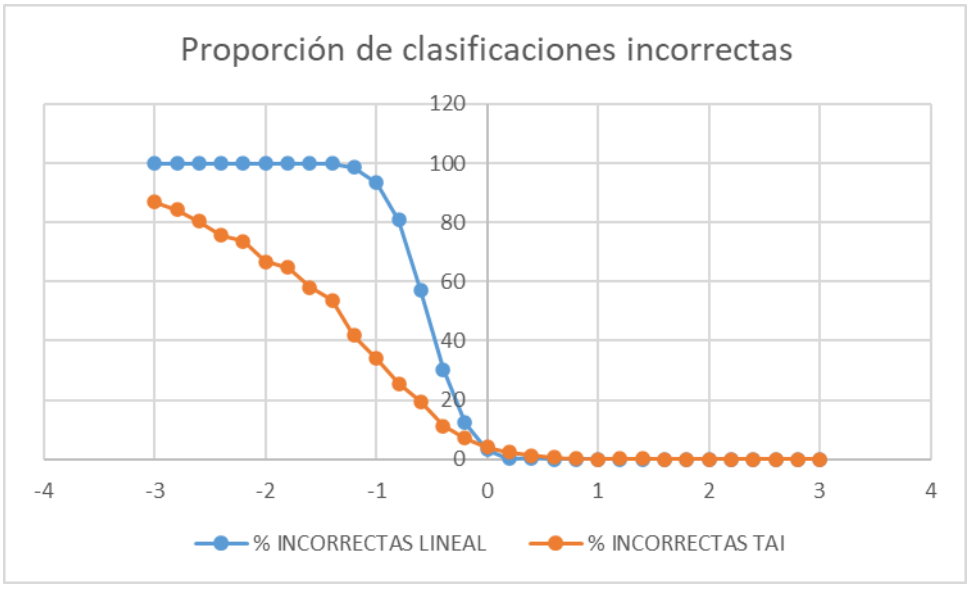


Figura 6.11. Porcentajes de clasificaciones incorrectas para la prueba de IAD.

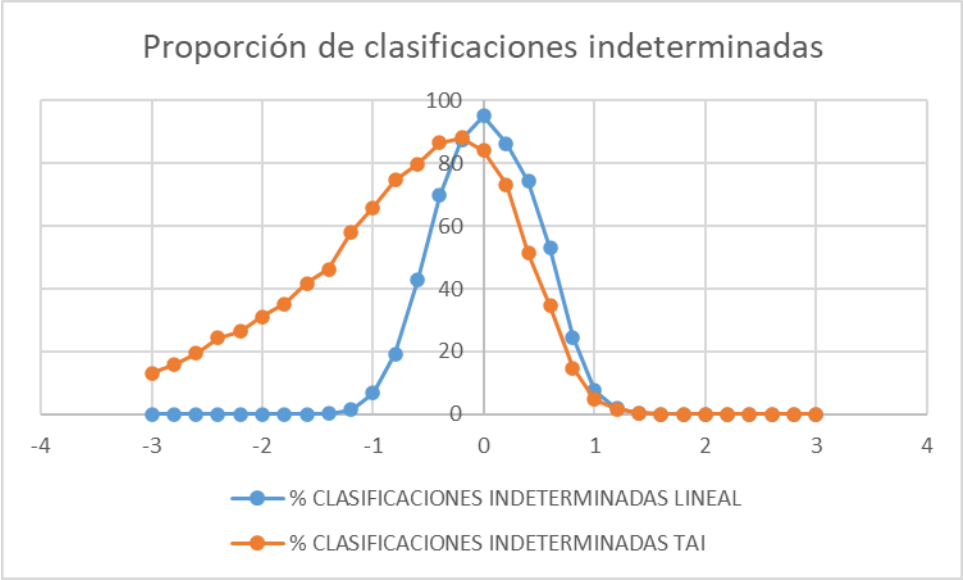


Figura 6.12. Porcentajes de clasificaciones indeterminadas para la prueba de IAD.

Además, en la figura 6.13 se muestra el gráfico de cajas de las distribuciones de las longitudes de las pruebas TAI por nivel de capacidad.

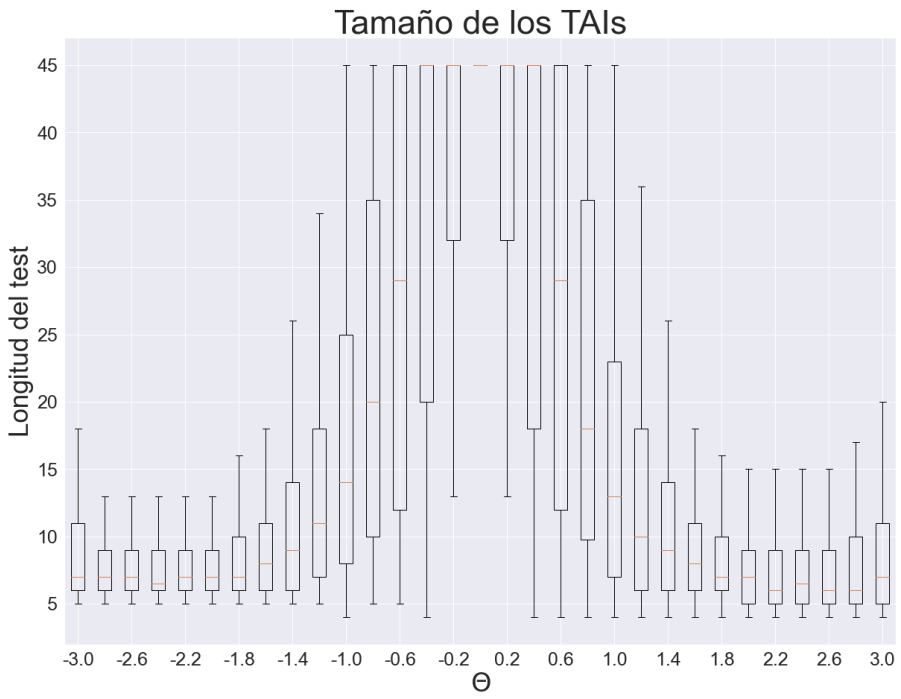


Figura 6.13. Longitudes de las pruebas adaptativas para cada nivel de capacidad real ϴ para la prueba de IAD

En primer lugar, se puede apreciar como las diferencias entre las evaluaciones lineal y la adaptativa se producen principalmente en torno al umbral de clasificación. En el

límite superior de esta escala de capacidad, el porcentaje de clasificaciones correctas son casi máximas y consecuentemente, los otros dos tipos de porcentajes se aproximan a cero.

Al igual que vimos en el TAI de Netiqueta, cuando el nivel de capacidad se acerca al umbral de clasificación, disminuye significativamente la proporción de clasificaciones correctas, aumentando las clasificaciones indeterminadas, ya que discriminar niveles de habilidad alrededor del umbral de clasificación es mucho más difícil que para niveles de habilidad alejados del umbral.

Finalmente, cabe destacar las diferencias entre ambos tipos de evaluación en torno al umbral de clasificación. En primer lugar, la proporción de clasificaciones correctas es mayor para el diseño adaptativo (11,9%) que para el diseño lineal (1,7% en el mismo nivel). En segundo lugar, la proporción de clasificaciones incorrectas es ligeramente mayor para el diseño adaptativo que para las pruebas lineales: 4% frente a 3,3% en torno al umbral de clasificación. Por último, el porcentaje de clasificaciones indeterminadas es mayor para el ensayo lineal (valor máximo de 95%) que para el diseño adaptativo (porcentaje correspondiente de 84,1%).

En conjunto, estos resultados indican que, a pesar de administrar menos ítems que en las pruebas lineales, los diseños adaptativos devuelven con menos frecuencia alguna clasificación final (es decir, más conclusiones indeterminadas) (Ver tabla 6.3). En cambio, el diseño adaptativo da lugar a más clasificaciones correctas y reduce considerablemente la cantidad de clasificaciones incorrectas.

Tipo de enfoque	Clasificaciones indeterminadas	Clasificaciones correctas	Clasificaciones incorrectas
Lineal	5.705	12.537	12.759
Adaptativo	9.715	13.352	7.917

Tabla 6.3. Diferencias de clasificaciones totales entre ambos enfoques.

En resumen, el diseño adaptativo mejoró o mantuvo la clasificación de la mayoría de los examinados con una reducción de la longitud de la prueba de al menos un 25%, incluso más cuando el nivel de capacidad real se encuentra muy lejos del umbral de clasificación, como puede verse en la figura 6.13. El TAI devuelve las clasificaciones finales con menos frecuencia que las pruebas lineales, pero aumenta la tasa de clasificación correcta y disminuye considerablemente la tasa de clasificación incorrecta.

6.4. Conclusiones y discusión

El uso de las simulaciones como las llevadas a cabo en este estudio nos sirvieron para comprobar la viabilidad de diseñar una prueba adaptativa para evaluar la CD de la ciudadanía, y dar respuesta a la pregunta de investigación:

- PI-K5. *¿Qué consideraciones deben tenerse en cuenta para pasar de un diseño lineal a un diseño adaptativo en este tipo de sistemas?*

  El uso de simulaciones permite la compilación inmediata de resultados y no provoca dificultades en los participantes que tendrían que realizar las pruebas. Los resultados preliminares presentados muestran que pruebas con una reducción de al menos un 25% en la longitud de las pruebas, parece ser suficiente para estimar satisfactoriamente el rasgo latente, corroborando los hallazgos de Vam der Linden y Pashley (2009). El uso de TAI es una forma de evaluación moderna y eficiente, y supondrá un menor consumo de tiempo para la realización de las pruebas y proporcionará una forma gratuita de evaluar el nivel de CD de cualquier ciudadano.

  Según los resultados, las versiones adaptativas de ambas pruebas han reducido su tamaño considerablemente, así como ofrecido una mejora notable en la proporción de clasificaciones recibidas. Aun así, sería conveniente continuar llevando a cabo nuevos análisis como, p. ej., aplicando diferentes reglas de parada con diferentes longitudes de pruebas, aplicando diferentes reglas para seleccionar el siguiente ítem o para estimar los niveles de habilidad, etc. También convendría examinar la influencia de las características de los ítems de los bancos de ítems en la precisión de la estimación de los parámetros latentes.

  Además, cabe destacar que en las simulaciones diseñadas no hemos tenido en cuenta imponer restricciones para balancear los contenidos a nivel de CD y/o SC, ya que se tratan de simulaciones a pequeña escala con bancos de ítems reducidos. En futuros estudios deberíamos considerar la ampliación del número de ítems de los bancos y, así se podrían añadir restricciones más complejas para el diseño de la prueba, como p. ej. distribuciones determinadas a nivel de SC. La función `randomCAT()` que hemos utilizado en las simulaciones permite implementar el control de equilibrio de contenidos. Esta función admite como argumento de entrada el parámetro `cbControl`, que es una lista que sirve para determinar la proporción de elementos a determinar en base a una serie de categorías que se han debido de establecer previamente en los bancos de ítems. De esta forma, p. ej., podríamos configurar las pruebas de *IAD* para que se mostrase a los participantes al menos un mínimo de preguntas similar por cada una de las 3 CD que engloba.

Si el objetivo del TAI no es estimar las competencias de los participantes, sino clasificarlos en "aprobado" o "suspendido", como puede ser el caso de una evaluación orientada a la certificación como BAIT, sería conveniente establecer un punto de corte en la escala de rasgos latentes. Actualmente en BAIT, es requerido alcanzar al menos un 70% de la puntuación total de la prueba para obtener la certificación. Con el fin de equiparar este requerimiento al contexto del TAI, llevaremos a cabo futuros trabajos utilizando los datos de la aplicación real para establecer el punto de corte en cada prueba.

Una vez que profundicemos en la configuración más adecuada, los siguientes pasos que realizaremos en futuros trabajos tendrán como objetivo dotar a ETCD de un motor adaptativo. Teniendo en cuenta que ETCD ha sido desarrollado haciendo uso del lenguaje Java, integraremos las librerías del paquete *catR* mediante alguna tecnología para integrar R en Java, como p. ej. *RCaller*[56]. La arquitectura resultante se muestra en la figura 6.14.

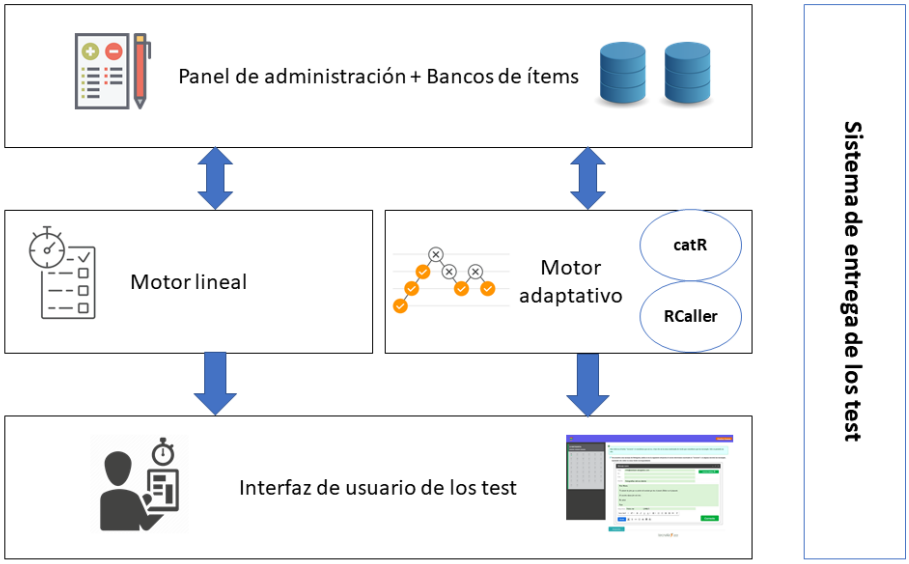


Figura 6.14. Arquitectura de ETCD incluyendo el motor adaptativo.

[56] https://github.com/jbytecode/rcaller

No seas menos de lo que eres capaz o serás profundamente infeliz.

Abraham Maslow

7.

Conclusiones

A lo largo del estudio, hemos explicado los principios de diseño aplicados en cada fase de desarrollo de la herramienta de evaluación de CD presentada. Realizamos una implementación a medida basada en DigComp, que contribuyó a crear un entendimiento común para debatir sobre la CD como un componente transversal necesario para distintos PD. Esta contribución fue esencial para la valoración de las pruebas de evaluación, el diseño de los ítems, su nivel de dificultad y adecuación, y la información proporcionada al final de las pruebas. Administramos las pruebas a expertos y a usuarios finales. Con los datos recogidos, realizamos una serie de estudios de validación para evaluar la calidad de la herramienta. Una vez analizados los resultados de la primera versión propuesta, desarrollamos una segunda propuesta diseñando una prueba basada en una CD y otra prueba basada en un AC. Por último, llevamos a cabo un estudio basado en simulaciones para examinar la viabilidad de convertir las pruebas en TAI. Los resultados preliminares apuntan que con una reducción del 25% de la longitud de las pruebas, se puede estimar satisfactoriamente el rasgo latente, mejorando notablemente la proporción de clasificaciones recibidas.

Recogemos en el presente capítulo las conclusiones de nuestra investigación, resumiendo el proceso llevado a cabo, indicando las limitaciones, resumiendo los resultados obtenidos, indicando algunas de las posibles aplicaciones y trazando algunas de las líneas futuras de investigación que planteamos.

7.1. Resumen del proceso de investigación

El lanzamiento de la primera versión de DigComp por Ferrari y Punie (2013), supuso un punto de inflexión en materia de CD. La CD de la ciudadanía pasó a ser una de las principales prioridades de la agenda política europea, que buscaba mejorar las capacidades y CD para la transformación digital. DigComp, el cual proporciona un lenguaje común para identificar y describir las áreas clave de la CD, se diseñó para mejorar la CD de la ciudadanía, ayudando a los responsables políticos a diseñar políticas e iniciativas que apoyen su desarrollo.

Desde el Gobierno Vasco se apostó por desarrollar BAIT, incorporando el marco DigComp alineado con las directrices estratégicas definidas en la Agenda Digital Europea[57], en un nuevo desarrollo que formaba parte del proyecto IKANOS[58]. IKANOS, desarrollado en el marco de despliegue de la Agenda Digital de Euskadi 2020[59], tiene como objetivo principal promover la difusión e impulsar la adopción en Euskadi del Marco Europeo de CD, nuevas formas de aprendizaje y un sistema de certificación. Con este objetivo se trazó un plan de trabajo que incluía varias líneas de trabajo, parte de las cuales son parte de este estudio. A continuación, resumimos las distintas fases en las que hemos llevado a cabo nuestro trabajo.

Fase 1. Desarrollo de P4E para la evaluación y acreditación de PD y establecimiento de hipótesis

En 2017 comenzó tanto el desarrollo de BAIT, como el proyecto P4E (ERASMUS+ 2016-1-ES01 KA204-024983). Ambos partieron de una revisión de literatura científica para identificar los principales avances y carencias en la definición, evaluación y acreditación de PD (ver punto 3.2). De ese análisis, consideramos oportuno diseñar una nueva herramienta de evaluación y acreditación de CD abordando las principales carencias detectadas, con la idea de ir aplicando en BAIT todos los progresos alcanzados. Definimos nuestro conjunto de hipótesis, detalladas en el capítulo 1, de acuerdo con el estado del arte en ese momento. En esta fase, nos centramos en el primer objetivo de diseñar una herramienta de evaluación de CD, suficientemente flexible como para permitir la incorporación de distintos formatos de ítems que nos permitan evaluar habilidades de orden cognitivo de distinto orden. Ante la ausencia de tecnologías configurables que nos permitiesen la incorporación de PD para su posterior evaluación y acreditación, decidimos diseñar de forma personalizada nuestra propia herramienta P4E llevando a cabo una implementación a medida basada en DigComp. Seguimos la misma arquitectura de software que en BAIT para poder transferir a BAIT los avances realizados de la

[57] https://ec.europa.eu/digital-single-market/

[58] https://ikanos.eus

[59] http://www.spri.eus/archivos/19777/Agenda-Digital-de-Euskadi-2020.pdf

manera más rápida y eficaz. Como resultado final, quedó disponible la herramienta web P4E, con dos PD disponibles definidos con expertos (emprendedor y teletrabajador) y con 10 pruebas de evaluación disponibles, una por cada AC y PD.

Fase 2. Desarrollo de ETCD para la evaluación de CD y validación de hipótesis

En 2019 comenzamos con el desarrollo de ETCD siguiendo la metodología DBR y las fases descritas en capítulo 4 teniendo presente la siguiente pregunta que orientaba el desarrollo de la tesis: ¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?

Diseñamos y desarrollamos tanto las pruebas de evaluación de CD, como la propia herramienta de evaluación. Como parte del desarrollo, realizamos una serie de estudios de validación para dar respuesta al resto de las preguntas de investigación planteadas (PI_K2, PI_K3 y PI_K4), destacando los estudios de validación de constructo basado en expertos, o el estudio de validación de RP con tecnología de seguimiento ocular. Finalmente, en la herramienta quedó disponible una prueba de evaluación basada en una CD (Netiqueta) y otra prueba de evaluación basada en un AC (IAD). Ambas pruebas fueron administradas a usuarios finales con el objetivo de analizar sus respuestas y evaluar su calidad.

Fase 3. Análisis de una implementación adaptativa de la herramienta de evaluación y validación de hipótesis

A lo largo del 2022 llevamos a cabo un análisis detallado de los distintos aspectos a considerar para llevar a cabo una implementación adaptativa de la herramienta de evaluación ETCD y evaluar su viabilidad, para dar respuesta a la pregunta de evaluación PI_K5: ¿Qué consideraciones deben tenerse en cuenta para pasar de un diseño lineal a un diseño adaptativo en este tipo de sistemas?

A partir del análisis preliminar llevado a cabo, analizamos la viabilidad de pasar a un diseño adaptativo los test lineales de ambas pruebas (de clasificación), haciendo uso del mismo conjunto de datos simulados. Más concretamente, comparamos las clasificaciones de los examinados con el banco de ítems completo de la prueba lineal (modelo 1PL de la TRI), con el diseño TAI, usando simulaciones con el paquete R catR (Magis y Barrada, 2017). Diseñamos las simulaciones con una serie de parámetros para los componentes claves del TAI, que identificamos en la revisión bibliográfica como los más utilizados y adecuados para nuestro contexto. Los resultados apoyaron la viabilidad de un TAI con una serie de puntos clave identificados.

Fase 4. Nueva revisión de la literatura y escritura de tesis

Para finalizar, llevamos a cabo una revisión complementaria actualizando la literatura disponible y comenzamos la escritura de la memoria de tesis doctoral presentada en este documento.

7.2. Resultados de la investigación

Durante el desarrollo de este estudio, hemos hallado resultados relevantes que nos han permitido dar respuesta a nuestras preguntas de investigación:

- PI_K1. *¿Qué consideraciones deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación?*

Realizamos una serie de consideraciones que deben tenerse en cuenta para diseñar este tipo de herramientas de evaluación, en base a los resultados de nuestro trabajo:

  - Es crucial la adopción de un marco de referencia, y en concreto DigComp, el cual proporciona criterios claros y bien definidos, para debatir sobre la CD como un componente transversal necesario para una gran variedad de perfiles laborales. Además, es necesario adaptarlo a las necesidades específicas de cada implementación en todas las fases del desarrollo (ej. Definición de PD, diseño del banco de ítems, información proporcionada al finalizar las pruebas, etc.). Más aún, la credibilidad y fiabilidad del marco DigComp debido a su origen y aval de la UE, facilita su difusión y adopción.
  - Es importante considerar la adopción de un enfoque pragmático, que facilite el diseño e implementación de la solución. Garantizar la fiabilidad y la validez de una prueba de evaluación es un gran reto, especialmente cuando se la evaluación sigue un enfoque basado en el rendimiento, y más aún, cuando la se pretende evaluar un constructo complejo como la CD. Por lo tanto, evaluar cada CD como un constructo independiente, sin tener en cuenta los puntos de solapamiento a nivel de CD (como se identifican en el propio marco DigComp), puede simplificar el diseño de manera notable. Con este enfoque, se va a priorizar la validez externa de la herramienta, siempre y cuando se describan detalladamente todos los pasos y decisiones tomadas, con el objetivo de que sean comprendidas y aceptadas por un público más amplio.
  - Si el objetivo de la herramienta es no solo evaluar habilidades cognitivas de orden bajo, los que corresponden a los niveles básicos de acuerdo con DigComp, será necesario incorporar distintos formatos de ítems, incluyendo en las pruebas distintos formatos de ítems que permitan evaluar las habilidades de orden superior como, por ejemplo, mediante

simulaciones interactivas u otro tipo de formatos dinámicos. Asimismo, es necesario diseñar estos tipos de ítems correctamente para lograr desencadenar en los participantes los comportamientos previstos. Este hecho es especialmente importante cuando se evalúan constructos cognitivos complejos como la CD, donde se pretende que los participantes pongan en acción sus conocimientos para llevar a cabo la tarea solicitada, lo que proporciona una imagen más precisa de su nivel de CD que la obtenida habitualmente mediante preguntas de opción múltiple.

- Un factor clave en el desarrollo del banco de ítems, es identificar el contenido requerido para el PD objetivo e inicialmente asignar los niveles de los ítems de acuerdo con los descriptores del marco de referencia. Los descriptores de DigComp utilizan la taxonomía de Bloom, lo cual facilita la asignación del nivel de los ítems analizando los verbos de los enunciados de estos. Cabe mencionar la relevancia de prestar especial atención a la comprensión lectora de los ítems, ya que la complejidad lingüística es un factor clave que incide directamente en la tasa de éxito de las respuestas.
- Otro punto a tener presente en el desarrollo de los ítems es que hay que tratar de diseñar los ítems de las pruebas sin basarse en aplicaciones o dispositivos específicos. Sin embargo, al diseñar ciertos formatos de ítems como las preguntas de simulación interactivas, es necesario diseñarlas basándose en la interfaz de alguna aplicación o dispositivo en concreto. Lo ideal es seleccionar las herramientas y dispositivos más utilizados, y que tengan comportamientos ampliamente aceptados en lo que a términos de usabilidad se refiere.
- Debe estudiarse bien el grado de profundidad que se pretende alcanzar en las pruebas. Hemos podido identificar varios enfoques en las implementaciones que se han llevado a cabo: pruebas basadas en CD, pruebas basadas en AC y pruebas basadas en el marco DigComp completo. Las 3 opciones son válidas, pero normalmente difieren en el número de preguntas por CD, y por lo tanto en el grado de profundidad que se pretende alcanzar en la evaluación.

- PI_K2. *¿Qué propiedades psicométricas tienen las pruebas? ¿Qué evidencias se pueden presentar que soporten las inferencias realizadas de las puntuaciones obtenidas?*

  - En el desarrollo de un instrumento cuantitativo con fines de evaluación, es crucial medir su calidad. En vista de ello, conviene planificar diversos estudios a lo largo de las diferentes fases del desarrollo para obtener suficientes evidencias que aseguren la calidad de la herramienta.

- Tras la revisión de los conocimientos y las prácticas actuales, definimos nuestros objetivos basándonos en los principios de diseño identificados en varios estudios clave.
- A continuación, mediante un estudio de validación basado en el juicio de expertos (facilitadores de KZgunea) confirmamos que el contenido de las pruebas representaba el constructo que se pretendía evaluar y cumplía los objetivos de la prueba.
- Además, llevamos a cabo un estudio para validar los RP de una selección de ítems incluidos en las pruebas, obteniendo información útil para comprender el rendimiento de los participantes e investigar si los criterios de evaluación de los ítems estaban correctamente establecidos. Para el resto de los ítems incluidos en las pruebas, aplicamos los principios de diseño identificados en el estudio.
- Respecto a la estructura interna de las pruebas, a pesar de que la multidimensionalidad del constructo de CD ha sido identificada en varios estudios, al seguir un enfoque pragmático con el objetivo de simplificar el diseño, tomamos las siguientes decisiones: para la prueba de *IAD*, consideramos las tres CD del AC como dimensiones independientes y mostramos como el modelo Rasch tridimensional se ajusta mejor a los datos que el modelo Rasch unidimensional; para la prueba de *Netiqueta* consideramos las cuatro SC seleccionadas como dimensiones independientes, y mostramos como el modelo Rasch cuatridimensional se ajusta mejor a los datos que el modelo Rasch unidimensional. Sin embargo, teniendo en cuenta las altas correlaciones obtenidas, parece que todos los ítems se relacionan con un factor fuerte, que puede interpretarse como CD general.
- También calculamos la estimación EAP/PV para investigar la consistencia interna de todas las dimensiones, con valores entre 0.84 y 0.88 para la prueba de *IAD* y valores entre 0.78 y 0.86 para la prueba de *Netiqueta*, siendo todos los coeficientes superiores a 0.70 lo que indica una buena consistencia interna. El alfa de Cronbach para las pruebas globales fue de 0.93 para la prueba de *IAD* y 0.89 para la prueba de *Netiqueta*. Ambas pruebas mostraron propiedades psicométricas sólidas que los convierten en instrumentos fiables y válidos para medir la CD. Es decir, ambos enfoques (basados en AC o en CD) son válidos, aunque difieran en el grado de profundidad de evaluación de las CD cubiertas por sus respectivas pruebas.

- PI_K3. *¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos generar y probar inferencias sobre el constructo de interés?*

  - Encontramos muy interesante la fuente de datos que usamos proveniente del dispositivo de seguimiento ocular utilizado, no disponible en un TEA

tradicional, para generar y probar inferencias sobre el constructo de interés con fines de validación. Lo que hicimos fue evaluar una interpretación alternativa de las puntuaciones en las pruebas, que podría poner en entredicho los criterios de evaluación de los ítems. Para ello, examinamos las rutas de exploración de los ítems basados en una imagen o simulación. Investigamos cómo los participantes procesaban las diferentes AOI de las imágenes incluidas en los ítems, para evaluar la situación y elegir la respuesta correcta.

- En primer lugar, examinamos qué AOI dentro de los ítems fueron examinados, con el objetivo de comprobar si patrones específicos de AOI visitados podrían poner en evidencia que los ítems requieran el mismo proceso cognitivo que se requiere en las tareas del mundo real según los criterios de evaluación previamente definidos. Los datos del seguimiento ocular nos permitieron confirmar la no existencia de patrones de respuesta con una tasa de visita inesperada, difícil de explicar que hubiera planteado dudas sobre si el ítem estaba generando los RP esperados.
- En segundo lugar, llevamos a cabo una agrupación de las respuestas en términos de AOI visitadas, aplicando un algoritmo de clasificación k-means no supervisado para examinar si patrones específicos de AOI visitadas suponían mayores tasas de éxito en el rendimiento global (la puntuación global considerando la tarea relacionada con la misma CD de la tarea en cuestión) y de manera inconsistente con las expectativas para cada ítem. Pudimos identificar clusters con mayores tasas de éxito, sin embargo, los resultados estaban lejos de ser significativos y, por lo tanto, no fue posible concluir que alguno de los patrones de comportamiento fuera más eficiente que el resto. Sería muy interesante volver a realizar este análisis con un mayor número de participantes.

- PI_K4. *¿Permite el análisis de las interacciones del participante durante la prueba entender su comportamiento, de manera que podamos utilizar esta información para mejorar los diseños de los ítems?*

  - Pudimos comprobar como examinando las interacciones de los participantes provenientes de los datos del seguimiento ocular, identificamos notables diferencias entre quienes acertaron y quienes fallaron los ítems. Por ejemplo, los participantes que fallan tenían rutas de exploración más largas que los que aciertan, y la varianza es mayor en el grupo de participantes que fallaron. Esta información la pudimos utilizar para la mejora del diseño de los ítems.
  - Calculando las rutas de exploración comunes de los participantes que acertaron, pudimos profundizar en las estrategias de resolución que habían seguido, obteniendo una representación visual que nos permitió comprobar de forma sencilla que los participantes que respondieron correctamente no mostraron un comportamiento imprevisto que invalidara los criterios de

evaluación del ítem. Encontramos esta representación gráfica de especial interés para poder complementar el diseño de los ítems y poder mostrarla en las revisiones de los ítems a los participantes que la hubiesen fallado, añadiendo un valor extra al proceso de revisión.

- Examinando donde se realizaron las fijaciones más largas, pudimos confirmar que los participantes que respondieron correctamente las realizaron en las zonas esperadas, y no pudimos identificar una interpretación alternativa que socavara los criterios de evaluación definidos. Además, basándonos en la distribución de las duraciones de las fijaciones, pudimos validar el diseño de los ítems identificando dos enfoques diferentes (sistemático y no sistemático). Mientras que en el enfoque sistemático sólo algunas AOIs eran más fijadas, en el enfoque no sistemático la distribución de las duraciones de las fijaciones era homogénea, lo que sugería que todas las AOIs se examinaban de forma similar.
- Por último, a partir del análisis de las interacciones producidas por los participantes a nivel de simulación, examinando el número de clics erróneos realizados y en que paso se producían, pudimos definir mejor el nivel de dificultad de ciertos ítems para que se ajustasen mejor al nivel de dificultad que les habíamos asignado inicialmente.

- PI-K5. *¿Qué consideraciones deben tenerse en cuenta para pasar de un diseño lineal a un diseño adaptativo en este tipo de sistemas?*

  - Primero, cabe mencionar que es recomendable el uso de simulaciones como las que hemos planteado en este estudio en las fases preliminares para analizar la viabilidad de este tipo de diseños, ya que permite la compilación inmediata de resultados y no provoca dificultades en los participantes que tendrían que realizar las pruebas.
  - Segundo, con un diseño adaptativo lo que buscamos es que para un participante determinado no se le presenten los ítems que le resultarían muy fáciles o difíciles, lo que reducirá la duración y el tiempo de ejecución de la prueba, mejorando su eficiencia y objetividad, y produciendo un resultado inmediato.
  - Realizamos una serie de consideraciones que deben tenerse en cuenta, en base a los resultados de nuestro trabajo, que pueden servir para diseñar herramientas similares:
    - El banco de ítems debidamente calibrado desempeña un papel fundamental en el TAI. Calibrar un banco de ítems es un proceso que requiere un gran esfuerzo, que depende directamente del modelo de la TRI que se quiera aplicar. Cuantos más parámetros tenga ese modelo, más respuestas se requerirán para poder llevar a cabo el proceso de calibración. En nuestro caso viendo la cantidad de respuestas que habíamos logrado en la fase de pruebas con usuarios finales, optamos

por utilizar el modelo de Rasch (1 parámetro) para llevar a cabo su calibración, por ser un modelo que da buenos resultados con muestras relativamente pequeñas.

- Dicho banco de ítems debería incluir ítems para todo el rango de niveles de dificultad que se desea evaluar, desde ítems fáciles a difíciles. De esta manera, será posible realizar una estimación precisa de toda la gama de niveles de habilidad, ya que los ítems fáciles son más informativos para los niveles de habilidad bajos y, por el contrario, los ítems difíciles son más informativos para los niveles de habilidad altos.
- Los test adaptativos por el mero hecho de contar con un banco de ítems calibrado y su naturaleza adaptativa, ofrecen una serie de ventajas no proporcionadas por los test lineales como, por ejemplo, incrementando el nivel de seguridad de las pruebas, ya que no todos los participantes van a ver los mismos ítems, dificultando que estos puedan resolverlas por reconocimiento y memoria.
- Además, es necesario comprobar si el banco de ítems tiene el suficiente tamaño para un número previsto de participantes, o si la longitud de la prueba, después de haber seleccionado un criterio de parada, es la apropiada para lograr el grado de fiabilidad buscado.
- Cabe destacar que en las simulaciones diseñadas no hemos impuesto restricciones para balancear los contenidos a nivel de CD y/o SC, ya que se tratan de simulaciones a pequeña escala con bancos de ítems reducidos. En futuros estudios deberíamos considerar la ampliación del número de ítems de los bancos y, así se podrían añadir restricciones más complejas para el diseño de la prueba, como por ejemplo distribuciones determinadas a nivel de SC. De esta forma, por ejemplo, podríamos configurar los test de IAD para que se mostrase a las personas que se examinan al menos un mínimo de preguntas similar por cada una de las 3 CD que engloba.
- Si el objetivo del TAI no es estimar las competencias de los participantes, sino clasificarlos en "aprobado" o "suspendido", como puede ser el caso de una evaluación orientada a la certificación como BAIT, sería conveniente establecer un punto de corte en la escala de rasgos latentes. Actualmente en BAIT, es requerido alcanzar al menos un 70% de la puntuación total del test para obtener la certificación. Si se pretende equiparar este requerimiento a un contexto TAI, será necesario realizar varios estudios adicionales para establecer el punto de corte en cada prueba.
- Si el objetivo es proporcionar a los participantes un diagnóstico al final de la prueba, los TAI proporcionan una información diagnóstica escasa. Tradicionalmente, los modelos utilizados que siguen un enfoque adaptativo han sido principalmente sumativos.
- Los resultados preliminares obtenidos de las simulaciones muestran que reduciendo al menos un 25% la longitud de las pruebas es posible

estimar satisfactoriamente el rasgo latente. Lo cual, supondrá una reducción del tiempo necesario para la realización de las pruebas de evaluación del nivel de CD de cualquier ciudadano. Este hecho, ayuda a reducir la fatiga y la frustración de los participantes durante la realización de las pruebas, si se ven obligados a responder muchos ítems. Además, no sólo las versiones adaptativas de ambas pruebas han reducido su tamaño considerablemente, sino que han ofrecido una mejora notable en la proporción de clasificaciones recibidas.

- Por último, nosotros en las pruebas optamos por una configuración ampliamente aceptada, como describimos en el capítulo 6, pero es recomendable seguir profundizando en la configuración más adecuada y que ofrezca mejores resultados, por ejemplo, aplicando diferentes reglas de parada con diferentes longitudes de pruebas, aplicando diferentes reglas para seleccionar el siguiente ítem o para estimar los niveles de habilidad, etc.
- Además, hay que tener presente la gestión del TAI en sí misma (que hemos pospuesto para futuras líneas de trabajo), que incluye el equilibrio de los contenidos, el análisis de los ítems, la puntuación de los ítems, el establecimiento de normas, el análisis de las prácticas y la actualización del banco de ítems. Con lo cual, llevar a cabo estudios de simulación pueden ayudar a evaluar las cuestiones previamente listadas, así como apoyar la toma de decisiones.

De acuerdo con todo esto, consideramos validada la hipótesis de esta tesis doctoral:

Es posible evaluar adecuadamente constructos cognitivos complejos dentro de la Competencia Digital mediante herramientas que empleen diferentes formatos dinámicos, el registro detallado de la interacción de los participantes y test adaptativos informatizados.

7.3. Limitaciones de la investigación

Esta investigación constituye un primer paso en un largo camino hacia el desarrollo de pruebas de evaluación adaptativas de CD. Durante su desarrollo, hemos detectado algunas limitaciones que pueden entenderse también como líneas futuras de investigación, tal y como describimos a continuación.

La primera limitación es debida al modo en el que tuvimos que llevar a cabo las pruebas con los usuarios finales. Inicialmente íbamos a llevarlas a cabo en sesiones tutorizadas en ordenadores con controles limitados. Es decir, en condiciones similares a cómo se llevan a cabo las pruebas de evaluación de IT Txartela y BAIT. De esta manera, los usuarios recibían una serie de instrucciones al inicio de la

prueba y a continuación llevaban a cabo la prueba con un navegador minimizado (sin barra de navegación) y optimizado a nivel de resolución de pantalla. Desafortunadamente la pandemia nos hizo modificar el plan establecido y tuvimos que optar por pruebas en abierto en entornos no controlados. La mayoría de los usuarios realizaron las pruebas desde sus casas con todo tipo de navegadores, resoluciones y sistemas operativos. Cabe destacar que las distintas resoluciones de los usuarios en sus ordenadores pudo ser un factor clave, ya que cuando estaba optimizada se mostraban las pruebas de manera que se aprovechase el máximo de la pantalla, facilitando su lectura sin requerir desplazamientos de pantalla. En los comentarios y sugerencias recibidos pudimos constatar este hecho. Afortunadamente, las incidencias fueron muy bajas y los usuarios pudieron seguir adelante con las pruebas. Marcamos la pauta de que la prueba debía llevarse a cabo de manera individual y sin acceder a Internet en otras pestañas del navegador para buscar información que les ayudase a responder a las preguntas correctamente. A pesar de esto, somos conscientes de que no pudimos garantizar el nivel de exigencia que podría asegurar resultados totalmente individuales y en un entorno controlado.

Otra limitación relacionada, tiene que ver con la variedad de la muestra de usuarios en nuestras pruebas. La manera en la que realizamos las invitaciones a participar en las pruebas, organizando un evento online durante la All Digital Week, no nos garantiza que hayamos podido contar con usuarios con niveles bajos de CD. Por tratarse de una prueba en línea, los usuarios que decidieron participar probablemente fueron usuarios que se sentían cómodos y seguros con este tipo de pruebas y requerimientos, y probablemente serían usuarios que no tenían niveles bajos de CD. Además, dadas las restricciones éticas y legales con respecto a la privacidad de los usuarios en las pruebas, no podemos asegurar que los usuarios introdujesen la información correctamente (género y rango de edad) a pesar de que las respuestas fuesen anónimas y no hubiese manera de relacionar esos datos con el usuario en cuestión.

En lo que respecta al formato de las pruebas, pudimos constatar que muchos usuarios decidieron abandonar prematuramente la realización de las pruebas. Probablemente una vez que vieron de que se trataba y que les iba a llevar cierto tiempo en realizar la prueba completa. Consecuentemente, tuvimos que desechar muchos resultados de pruebas, ya que consideramos como resultados de pruebas aceptables los que al menos habían respondido el 75% de las preguntas de la prueba. En pruebas supervisadas la ratio de pruebas desechadas probablemente disminuiría considerablemente. Por otra parte, hay que reconocer que, al realizar pruebas abiertas en línea, pudimos acceder a un mayor grupo de usuarios en un periodo corto de tiempo. Aunque el abandono y la desmotivación son elementos inevitables en cualquier proceso de aprendizaje como las pruebas llevadas a cabo, en el estudio realizado con tecnología de seguimiento ocular pudimos constatar que los usuarios cuando se enfrentaban a ítems con formatos dinámicos

(simulaciones interactivas, basados en imagen o simulación, etc.) mostraban un mayor interés y compromiso con la prueba, que cuando tenían que responder un ítem de opción múltiple, ampliamente utilizado en pruebas de evaluación.

7.4. Aplicaciones de la investigación

Como resultado de la investigación hemos realizado una serie de contribuciones científicas que desglosamos a continuación:

- En relación con el capítulo 3, se ha publicado el siguiente reporte técnico colaborando con el Joint Research Centre: Centeno, C., Vuorikari, R., Punie, Y., Kluzer, S., Vitorica, A., Lejarzegi, R., ... y Bartolomé, J. (2019). Developing digital competence for employability: Engaging and supporting stakeholders with the use of DigComp (No. JRC118711). Joint Research Centre (Seville site).

- En relación con el capítulo 3, se ha publicado el siguiente artículo de revista (indexada Emerging Sources Citation Index y Scopus): Bartolomé, J., Garaizar, P., y Larrucea, X. (2022). A pragmatic approach for evaluating and accrediting digital competence of digital profiles: A case study of entrepreneurs and remote workers. Technology, Knowledge and Learning, 27(3), 843-878.

- En relación con el capítulo 4, se ha publicado el siguiente artículo de revista (indexada WoS y Scopus), (Journal Citation Reports (JCR)): Bartolomé, J., y Garaizar, P. (2022). Design and Validation of a Novel Tool to Assess Citizens' Netiquette and Information and Data Literacy Using Interactive Simulations. Sustainability, 14(6), 3392.

- En relación con el capítulo 5, se ha enviado el siguiente artículo de revista (indexada WoS y Scopus), (Journal Citation Reports (JCR)) el 04/01/2023 a la revista PLOS ONE, que se encuentra pendiente de evaluación: Bartolomé, J., Garaizar, P., Loizaga, E. y Bastida, L. (2023). Using Eye-Tracking Data to Examine Response Processes in Digital Competence Assessment for Validation purposes.

También hemos realizado varias participaciones en congresos internacionales relacionadas con este trabajo:

- Bartolomé, J., de Soria, I. M., Jakobsone, M., Fernández, A., Ruseva, G., Koutoudis, P., ... y Vaquero, M. (2018, March). Developing a digital competence assessment and accreditation platform for digital profiles. In Proceedings of the 12th International Technology, Education and Development Conference (INTED), Valencia, Spain (pp. 5-7).

- Bartolomé, J., y Garaizar, P. (2018). Detecting confusing items in a digital skills certification system through assessment analytics. In INTED2018 Proceedings (pp. 4522-4528). IATED.

- Bartolomé, J., Garaizar, P., y Bastida, L. (2020, Octubre). Validating item response processes in digital competence assessment through eye-tracking techniques. In Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 738-746).

Creemos que desde los resultados de esta investigación hay tres líneas de aplicabilidad muy claras, que pueden tener en cuenta los desarrolladores de herramientas de evaluación de CD.

El primer campo de aplicación se refiere a la importancia de seguir una metodología que incluya diferentes tipos de estudios de validación durante el proceso de desarrollo la herramienta, con el objetivo de garantizar la calidad de las propiedades psicométricas de la prueba. Incorporar en este diseño el marco de referencia DigComp es crucial, y más cuando se pretende hablar de la CD como una competencia transversal presente en multitud de PD. La metodología seguida para el diseño de los PD y la elaboración del banco de ítems puede servir de orientación a desarrolladores de herramientas de evaluación de CD. Detallar las decisiones tomadas, así como los principios de diseño seguidos, pueden servir a los desarrolladores para que puedan evaluar si tales decisiones son aplicables a su contexto específico, o si necesitan llevar a cabo ajustes.

El segundo campo de aplicación es el desarrollo de nuevos formatos de ítems que ofrezcan nuevas posibilidades de evaluación de habilidades cognitivas de orden superior. Hemos mostrado como puede utilizarse la información registrada durante la realización de las pruebas con distintos fines. Por ejemplo, validando los RP de los participantes, comprobando que muestran los comportamientos esperados de acuerdo con los criterios de evaluación establecidos, o utilizando la información recogida para mejorar el diseño de los ítems, ajustando el nivel de dificultad según lo esperado inicialmente, etc. Esta información puede ser utilizada por desarrolladores y grupos de investigación interesados en trabajar en nuevos formatos de ítems que requieran de RP más complejos.

El tercer campo de aplicación es de la evaluación adaptativa de CD. Como hemos mostrado el estudio exploratorio de nuestra investigación, puede incorporarse este diseño a la evaluación de la CD. Es un proceso complejo que requiere de la realización de diversos estudios dependiendo de los objetivos que se tengan en mente. Nosotros hemos propuesto una serie de consideraciones que tratan de simplificar este diseño y hacerlo más abarcable. Las posibilidades que ofrece un diseño adaptativo son notables. Desarrolladores de pruebas de evaluación de CD pueden estar interesados en partir de las consideraciones identificadas en nuestro

trabajo con el objetivo de seguir avanzando en el diseño de su propia herramienta de acuerdo con sus necesidades.

7.5. Líneas futuras de trabajo

Tras los resultados de esta investigación, describimos a continuación una serie de líneas de trabajo futuras relacionadas con los objetivos de nuestro estudio.

Una de las ventajas de diseño que hemos adoptado para la herramienta de evaluación es que posibilita la integración de nuevos formatos de ítems. En concreto, sería muy interesante diseñar nuevos formatos de ítems que produzcan en los participantes RP más complejos y sirvan para evaluar habilidades cognitivas de orden superior como, por ejemplo, diseñando e integrando juegos interactivos diseñados a medida. Como hemos mencionado anteriormente, cada una de las 21 CD definidas en DigComp tienen sus propias peculiaridades y si queremos evaluarlas, tendremos que elegir los formatos de ítems que mejor se ajusten a los criterios de evaluación que establezcamos, para que los participantes puedan activar los comportamientos esperados. Por ejemplo, no le vamos a administrar el mismo tipo de ítems en una prueba de "Protección de dispositivos y contenidos digitales", en la que entre otras cosas le vamos a pedir tome las medidas necesarias para asegurar la fiabilidad y la privacidad que, en una prueba de "Programación", en la que le vamos a pedir, por ejemplo, que desarrolle secuencias de instrucciones para solucionar un problema dado, o ejecutar una tarea determinada.

Cabe destacar que a lo largo del trabajo realizado en la tesis hemos examinado una serie de métricas de los datos recogidos durante la realización de las pruebas por parte de los participantes, pero como futura línea de trabajo se podría considerar el estudio de nuevas métricas que no hayan sido exploradas aún como, por ejemplo, los desplazamientos del mouse durante la resolución de los ítems con formatos dinámicos. El análisis de nuevas métricas podría servir para entender mejor el comportamiento de los participantes durante la realización de las pruebas y ser usado con fines de validación o para ajustar el diseño de los ítems.

Otra línea de trabajo identificada está relacionada con las revisiones de las pruebas de evaluación, y que consideramos como parte del diseño de los ítems. Las personas que se examinen y no superen una prueba en BAIT, podrán revisar los ítems fallados. Llevarlo a cabo con ítems de opción múltiple es relativamente sencillo, mostrándoles su respuesta y la correcta. Pero con otros tipos de formatos ya no es tan inmediato. Por ejemplo, en una simulación interactiva podemos reproducir sus acciones y luego mostrarle el camino correcto. Pero ¿qué camino le mostramos cuando hay varios posibles? ¿Hay uno mejor que otro? En el estudio que llevamos a cabo con tecnología de seguimiento ocular, conseguimos crear una representación gráfica muy interesante para los ítems basados en una imagen o simulación, mostrando un modelo del orden de las áreas examinadas por los participantes que habían respondido correctamente ese ítem. Este modelo aparte

de ser utilizado para validar el criterio de evaluación del ítem de una manera muy sencilla, pretendemos incorporarlo en las revisiones para este formato de preguntas.

Por último, en la última parte de la tesis identificamos las consideraciones para tener en cuenta en el diseño de un sistema adaptativo de evaluación de CD. Una futura línea de trabajo es seguir profundizando en la configuración del TAI, aplicando diferentes reglas de parada con diferentes longitudes de pruebas, aplicando diferentes reglas para seleccionar el siguiente ítem o para estimar los niveles de habilidad, etc. Además, llevar a cabo estudios de simulación pueden ayudar a evaluar las distintas opciones y apoyar la toma de decisiones. Y en paralelo, implementar en la plataforma el motor adaptativo siguiendo las especificaciones identificadas en nuestro trabajo, con el objetivo de poder administrar pruebas adaptativas a usuarios finales. Más aún, existen otros aspectos como la gestión del TAI en sí misma, que no puede ser ignorada y hemos pospuesto para futuras líneas de trabajo.

7.6. Comentarios finales

Últimos días de febrero de 2023, y aquí estoy escribiendo los últimos textos de la tesis y dándole los últimos retoques. Pensé que este momento nunca iba a llegar. Cuando empecé, tuve claro que realizar una tesis doctoral sería duro por mis condiciones de padre de familia y trabajador a jornada completa. Lo que no me imaginé es que los vaivenes de la vida y sus improvistos, podrían hacer ese largo camino más duro todavía. Y por si fuese poco, se coló una pandemia y una guerra mundial encubierta nada más y nada menos. El esfuerzo ha merecido la pena. Tarde, pero he podido descubrir una parte de la ciencia que desconocía y que me tiene enganchado. Espero que esta tesis sea sólo la primera semilla de un nuevo y apasionante mundo por recorrer. Como dijo Diógenes de Sínope: "*Más vale tarde que nunca*".

No puedes parar las olas, pero puedes aprender a surfear

Sharif Ahnor

8.

Glosario y Abreviaturas

8.1. Glosario

AOI	Área de interés
API	Conjunto de reglas de comunicación entre aplicaciones (*Application Programming Interface*).
App	Programa instalable, especialmente para el caso de dispositivos móviles (abreviatura de *Application*).
BAIT	Sistema de Evaluación y Certificación de CDs del Gobierno Vasco. Accesible en la web https://bait.eus
CBA	Evaluación basada por ordenador (*Computer-based assessment*).
DigComp	Marco Europeo de Competencias Digitales para la Ciudadanía
DINA	*Deterministic Input, Noisy And*
DLGF	Marco Global de Alfabetización Digital
DSM	*Digital Single Market*
EAP	Estimación bayesiana Esperada a Posteriori
ECDL	*European Computer Driver License*
emoticono	Icono que transmite expresión facial y/o emocional (neologismo derivado del inglés *emoticon*, contracción de emoción e icono).

EntreComp	El marco de competencias del espíritu empresarial
ETCD	Herramienta online desarrollada por nuestro equipo de investigación. Accesible en la web http://evaluatucompetenciadigital.com
ESCO	*European multilingual classification of Skills, Competences, Qualifications and Occupations*
GDI	Índice de discriminación global
KL	*Kullback-Leibler*
feedback	Retroalimentación, devolución de información al usuario.
iBATIS	Framework desarrollado por *Apache Software Foundation*, que se ocupa de la capa de persistencia.
IT Txartela	Sistema de Certificación de Competencias básicas en Tecnologías de la Información del Gobierno Vasco. Accesible en la web https://it-txartela.net
Java	Lenguaje de programación (utilizado en ETCD)
JavaScript	Lenguaje de programación (utilizado en *Articulate Storyline*)
MAP	Estimación bayesiana Máxima A Posteriori
MIRT	Teoría Multidimensional de Respuesta al Ítem
MOOC	Cursos Online Masivos y Abiertos (*Massive Online Open Courses*)
MySQL	Sistema de gestión de bases de datos relacional desarrollado por *Oracle Corporation* y considerada como la base de datos de código abierto más popular del mundo.
ODS	Objetivos para el Desarrollo Sostenible. Accesible en la web https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/
OCDE	Cooperación y el Desarrollo Económico
O*NET	*Occupational Information Network*
open source	Código abierto, modelo de software basado en la colaboración abierta.
PIAAC	Programa para la Evaluación Internacional de las Competencias de los Adultos
PISA	Programa para la Evaluación Internacional de Estudiantes
P4E	Herramienta online desarrollada por nuestro equipo de investigación. Accesible en la web http://pathwaysforemploy.com
storytelling	Narrativa atrapante de sucesos, con un mensaje final que deja un aprendizaje o concepto
Struts	Struts es una herramienta de soporte para el desarrollo de aplicaciones Web del patrón MVC bajo la plataforma Java EE.
TAI	Test Adaptativo Informatizado

TEA	Evaluación Mediada por la Tecnología
TIC	Tecnologías de la Información y la Comunicación
TOI	Tiempo de interés
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

8.2. Abreviaturas utilizadas en el documento

AA	Analíticas de evaluación
AC	Área competencial
AFE	Análisis factorial exploratorio
AFC	Análisis factorial confirmatorio
CC	Componentes del conocimiento
CD	Competencia digital
CFI	Ajuste comparativo
CPS	Resolución de problemas complejos
DBR	Metodología de investigación basada en el diseño
DC	Diagnóstico cognitivo
DC-TAI	Pruebas adaptativas informatizadas de diagnóstico cognitivo
ET	Seguimiento ocular (*eye-tracking*)
IAD	Información y alfabetización digital
LA	Analíticas de aprendizaje
M	Media
ML	Máxima verosimilitud marginal
MLA	Analítica de aprendizaje multimodal
Mos	Movimientos oculares
MRCML	Modelo multidimensional random coefficient multinomial logit
S	Segundos
SC	Sub-Competencia

PD	Perfil Digital
SD	Desviación estándar
RMSEA	Error cuadrático medio de aproximación
RP	Proceso de respuesta
Satisf	Satisfacción
TCT	Teoría clásica de los test
TLI	Ajuste no normado
TME	Pruebas multietapa
TPL	Tobii Pro Lab
TRI	Teoría de la respuesta al ítem

9. Bibliografía

9.1. Referencias

Abad, F. J., Olea, J., Real, E. y Ponsoda, V. (2002). "Estimación de habilidad y precisión en tests adaptativos informatizados y tests óptimos: un caso práctico." Revista Electrónica de Metodología Aplicada 7(1): 1-20.

Abad, F. J., Olea, J., Ponsoda, V., y García, C. (2011). Medición en ciencias del comportamiento y de la salud. Madrid: Editorial Síntesis.

Abidin, A. Z., Istiyono, E., Fadilah, N., & Dwandaru, W. S. B. (2019). A Computerized Adaptive Test for Measuring the Physics Critical Thinking Skills. International Journal of Evaluation and Research in Education, 8(3), 376-383.

Abidoye, R., Lim, B. T. H., Lin, Y. C., & Ma, J. (2022). Equipping Property Graduates for the Digital Age. Sustainability, 14(2), 640.

Adams, R.J. (2005). Reliability as a measurement design effect. Stud. Educ. Eval. 31, 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008.

Adams, R.J. & Khoo, S.T. (1996). Quest; ACER: Melbourne, Australia.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. Applied psychological measurement, 21(1), 1-23.

Aesaert, K., van Nijlen, D., Vanderlinde, R., & van Braak, J. (2014). Direct measures of digital information processing and communication skills in primary education: Using item response theory for the development and validation of an ICT competence scale. Computers & Education, 76, 168-181. doi:10.1016/j.compedu.2014.03.013

Aesaert, K., Voogt, J., Kuiper, E., & van Braak, J. (2017). Accuracy and bias of ICT self-efficacy: An empirical study into students' over-and underestimation of their ICT competences. Computers in human behavior, 75, 92-102.

Akyar, Ö. Y., & Demirhan, G. (2022). Assessment of negotiation styles in higher education through a game-based assessment tool. Education and Information Technologies, 1-18.

Ala-Mutka, K. (2011). Mapping digital competence: Towards a conceptual understanding. Sevilla: Institute for Prospective Technological Studies, 7-60.

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. Computers & Education, 125, 413-428.

Alonso-Fernandez, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. Game Learning Analytics: Blending visual and data mining techniques to improve serious games and to better understand players learning. (lo publican en noviembre, es un draft)

ALRABABAH, S., Wu, H., & Molnár, G. (2022). Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity and Behaviour Pattern Analyses.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). Standards for educational and psychological testing. Amer Educational Research Assn.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME] (2014). Washington, DC: American Educational Research Association. Standards for educational and psychological testing.

Andrzejewska, M., & Stolińska, A. (2016). Comparing the difficulty of tasks using eye tracking combined with subjective and behavioural criteria. Journal of Eye Movement Research, 9(3). https://doi.org/10.16910/jemr.9.3.3

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record, 28(2), 49-60.

Arouri, Y. M., & Hamaidi, D. A. (2017). Undergraduate Students' Perspectives of the Extent of Practicing Netiquettes in a Jordanian Southern University. International Journal of Emerging Technologies in Learning, 12(3).

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. Medical teacher, 40(1), 62-69. https://doi.org/10.1080/0142159X.2017.1391373

Association of College & Research Libraries [ACRL] (2016). Framework for Information Literacy for Higher Education; American Library Association: Chicago, IL, USA, 2016; http://www.ala.org/acrl/standards/ilframework (accedido el 15 de octubre 2022).

Attali, Y. (2018, June). Automatic item generation unleashed: An evaluation of a large-scale deployment of item models. In International Conference on Artificial Intelligence in Education (pp. 17-29). Springer, Cham.

Aybek, E. C., y Demirtasli, R. N. (2017). Computerized Adaptive Test (CAT) Applications and Item Response Theory Models for Polytomous Items. International Journal of Research in Education and Science, 3(2), 475-487.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. Educational psychologist, 50(1), 84-94. https://doi.org/10.1080/00461520.2015.1004069

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. Computers in Human Behavior, 96, 207-210.

Bacigalupo, M., Kampylis, P., Punie, Y., & Van den Brande, G. (2016). EntreComp: The entrepreneurship competence framework. Luxembourg: Publication Office of the European Union, 10, 593884.

Badaracco, M., y Martínez, L. (2013). A fuzzy linguistic algorithm for adaptive test in Intelligent Tutoring System based on competences. Expert Systems with Applications, 40(8), 3073-3086.

Baek, C., & Doleck, T. (2021). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. Interactive Learning Environments, 1-23.

BAIT—Servicio de Evaluación y Certificación de Competencias Digitales. Disponible en línea: http://www.bait.eus (accedido el 17 de noviembre de 2022).

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques. CRC press.

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing Early At-Risk Factors in Higher Education E-Learning Courses. International Educational Data Mining Society.

Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications, 379-396.

Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. Guilford Publications.

Barnard, J. J. (2018). From simulation to implementation: two CAT case studies. Practical Assessment, Research, and Evaluation, 23(1), 14.

Barthakur, A., Kovanovic, V., Joksimovic, S., Siemens, G., Richey, M., & Dawson, S. (2021). Assessing program-level learning strategies in MOOCs. Computers in Human Behavior, 117, 106674.

Barthakur, A., Kovanovic, V., Joksimovic, S., Zhang, Z., Richey, M., & Pardo, A. (2022). Measuring leadership development in workplace learning using automated assessments: Learning analytics and measurement theory approach. British Journal of Educational Technology.

Bartolomé, J., Garaizar, P., & Bastida, L. (2020, October). Validating item response processes in digital competence assessment through eye-tracking techniques. In Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 738-746).

Bartolomé, J., Garaizar, P., & Larrucea, X. (2022). A pragmatic approach for evaluating and accrediting digital competence of digital profiles: A case study of entrepreneurs and remote workers. Technology, Knowledge and Learning, 27(3), 843-878.

Bartolomé, J., & Garaizar, P. (2022). Design and Validation of a Novel Tool to Assess Citizens' Netiquette and Information and Data Literacy Using Interactive Simulations. Sustainability, 14(6), 3392.

Barrios, M., & Cosculluela, A. (2013). Fiabilidad. Psicometría, 75-140.

Bashir, S., & Miyamoto, K. (2020). Digital Skills: Frameworks and Programs. World Bank.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: issues and practice, 17(3), 37-45.

Bayes, T. (1763). "Essay towards solving a problem in the doctrine of chances." Philosophical Transactions of the Royal Society of London 53: 370-418.

Beliaeva, T., Ferasso, M., Kraus, S., & Damke, E. J. (2019). Dynamics of digital entrepreneurship and the innovation ecosystem: A multilevel perspective. International Journal of Entrepreneurial Behavior & Research.

Benchoff, D. E., González, M. P., & Huapaya, C. R. (2018). Personalization of Tests for Formative Self-Assessment. IEEE Journal of Latin-American Learning Technologies, 13(2), 70–74.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automad scoring: It's not only the scoring. Educational Measurement: Issues and Practice, 17(4), 9–17

Bernardi, A. et al. (2019). On the Design and Development of an Assessment System with Adaptive Capabilities. In T. Di Mascio (Ed.), Methodologies and Intelligent Systems for Technology Enhanced Learning, Advances in Intelligent Systems and Computing (pp. 190–199). Cham: Springer.

Bi, R., Davison, R. M., & Smyrnios, K. X. (2017). E-business and fast growth SMEs. Small Business Economics, 48(3), 559-576.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In Assessment and teaching of 21st century skills (pp. 17-66). Springer, Dordrecht.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores. F. M. Lord y M. R. Novick. Reading (USA), Addison-Wesley: chapters 17-20.

Birnbaum, A. (1969). "A statistical theory for logistic mental test models with a prior distribution of ability." Journal of Mathematical Psychology 6: 258-276.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., & Ertl, T. (2017, December). Visualization of eye tracking data: A taxonomy and survey. In Computer Graphics Forum (Vol. 36, No. 8, pp. 260-284). https://doi.org/10.1111/cgf.13079

Bock, R. D. y Mislevy, R. J. (1982). "Adaptive EAP estimation of ability in a microcomputer environment." Applied Psychological Measurement 6: 431-444.

Boisvert, J. F., & Bruce, N. D. (2016). Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. Neurocomputing, 207, 653-668. https://doi.org/10.1016/j.neucom.2016.05.047

Brown, T. A. (2015). Confirmatory factor analysis for applied research. Guilford publications.

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. Cognitive

research: principles and implications, 4(1), 1-16. https://doi.org/10.1186/s41235-019-0159-2

Brusco, J. M. (2011). Know your netiquette. AORN journal, 94(3), 279-286.

Cabezas-González, M., Casillas-Martín, S., & García-Valcárcel Muñoz-Repiso, A. (2021). Basic education students' digital competence in the area of communication: The influence of online communication and the use of social networks. Sustainability, 13(8), 4442.

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. Studies in higher education, 36(4), 395-407.

Carter, M., & Egliston, B. (2021). What are the risks of virtual reality data? Learning analytics, algorithmic bias and a fantasy of perfect data. new media & society, 14614448211012794.

Cascio, W. F., & Montealegre, R. (2016). How technology is changing work and organizations. Annual review of organizational psychology and organizational behavior, 3(1), 349-375.

Castillo-Abdul, B., Romero-Rodríguez, L. M., & Larrea-Ayala, A. (2020). Kid influencers in Spain: understanding the themes they address and preteens' engagement with their YouTube channels. Heliyon, 6(9), e05056.

Catalano, A. (2015). The effect of a situated learning environment in a distance education information literacy course. The Journal of Academic Librarianship, 41(5), 653-659.

Catalano, A. J. (2016). Streamlining LIS Research: A Compendium of Tried and True Tests, Measurements, and Other Instruments: A Compendium of Tried and True Tests, Measurements, and Other Instruments. ABC-CLIO.

Cedefop, (2016). European guidelines for validating non-formal and informal learning, Publications Office. https://data.europa.eu/doi/10.2801/378817

Chalmers, P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. Journal of Statistical Software, 71(5), 1–38.

Chang, H.-H. y Ying, Z. (1996). "A global information approach to computerized adaptive testing." Applied Psychological Measurement 20(3): 213-229.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. International journal of Technology Enhanced learning, 4(5-6), 318-331.

Chen, J., y Wang, L. (2010). Computerized Adaptive Testing: A New Trend in Language Testing. In International Conference on Artificial Intelligence and Education (ICAIE) (pp. 725–728).

Cheng, Y. y Chang, H.-H. (2007). The modified maximum global discrimination index method for cognitive diagnostic computerized adaptive testing. 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, Minnesota (USA).

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. Psychometrika, 74(4), 619-632.

Cho, J. H., Xu, S., Hurley, P. M., Mackay, M., Benjamin, T., & Beaumont, M. (2019). Stram: Measuring the trustworthiness of computer-based systems. ACM Computing Surveys (CSUR), 51(6), 1-47.

Chrysafiadi, K., Troussas, C., & Virvou, M. (2020). Combination of fuzzy and cognitive theories for adaptive e-assessment. Expert Systems with Applications, 161, 113614.

Clifford, I., Kluzer, S., Troia, S., Jakobsone, M., & Zandbergs, U. (2020). DigCompSat. A Self-reflection Tool for the European Digital Framework for Citizens (No. JRC123226). Joint Research Centre (Seville site).

Colbert, A., Yee, N., & George, G. (2016). The digital workforce and the workplace of the future. Academy of management journal, 59(3), 731-739.

Comisión Europea. (2019). Erasmus+ Programme Guide. https://ec.europa.eu/programmes/erasmus-plus/resources/documents/treoirleabhar-erasmus-2019_en (accedido el 12 de octubre 2022)

Coskun, A., & Cagiltay, K. (2022). A systematic review of eye-tracking-based research on animated multimedia learning. Journal of Computer Assisted Learning, 38(2), 581-598.

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. Behavior research methods, 50(1), 362-379. https://doi.org/10.3758/s13428-017-0876-8

Cronbach, L. J. (1949). Essentials of psychological testing 2nd ed. Harper & brothers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334.

Cronbach, L. J. (1980). Validity on parole: How can we go straight. New directions for testing and measurement, 5(1), 99-108.

Cronbach, L. (1989). After Thirty Years. In Intelligence: Measurement, theory, and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys (p. 147). University of Illinois Press.

Cuhadar, I., & Binici, S. (2022). Modeling Slipping Effects in a Large-Scale Assessment with Innovative Item Formats. Educational Measurement: Issues and Practice, 41(3), 48-57.

Cúri, M., & Silva, V. (2019). Academic English proficiency assessment using a computerized adaptive test. TEMA (São Carlos), 20, 381-401.

David, H. J. J. O. E. P. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of economic perspectives, 29(3), 3-30.

Debuse, J. C., & Lawley, M. (2016). Benefits and drawbacks of computer-based assessment and feedback systems: Student and educator perspectives. British Journal of Educational Technology, 47(2), 294-301.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. Frontiers in psychology, 10, 102.

De La Torre, J. (2011). The generalized DINA model framework. Psychometrika, 76(2), 179-199.

De la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. Psychometrika, 81(2), 253-273.

de Macêdo, T. A. M., Cabral, E. L. D. S., Silva Castro, W. R., de Souza Junior, C. C., da Costa Junior, J. F., Pedrosa, F. M., ... & Másculo, F. S. (2020). Ergonomics and telework: A systematic review. Work, 66(4), 777-788.

Desmarais MC, Baker RS (2012) A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction 22(1-2):9–38

Dessart, L. (2017). Social media engagement: a model of antecedents and relational outcomes. Journal of Marketing Management, 33(5-6), 375-399.

del Carmen García Galera, M., Muñoz, C. F., & Pedrosa, L. P. (2017). Youth empowerment through social networks. Creating participative digital citizenship. Communication & Society, 30(3).

Dhaliwal, A., & Sahay, A. (2020). Factors influencing the success of women entrepreneurs in Emerging Markets: A Study of Indian women entrepreneurs. Journal of Asia Entrepreneurship and Sustainability, 16(2), 21–72

Dingel, J. I., & Neiman, B. (2020). How many jobs can be done at home?. Journal of Public Economics, 189, 104235.

Dolezalova, J., & Popelka, S. (2016). Scangraph: A novel scanpath comparison method using visualisation of graph cliques. Journal of Eye Movement Research, 9(4). https://doi.org/10.16910/jemr.9.4.5

Dormezil, S., Khoshgoftaar, T. M., & Robinson-Bryant, F. (2019). Differentiating between Educational Data Mining and Learning Analytics: A Bibliometric Approach. In EDM (Workshops) (pp. 17-22).

Dowell, N. M., & Poquet, O. (2021). SCIP: Combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments. Computers in Human Behavior, 119, 106709.

Drasgow, F. (2016). Technology and testing: Improving educational and psychological measurement (p. 376). Taylor & Francis.

Duchowski, A. T., & Duchowski, A. T. (2017). Eye tracking methodology: Theory and practice. Springer.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. Educational Researcher, 44(4), 237-251.

Dumas, D., Dong, Y., & McNeish, D. (2022). How fair is my test? A ratio coefficient to help represent consequential validity. European Journal of Psychological Assessment.

Durkin, M., McGowan, P., & McKeown, N. (2013). Exploring social media adoption in small to medium-sized enterprises in Ireland. Journal of Small Business and Enterprise Development.

D'Mello, S. (2017). Emotional learning analytics. Handbook of learning analytics, 115.

EC. (2015). A Digital Single Market Strategy for Europe—Analysis and Evidence—Commission Staff Working Document.

EC. (2016). A new skills agenda for Europe: Working together to strengthen human capital, employability and competitiveness.

EC. (2016b). European Commission. Digital Economy and Society Index (DESI). Retrieved October 28, 2019, from https:// ec. europa. eu/ digit al- single- market/ en/ desi.

EC. (2017b). The digital skills and jobs coalition. https://ec.europa.eu/digital-singlemarket/en/digital-skills-jobscoalition.

Eder, T. F., Scheiter, K., Richter, J., Keutel, C., & Hüttig, F. (2022). I see something you do not: Eye movement modelling examples do not improve anomaly detection in interpreting medical images. Journal of Computer Assisted Learning, 38(2), 379-391.

Eggen, T. J., van der Kleij, F. M., & Timmers, C. F. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. The Effectiveness of Methods for Providing Written Feedback Through a Computer-Based Assessment for Learning: a Systematic Review, 21-38.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. Psicologica, 32, 107–132.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Emerson, A., Min, W., Azevedo, R., & Lester, J. (2022). Early prediction of student knowledge in game-based learning with distributed representations of assessment questions. British Journal of Educational Technology.

Engelhardt, L., Goldhammer, F., Naumann, J., & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. Computers in Human Behavior. http://dx.doi.org/10.1016/j.chb.2017.02.020.

Eraslan, S., Yesilada, Y., & Harper, S. (2016). Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. Journal of Eye Movement Research, 9(1). https://doi.org/10.16910/jemr.9.1.2

Eraslan, S., Yaneva, V., Yesilada, Y., & Harper, S. (2019). Web users with autism: eye tracking evidence for differences. Behaviour & Information Technology, 38(7), 678-700. https://doi.org/10.1080/0144929X.2018.1551933

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). Validation of score meaning for the next generation of assessments: The use of response processes. Taylor & Francis. https://doi.org/0.4324/9781315708591-11

Ercikan, K., Seixas, P., Lyons-Thomas, J., & Gibson, L. (2015). Cognitive validity evidence for validating assessments of historical thinking. In New directions in assessing historical thinking (pp. 228-242). Routledge.

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. Educational Assessment, 25(3), 179-197. https://doi.org/10.1080/10627197.2020.1804353

Eurofound and the International Labour Office (2017), Working anytime, anywhere: The effects on the world of work, Publications Office of the European Union, Luxembourg, and the International Labour Office, Geneva.

European Commission. (2016). A new skills agenda for Europe: Working together to strengthen human capital, employability and competitiveness.

European Commission. (2017b). The digital skills and jobs coalition. https://ec.europa.eu/digital- singlemarket/en/digital-skills-jobscoalition.

European Parliament and the Council. (2006). Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning. Official Journal of the European Union, L394/310.

European Parliament and the Council. (2008b). Recommendation of the European Parliament and of the Council on the establishment of the European Qualifications Framework for lifelong learning. Official Journal of the European Union, C111/111.

Eurostat, 2017. Being Young in Europe Today–Digital World. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Being_young_in_Europe_today (accedido el 12 de octubre 2022).

Eurostat, 2021. Labour Force Survey. https://ec.europa.eu/eurostat/web/lfs/data/database (accedido el 14 de octubre 2022)

Eurobarometer, S. (2020). Public opinion in the European Union., Report, (93).

Evangelinos, G., & Holley, D. (2014, June). Developing a digital competence self-assessment toolkit for nursing students. In EDEN Conference Proceedings (No. 1, pp. 206-212).

Evans, D. S., & Schmalensee, R. (2016). Matchmakers: The new economics of multisided platforms. Harvard Business Review Press.

Fan, Y., van der Graaf, J., Lim, L., Raković, M., Singh, S., Kilgour, J., ... & Gašević, D. (2022). Towards investigating the validity of measurement of self-regulated learning based on trace data. Metacognition and Learning, 1-39.

Fang, Y., Huanrui, Z., & Mingrui, Y. (2017, August). Enhancing english vocabulary learning via computerized adaptive testing. In 2017 12th International Conference on Computer Science and Education (ICCSE) (pp. 3-5). IEEE.

Felstead, A., & Henseke, G. (2017). Assessing the growth of remote working and its consequences for effort, well-being and work-life balance. New Technology, Work and Employment, 32(3), 195–212

Ferguson, R., Hoel, T., Scheffel, M., & Drachsler, H. (2016). Guest editorial: Ethics and privacy in learning analytics. Journal of learning analytics, 3(1), 5-15.

Fernández-Macías, E., Hurley, J., & Bisello, M. (2016). What do Europeans do at work? A task-based analysis: European Jobs Monitor 2016.

Ferrari, A. (2012). Digital competence in practice: An analysis of frameworks. Sevilla: JRC IPTS, 10, 82116.

Ferrari, A., & Punie, Y. (2013). DIGCOMP: A framework for developing and understanding digital competence in Europe.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movements of aircraft pilots during instrument-landing approaches. Aeronautical engineering review.

Foo, S., Majid, S., & Chang, Y. K. (2017). Assessing information literacy skills among young information age students in Singapore. Aslib Journal of Information Management.

Fraillon, J. (2018). International large-scale computer-based studies on information technology literacy in education. Second handbook of information technology in primary and secondary education, 1161-1179.

Handel, M. J. (2016). The O* NET content model: strengths and limitations. Journal for Labour Market Research, 49(2), 157-176.

Forkosh Baruch, A., & Erstad, O. (2018). Upbringing in a digital world: Opportunities and possibilities. Technology, Knowledge and Learning, 23(3), 377-390.

Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. Assessment & Evaluation in Higher Education, 45(6), 842-854.

Gane, B. D., Israel, M., Elagha, N., Yan, W., Luo, F., & Pellegrino, J. W. (2021). Design and validation of learning trajectory-based assessments for computational thinking in upper elementary grades. Computer Science Education, 31(2), 141-168.

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. User Modeling and User-Adapted Interaction, 28(2), 127-203.

Gardner, J., Brooks, C., & Baker, R. (2019, March). Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th international conference on learning analytics & knowledge (pp. 225-234).

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., y Frank, E. (2004). Computerized adaptive measurement of depression: a simulation study. BMC psychiatry, 4(1), 1-11.

Garrett, R. K., & Danziger, J. N. (2007). Which telework? Defining and testing a taxonomy of technology-mediated work at a distance. Social Science Computer Review, 25(1), 27-47.

Garrison, W. M., & Baumgarten, B. S. (1986). An Application of Computer Adaptive Testing with Communication Handicapped Examinees. Educational and Psychological Measurement, 46(1), 23–35. https://doi.org/10.1177/0013164486461003

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64-71.

Gašević, D., Greiff, S., & Shaffer, D. W. (2022). Towards Strengthening Links between Learning Analytics and Assessment: Challenges and Potentials of a Promising New Bond. Computers in Human Behavior, 107304.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. Journal of Technology, Learning, and Assessment, 5(8), n8.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. Annual review of clinical psychology, 12, 83-104.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). Automatic item generation: Theory and practice. Routledge.

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. Medical education, 48(10), 950-962.

Gierl, M. J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. Applied psychological measurement, 42(1), 42-57.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. Applied psychological measurement, 14(1), 33-43.

Gilbert, N. (2019). 20 Best Office Software Solutions in 2019. Finances. https://financesonline.com/office-software/ (accedido el 12 de octubre 2022).

Glady, Y., Thibaut, J. P., & French, B. (2013). Visual strategies in analogical reasoning development: A new method for classifying scanpaths. In Proceedings of the annual meeting of the cognitive science society (Vol. 35, No. 35).

Global Entrepreneurship Monitor. (2018). https://www.gemconsortium.org/report/50012 (accedido el 13 de octubre 2022).

GMSA (2022). The Mobile Economy Europe 2022. https://www.gsma.com/mobileeconomy/wp-content/uploads/2022/10/051022-Mobile-Economy-Europe-2022.pdf (accedido el 12 de octubre 2022).

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002, March). Eye tracking in web search tasks: design implications. In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 51-58). https://doi.org/10.1145/507072.507082

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC.

Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. Computers in Human Behavior, 76, 607-616.

Graf, D., Oppl, S. and Eckmaier, A. (2017). Towards BPM Skill Assessment using Computerized Adaptive Testing. Proceedings of 9th ACM Conference on Subject-oriented Business Process Management, Darmstadt, Germany.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational measurement, 21(4), 347-360.

Greiff, S., Holt, D., & Funke, J. (2013). Perspectives on problem solving in cognitive research and educational assessment: analytical, interactive, and collaborative problem solving. Journal of Problem Solving, 5, 71-91.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. Computers & Education, 91, 92-105. https://doi.org/10.1016/j.compedu.2015.10.018

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. Computers in Human Behavior, 61, 36-46. https://doi.org/10.1016/j.chb.2016.02.095

Groner, R., Walder, F., & Groner, M. (1984). Looking at faces: Local and global aspects of scanpaths. In Advances in psychology (Vol. 22, pp. 523-533). North-Holland. https://doi.org/10.1016/S0166-4115(08)61874-9)

Grover, S., & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. Educational researcher, 42(1), 38-43.

Guillén-Gámez, F. D., Mayorga-Fernández, M., & Álvarez-García, F. J. (2020). A study on the actual use of digital competence in the practicum of education degree. Technology, Knowledge and Learning, 25(3), 667-684.

Haddon, L., & Brynin, M. (2005). The character of telework and the characteristics of teleworkers. New Technology, Work and Employment, 20(1), 34-46.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education, 15(3), 309-333.

Halszka, J., Holmqvist, K., & Gruber, H. (2017). Eye tracking in Educational Science: Theoretical frameworks and research agendas. Journal of eye movement research, 10(1). https://doi.org/10.16910/jemr.10.1.3

Hambleton, R. K. y Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. New horizons in testing. D. J. Weiss. New York (USA), Academic Press: 31-49.

Hambleton, R. K., y Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational measurement: issues and practice, 12(3), 38-47.

Hambleton, R. K. y Swaminathan, H. (1985). Item response theory: principles and applications. Boston (USA), Kluwer-Nijhoff Publishing.

Hambleton, R. K., & Swaminathan, H. (2013). Item response theory: Principles and applications. Springer Science & Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). Sage.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. (1991). Computerized adaptive testing: Theory, applications, and standards. In Advances in educational and psychological testing: Theory and applications (pp. 341-366). Springer, Dordrecht.

Han, K. C. T. (2018). Components of the item selection algorithm in computerized adaptive testing. Journal of Educational Evaluation for Health Professions, 15.

Handel, M. J. (2016). The O* NET content model: strengths and limitations. Journal for Labour Market Research, 49(2), 157-176.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. System, 33(1), 107-133.

Hair, N., Wetsch, L. R., Hull, C. E., Perotti, V., & Hung, Y. T. C. (2012). Market orientation in digital entrepreneurship: advantages and challenges in a Web 2.0 networked world. International Journal of Innovation and Technology Management, 9(06), 1250045.

Hammond, L., & Moseley, K. (2018). Reeling in proper "netiquette". Nursing made Incredibly Easy, 16(2), 50-53.

Harmer, B. M., & Pauleen, D. J. (2012). Attitude, aptitude, ability and autonomy: The emergence of 'offroaders', a special class of nomadic worker. Behaviour & Information Technology, 31(5), 439-451.

Hayter, C. S., Lubynsky, R., & Maroulis, S. (2017). Who is the academic entrepreneur? The role of graduate students in the development of university spinoffs. The Journal of Technology Transfer, 42(6), 1237-1254.

Heer, R. (2012). A model of learning objectives—based on A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Center for Excellence in Learning and Teaching, Iowa State University.

Heifetz, R. A., & Laurie, D. L. (1997). The work of leadership. Harvard business review, 75, 124-134.

Helfaya, A., & O'Neill, J. (2018). Using computer-based assessment and feedback: Meeting the needs of digital natives in the digital age. International Journal of Teacher Education and Professional Development (IJTEPD), 1(2), 46-71.

Helfaya, A. (2019). Assessing the use of computer-based assessment-feedback in teaching digital accountants. Accounting Education, 28(1), 69-99.

Herrington, J., McKenney, S., Reeves, T., & Oliver, R. (2007, June). Design-based research and doctoral students: Guidelines for preparing a dissertation proposal. In EdMedia+ Innovate Learning (pp. 4089-4097). Association for the Advancement of Computing in Education (AACE).

Hollis, H. (2018). Information literacy as a measurable construct: A need for more freely available, validated and wide ranging instruments. Journal of Information Literacy, 12(2).

Holsanova, J., Rahm, H., & Holmqvist, K. (2006). Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. Visual communication, 5(1), 65-93. https://doi-org/10.1177/1470357206061005

Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. Applied Psychological Measurement, 37(7), 563-582.

Hu, Y., Wu, B., & Gu, X. (2017). An eye tracking study of high-and low-performing students in solving interactive and analytical problems. Journal of Educational Technology & Society, 20(4), 300-311. https://doi.org/10.1037/a0020082

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In Understanding and investigating response processes in validation research (pp. 1-12). Springer, Cham. https://doi.org/10.1007/978-3-319-56129-5_1

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. Practical Assessment, Research, and Evaluation, 15(1), 3.

Hull, C. E. K., Hung, Y. T. C., Hair, N., Perotti, V., & DeMartino, R. (2007). Taking advantage of digital opportunities: a typology of digital entrepreneurship. International Journal of Networking and Virtual Organisations, 4(3), 290-303.

Hunt, T., & Jordan, S. (2016). I wish I could believe you: the frustrating unreliability of some assessment research. Practitioner Research in Higher Education, 10(1), 13-21.

Ifenthaler, D., & Schumacher, C. (2016). Student perceptions of privacy principles for learning analytics. Educational Technology Research and Development, 64(5), 923-938.

Ifenthaler, D., & Greiff, S. (2021). Leveraging learning analytics for assessment and feedback. In J. Liebowitz (Ed.), Online learning analytics (pp. 1–18). CRC Press.

Ifenthaler, D., Schumacher, C., & Kuzilek, J. (2022). Investigating students' use of self-assessments in higher education using learning analytics. Journal of Computer Assisted Learning.

Iglesias-Rodríguez, A., Hernández-Martín, A., Martín-González, Y., & Herráez-Corredera, P. (2021). Design, Validation and Implementation of a Questionnaire to

Assess Teenagers' Digital Competence in the Area of Communication in Digital Environments. Sustainability, 13(12), 6733.

Inal, Y. (2016). User-friendly locations of error messages in web forms: An eye tracking study. Journal of eye movement research, 9(5). https://doi.org/10.16910/jemr.9.5.1

Iñiguez-Berrozpe, T., & Boeren, E. (2020). Twenty-first century skills for all: Adults and problem solving in technology rich environments. Technology, Knowledge and Learning, 25(4), 929-951.

Israel-Fishelson, R., Hershkovitz, A., Eguíluz, A., Garaizar, P., & Guenaga, M. (2021). A log-based analysis of the associations between creativity and computational thinking. Journal of Educational Computing Research, 59(5), 926-959.

Istiyono, E., Dwandaru, W. S. B., y Faizah, R. (2018). Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT. REID (Research and Evaluation in Education), 4(2), 144-154.

Janssen, J., Stoyanov, S., Ferrari, A., Punie, Y., Pannekeet, K., & Sloep, P. (2013). Experts' views on digital competence: Commonalities and differences. Computers & Education, 68, 473-481.

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples. Instructional Science, 40(5), 813-827. https://doi.org/10.1007/s11251-012-9218-5

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. American Behavioral Scientist, 65(2), 371-388.

Jordan, S. (2014). Using e-assessment to learn about students and learning. International Journal of e-Assessment, 4(1).

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. The Internet and Higher Education, 33(4), 74-85.

Jovanović, J., Saqr, M., Joksimović, S., & Gašević, D. (2021). Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. Computers & Education, 172, 104251.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. Cognitive psychology, 8(4), 441-480. https://doi.org/10.1016/0010-0285(76)90015-3

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. Psychological review, 87(4), 329. https://doi.org/10.1037/0033-295X.87.4.329

Just, M. A., & Carpenter, P. A. (2018). Using eye fixations to study reading comprehension. In New methods in reading comprehension research (pp. 151-182). Routledge.

Kanan, C., Ray, N. A., Bseiso, D. N., Hsiao, J. H., & Cottrell, G. W. (2014, March). Predicting an observer's task using multi-fixation pattern analysis. In Proceedings of

the symposium on eye tracking research and applications (pp. 287-290). https://doi.org/10.1145/2578153.2578208

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In Validation of score meaning for the next generation of assessments (pp. 11-24). Routledge.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. Applied psychological measurement, 39(3), 167-188.

Kato, P. M., & de Klerk, S. (2017). Serious games for assessment: Welcome to the jungle. Journal of Applied Testing Technology, 18(S1), 1-6.

King, A. J., Bol, N., Cummins, R. G., & John, K. K. (2019). Improving visual behavior research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. Communication Methods and Measures, 13(3), 149-177. https://doi.org/10.1080/19312458.2018.1558194

Kingsbury, G. G. y Zara, A. R. (1989). "Procedures for selecting items for computerized adaptive tests." Applied Measurement in Education 2(4): 359-375.

Kingsbury, C. G., y Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. Applied measurement in education, 4(3), 241-261.

Kingsbury, G. G. y Houser, R. L. (1993). "Assessing the utility of item response models in computerized adaptive systems." Educational Measurement: Issues and Practice 12(1): 21-27.

Kitto, K., Cross, S., Waters, Z., & Lupton, M. (2015, March). Learning analytics beyond the LMS: the connected learning analytics toolkit. In Proceedings of the fifth international conference on learning analytics and knowledge (pp. 11-15).

Kluzer, S., & Priego, L. P. (2018). Digcomp into action: Get inspired, make it happen. a user guide to the european digital competence framework (No. JRC110624). Joint Research Centre (Seville site).

Knight, S., Buckingham Shum, S., & Littleton, K. (2013). In Epistemology, Pedagogy, assessment and learning analytics. Proceedings of the third International Conference on learning analytics and knowledge (pp. 75–84). https://doi.org/10.1145/2460296.2460312

Kok, E., Hormann, O., Rou, J., van Saase, E., van der Schaaf, M., Kester, L., & van Gog, T. (2022). Re-viewing performance: Showing eye-tracking data as feedback to improve performance monitoring in a complex visual task. Journal of Computer Assisted Learning.

Kopecky, K., Szotkowski, R., Aznar-Díaz, I., & Romero-Rodríguez, J. M. (2020). The phenomenon of sharenting and its risks in the online environment. Experiences from Czech Republic and Spain. Children and Youth Services Review, 110, 104812.

Kozík, T., & Slivová, J. (2014). Netiquette in electronic communication. International Journal of Engineering Pedagogy (iJEP), 4(3), 67-70.

Krathwohl, D.R. (2002) A revision of Bloom's taxonomy: An overview. Theory Pract. 41, 212–218. https://doi.org/10.1207/s15430421tip4104_2.

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. Behaviormetrika, 45(2), 527-563. https://doi.org/10.1007/s41237-018-0063-y

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of personality and social psychology, 77(6), 1121.

Kucharský, Š., Visser, I., Truțescu, G. O., Laurence, P. G., Zaharieva, M., & Raijmakers, M. E. (2020). Cognitive strategies revealed by clustering eye movement transitions. Journal of Eye Movement Research, 13(1). https://doi.org/10.16910/jemr.13.1.1

Kuo, C. Y., & Wu, H. K. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. Computers & Education, 68, 388–403

Kübler, T. C., Rothe, C., Schiefer, U., Rosenstiel, W., & Kasneci, E. (2017). SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies. Behavior research methods, 49(3), 1048-1064. https://doi.org/10.3758/s13428-016-0765-6

Laanpere, M. (2019). Recommendations on Assessment Tools for Monitoring Digital Literacy within UNESCO's Digital Literacy Global Framework. Information Paper No, 56. Disponible en línea: https://unesdoc.unesco.org/ark:/48223/pf0000366740 (accedido el 11 de Octubre de 2022).

Lan AS, Waters AE, Studer C, Baraniuk RG (2014) Sparse factor analysis for learning and content analytics. The Journal of Machine Learning Research 15(1):1959–2008

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). Handbook of test development (Vol. 2, pp. 3-18). New York, NY: Routledge.

Lang, C., Siemens, G., Wise, A., Gašević, D., & Merceron, A. (Eds.). (2022). Handbook of learning analytics (Second). Society for Learning Analytics and Research.

Larrondo Ureta, A., Peña Fernández, S., & Agirreazkuenaga Onaindia, I. (2020). Hacia una mayor participación de la audiencia: experiencias transmedia para jóvenes.

Law, N., Woo, D., & Wong, G. (2018). A global framework of reference on digital literacy skills for indicator 4.4. 2 (No. 51, p. 146). UNESCO.

Lee, Y. H., Hao, J., Man, K., & Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. Frontiers in psychology, 10, 906.

Leichner, N., Peter, J., Mayer, A. K., & Krampen, G. (2013). Assessing information literacy among German psychology students. Reference Services Review.

Lehmann, E. L., & Casella, G. (2006). Theory of point estimation. Springer Science & Business Media.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. Journal of educational measurement, 41(3), 205-237.

León, J. A., Moreno, J. D., Escudero, I., & Kaakinen, J. K. (2019). Selective attention to question-relevant text information precedes high-quality summaries: Evidence from eye movements. Journal of Eye Movement Research, 12(1). https://doi.org/10.16910/jemr.12.1.6

Lewandowski, D., & Kammerer, Y. (2021). Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research. Behaviour & Information Technology, 40(14), 1485-1515. https://doi.org/10.1080/0144929X.2020.1761450

Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: has it lived up to its promise and, if not, what would it take to do so. In Understanding and investigating response processes in validation research (pp. 159-177). Springer, Cham. https://doi.org/10.1007/978-3-319-56129-5_9

Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In Elements of adaptive testing (pp. 3-30). Springer, New York, NY.

Linden, W. J., van der Linden, W. J., & Glas, C. A. (Eds.). (2000). Computerized adaptive testing: Theory and practice. Springer Science & Business Media.

Linek, S. B., & Ostermaier-Grabow, A. (2018). Netiquette between students and their lecturers on Facebook: Injunctive and descriptive social norms. Social Media+ Society, 4(3), 2056305118789629.

Ling, G., Attali, Y., Finn, B., y Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test?. Applied psychological measurement, 41(7), 495-511.

List, A., Brante, E. W., & Klee, H. L. (2020). A framework of pre-service teachers' conceptions about digital literacy: Comparing the United States and Sweden. Computers & Education, 148, 103788.

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. Journal of Experimental Psychology: Applied, 16(3), 251.

Litchfield, B. C., & Dempsey, J. V. (2015). Authentic assessment of knowledge, skills, and attitudes. New Directions for Teaching and Learning, 142(142), 65-80.

Littlejohn, A. (2021). Professional Learning Analytics. Society for Learning Analytics and Research.

Liu, M., Kitto, K., & Shum, S. B. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. Computers in Human Behavior, 120, 106733.

Liu, T., & Israel, M. (2022). Uncovering students' problem-solving processes in game-based learning environments. Computers & Education, 182, 104462.

Lord, F. (1952). A theory of test scores. Psychometric monographs.

Lord, F. M. (1971a). Robbins–Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 2–31.

Lord, F. M. (1974). Practical methods for redesigning a homogeneous test, also for designing a multilevel test. ETS Research Bulletin Series, 1974(1), i-26.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, New Jersey (USA), Lawrence Erlbaum Associates.

Lord, F. M. (1983). Small N justifies Rasch model. In New horizons in testing (pp. 51-61). Academic Press.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. Journal of Educational Measurement, 21(3), 239-243.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 157-162.

Lord, F. M. (2012). Applications of item response theory to practical testing problems. Routledge.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. Applied Psychological Measurement, 22(3), 224-236.

Luecht, R. (2014). Design and implementation of large-scale multistage testing systems. Computerized multistage testing: Theory and applications, 69-83.

Lund, H. (2016). Eye tracking in library and information science: a literature review. Library Hi Tech. https://doi.org/10.1108/LHT-07-2016-0085

Lunz, M. E. y Bergstrom, B. A. (1994). "An empirical study of computer adaptive test administration formats." Journal of Educational Measurement 31(3): 251-263.

Lynch, D., & Howlin, C. P. (2014, November). Real world usage of an adaptive testing algorithm to uncover latent knowledge. In 7th international conference of education, research and innovation (ICERI2014 proceedings). IATED, Seville, Spain (pp. 504-511).

Ma, W. (2019). A diagnostic tree model for polytomous responses with multiple strategies. British Journal of Mathematical and Statistical Psychology, 72(1), 61-82.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. Journal of Statistical Software, 76, 1-19.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. Journal of Statistical Software, 48(8), 1–31.

Magis, D., Yan, D., & Von Davier, A. A. (2017). Computerized adaptive and multistage testing with R: Using packages catR and mstR. Springer.

Mahler, J. (2012). The telework divide: Managerial and personnel challenges of telework. Review of Public Personnel Administration, 32(4), 407-418.

Mahmood, K. (2016). Do people overestimate their information literacy skills? A systematic review of empirical evidence on the Dunning-Kruger effect. Communications in Information Literacy, 10(2), 3.

Mahmood, K. (2017). A systematic review of evidence on psychometric properties of information literacy tests. Library Review.

Makransky, G., Mayer, R., Nøremølle, A., Cordoba, A. L., Wandall, J., & Bonde, M. (2020). Investigating the feasibility of using assessment and explanatory feedback in desktop virtual reality simulations. Educational Technology Research and Development, 68(1), 293-317.

Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? The influence of experience and training on searching for chest nodules. Radiography, 12(2), 134-142. https://doi.org/10.1016/j.radi.2005.02.003

Markowski, B., McCartin, L., & Evers, S. (2018). Meeting students where they are: Using rubric-based assessment to modify an information literacy curriculum. Communications in Information Literacy, 12(2), 5.

Marsh, E. (2021). Understanding the effect of digital literacy on employees' digital workplace continuance intentions and individual performance. In Research anthology on digital transformation, organizational change, and the impact of remote work (pp. 1638-1659). IGI Global.

Martínez-Maldonado, R., Yan, L., Deppeler, J., Phillips, M., Gašević, D. (2022). Classroom Analytics: Telling Stories About Learning Spaces Using Sensor Data. In: Gil, E., Mor, Y., Dimitriadis, Y., Köppe, C. (eds) Hybrid Learning Spaces. Understanding Teaching-Learning Practice. Springer, Cham. https://doi.org/10.1007/978-3-030-88520-5_11

Masip, P., Suau, J., & Ruiz-Caballero, C. (2020). Perceptions on media and disinformation: Ideology and polarization in the Spanish media system. Profesional de la Información, 29(5).

Mason, L., Scheiter, K., & Tornatora, M. C. (2017). Using eye movements to model the sequence of text–picture processing for multimedia comprehension. Journal of Computer Assisted Learning, 33(5), 443-460.

Matcha, W., Gasevic, D., Uzir, N. A. A., Jovanovic, J., Pardo, A., Lim, L., ... & Tsai, Y. S. (2020). Analytics of Learning Strategies: Role of Course Design and Delivery Modality. Journal of Learning Analytics, 7(2), 45-71.

Mattar, J., Ramos, D. K., & Lucas, M. R. (2022). DigComp-based digital competence assessment tools: literature review and instrument analysis. Education and Information Technologies, 27(8), 10843-10867.

Mayrath, M. C., Clarke-Midura, J., & Robinson, D. H. (2012). Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern. Information Age Publishing: Charlotte, NC, USA.

McArthur, J. (2022). Rethinking authentic assessment: work, well-being, and society. Higher education, 1-17.

McCallum, E., Weicht, R., McMullan, L., & Price, A. (2018). EntreComp into action-Get inspired, make it happen: A user guide to the European Entrepreneurship Competence Framework (No. JRC109128). Joint Research Centre (Seville site).

McIntyre, D. P., & Srinivasan, A. (2017). Networks, platforms, and strategy: Emerging views and next steps. Strategic management journal, 38(1), 141-160.

McKenney, S., & Reeves, T. C. (2018). Conducting educational design research. Routledge. https://doi.org/10.4324/9781315105642.

McMurdo, G. (1995). Netiquettes for networkers. Journal of information science, 21(4), 305-318.

McNamara, B. R., & Nulsen, P. E. J. (2012). Mechanical feedback from active galactic nuclei in galaxies, groups and clusters. New Journal of Physics, 14(5), 055023.

Messick, S. (1987). Validity. ETS research report series, 1987(2), i-208.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational researcher, 18(2), 5-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. Educational Measurement: Issues and Practice, 14(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Millán, J. M., Lyalkov, S., Burke, A., Millán, A., & van Stel, A. (2021). 'Digital divide'among European entrepreneurs: Which types benefit most from ICT implementation?. Journal of Business Research, 125, 533-547.

Milligan, S., & Griffin, P. (2016). Understanding learning and learning design in MOOCs: A measurement-based interpretation. Journal of Learning Analytics, 3(2), 88–115. https://doi.org/10.18608/jla.2016.32.5

Milligan, S. (2020). Standards for developing assessments of learning using process data. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), Re-imagining University assessment in a digital World (pp. 179–192). Springer International Publishing. https://doi.org/10.1007/978-3-030-41956-1_13.

Millman, J., & Arter, J. A. (1984). Issues in item banking. Journal of Educational Measurement, 21(4), 315-330.

Misiejuk, K., Wasson, B., & Egelandsdal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. Computers in human behavior, 117, 106658.

Misiejuk, K., & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. Computers & Education, 175, 104319.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of educational statistics, 11(1), 3-31.

Mohammadyari, S., & Singh, H. (2015). Understanding the effect of e-learning on individual performance: The role of digital literacy. Computers & Education, 82, 11-25.

Molnár, G., Alrababah, S. A., & Greiff, S. (2022). How we explore, interpret, and solve complex problems: A cross-national study of problem-solving processes. Heliyon, 8(1), e08775.

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., y Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. Psychiatry research, 263, 250-256.

Mueller, R.O. y Knapp, T.R. (2018) Reliability and validity. In The Reviewer's Guide to Quantitative Methods in the Social Sciences. Routledge: London, UK. pp. 397–401.

Muñiz Fernández, J. (1997) Introducción a la teoría de respuesta a los ítems. Pirámide.

Muñiz, J. (2018). Introducción a la Psicometría: Teoría clásica y TRI. Pirámide.

Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. Computers & Education, 141, 103618.

Nambisan, S. (2017). Digital entrepreneurship: Toward a digital technology perspective of entrepreneurship. Entrepreneurship theory and practice, 41(6), 1029-1055.

Nambisan, S., Siegel, D., & Kenney, M. (2018). On open innovation, platforms, and entrepreneurship. Strategic Entrepreneurship Journal, 12(3), 354-368.

Nering, M. L., & Ostini, R. (Eds.). (2011). Handbook of polytomous item response theory models. Taylor & Francis.

Newman, D. (2016). Is mobility the answer to better employee productivity? Forbes. https://www.forbes.com/sites/danielnewman/2016/03/29/is-mobility-the-answer-to-better-employee-productivity/#7075a 0ce13 1c (accedido el 14 de octubre 2022).

Newton, P. E. (2019). What is response process validation evidence and how important is it? An essay reviewing Ercikan and Pellegrino (2017) and Zumbo and Hubley (2017). DOI: https://doi.org/10.1080/0969594X.2018.1447909

Nguyen, Q., Rienties, B., Toetenel, L., Ferguson, R., & Whitelock, D. (2017). Examining the designs of computer-based assessment and its impact on student engagement, satisfaction, and pass rates. Computers in Human Behavior, 76, 703-714. https://doi.org/10.1016/j.chb.2017.03.028

Nguyen, Q., Rienties, B., & Richardson, J. T. (2020). Learning analytics to uncover inequality in behavioural engagement and academic attainment in a distance learning setting. Assessment & Evaluation in Higher Education, 45(4), 594-606.

Nicolay, B., Krieger, F., Stadler, M., Gobert, J., & Greiff, S. (2021). Lost in transition–Learning analytics on the transfer from knowledge acquisition to knowledge application in complex problem solving. Computers in Human Behavior, 115, 106594.

Nisiforou, E. A., & Laghos, A. (2013). Do the eyes have it? Using eye tracking to assess students cognitive dimensions. Educational Media International, 50(4), 247-265. https://doi.org/10.1080/09523987.2013.862363

Oberländer, M., Beinicke, A., & Bipp, T. (2020). Digital competencies: A review of the literature and applications in the workplace. Computers & Education, 146, 103752.

Ochoa, X. (2019). Learning analytics in Latin America present an opportunity not to be missed. Nature human behaviour, 3(1), 6-7.

OECD. (2013). PISA 2012 assessment and analytical framework - Mathematics, reading, science, problem solving and financial literacy. Paris: OECD Publishing.

OECD. (2016). Skills for a digital world: Background Paper for Ministerial Panel 4.2 - DSTI/ICCP/IIS (2015)10/FINAL. Paris, France: OECD/Directorate for Science, Technology, and Innovation/Committee on Digital Economy Policy/Working Party on Measurement and Analysis of the Digital Economy.

Oggero, N., Rossi, M. C., & Ughetto, E. (2020). Entrepreneurial spirits in women and men. The role of financial literacy and digital skills. Small Business Economics, 55(2), 313-327.

Olea, J. y Ponsoda, V. (1996). Test adaptativos informatizados. Psicometría. J. Muñiz. Madrid (España), Universitas: 729-783.

Olsson, A. K., & Bernhard, I. (2020). Keeping up the pace of digitalization in small businesses–Women entrepreneurs' knowledge and use of social media. International Journal of Entrepreneurial Behavior & Research.

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. International Journal of Educational Technology in Higher Education, 14(2), 1–21.

Oranje A., Gorin J., Jia Y., & Kerr D. (2017). Collecting, Analyzing, and Interpreting Response Time, Eye-Tracking, and Log Data. 1st ed. Validation of score meaning for the next generation of assessments, Routledge, 39-51. https://doi.org/10.4324/9781315708591-4

Orso, D., Federici, N., Copetti, R., Vetrugno, L., & Bove, T. (2020). Infodemic and the spread of fake news in the COVID-19-era. European Journal of Emergency Medicine.

Osborne, R., Dunne, E., & Farrand, P. (2013). Integrating technologies into ''authentic''assessment design: an affordances approach. Research in Learning Technology, 21.

Owen, R. J. (1975). "A Bayesian sequential procedure for quantal response in the context of adaptive mental testing." Journal of the American Statistical Association(70): 351-356.

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. European Journal of Education, 53(2), 160-175.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

O'Sullivan, K., Clark, S., Marshall, K., & MacLachlan, M. (2021). A Just Digital framework to ensure equitable achievement of the Sustainable Development Goals. Nature communications, 12(1), 1-4.

Paiva, J. C., Leal, J. P., & Figueira, Á. (2022). Automated assessment in computer science education: A state-of-the-art review. ACM Transactions on Computing Education (TOCE), 22(3), 1-40.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. Contemporary educational psychology, 24(2), 124-139.

Pan, C. and Lin, C. (2018). Designing and implementing a computerized adaptive testing system with an MVC framework: A case study of the IEEE floating-point standard. IEEE International Conference on Applied System Invention (ICASI), Chiba, 609-612.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49-64.

Papamitsiou, Z., & Economides, A. A. (2016). Learning analytics for smart learning environments: A meta-analysis of empirical research results from 2009 to 2015. Learning, design, and technology: An international compendium of theory, research, practice, and policy, 1, 23.

Papamitsiou, Z., & Economides, A. A. (2017). Exhibiting achievement behavior during computer-based testing: What temporal trace data and personality traits tell us?. Computers in Human Behavior, 75, 423-438.

Pardo, A. (2018). A feedback model for data-rich learning experiences. Assessment & Evaluation in Higher Education, 43(3), 428-438.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. British journal of educational technology, 45(3), 438-450.

Pauli, M., & Ferrell, G. (2020). The future of assessment: five principles, five targets for 2025.

Pérez-Escoda, A., & Esteban, L. M. P. (2021). Retos del periodismo frente a las redes sociales, las fake news y la desconfianza de la generación Z. Revista Latina de Comunicación Social, (79), 67-85.

Pérez-Escoda, A., Pedrero-Esteban, L. M., Rubio-Romero, J., & Jiménez-Narros, C. (2021). Fake news reaching young people on social networks: Distrust challenging media literacy. Publications, 9(2), 24.

Peters, H., Kyngdon, A., & Stillwell, D. (2021). Construction and validation of a game-based intelligence assessment in minecraft. Computers in Human Behavior, 119, 106701.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., ... & Dye, D. M. (2001). Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. Personnel psychology, 54(2), 451-492.

Pina, J. A. L., & Montesinos, M. D. H. (1996). Bondad de ajuste y teoría de respuesta a los ítems. In Psicometría (pp. 643-704).

Plath, J., & Leiss, D. (2018). The impact of linguistic complexity on the solution of mathematical modelling tasks. ZDM, 50(1-2), 159-171.

Poquet, O., & Jovanovic, J. (2020, March). Intergroup and interpersonal forum positioning in shared-thread and post-reply networks. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (pp. 187-196).

Portillo, J., Garay, U., Tejada, E., & Bilbao, N. (2020). Self-perception of the digital competence of educators during the COVID-19 pandemic: A cross-analysis of different educational stages. Sustainability, 12(23), 10128.

Postigo, Á., Cuesta, M., Pedrosa, I., Muñiz, J., y García-Cueto, E. (2020). Development of a computerized adaptive test to assess entrepreneurial personality. Psicologia: Reflexão e Crítica, 33.

Prinsloo, P., & Slade, S. (2017). Ethics and learning analytics: Charting the (un) charted. SoLAR.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. IEEE Transactions on pattern analysis and machine intelligence, 22(9), 970-982. https://doi.org/10.1109/34.877520

Rahimi, S., Shute, V., Kuba, R., Dai, C. P., Yang, X., Smith, G., & Fernández, C. A. (2021). The use and effects of incentive systems on learning and performance in educational games. Computers & Education, 165, 104135.

Raîche, G., Blais, J.-G. y Magis, D. (2007). Adaptive estimators of trait level in adaptive testing: some proposals. 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, Minnesota (USA).

Raković, M., Bernacki, M. L., Greene, J. A., Plumley, R. D., Hogan, K. A., Gates, K. M., & Panter, A. T. (2022). Examining the critical role of evaluation and adaptation in self-regulated learning. Contemporary Educational Psychology, 68, 102027.

Rappa, N. A., Ledger, S., Teo, T., Wai Wong, K., Power, B., & Hilliard, B. (2019). The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: a scoping review. Interactive Learning Environments, 1-13.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Rasch, G. (1993). Probabilistic models for some intelligence and attainment tests. MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637; e-mail: MESA@ uchicago. edu; web address: www. rasch. org; tele.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological bulletin, 124(3), 372. https://doi.org/10.1037/0033-2909.124.3.372

Rayón, A., Guenaga, M., & Núñez, A. (2014, October). Supporting competency-assessment through a learning analytics approach using enriched rubrics. In Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 291-298).

Reckase, M.D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. Behav. Res. Methods 6(2), 208–212

Reckase, M. D. (2009). Multidimensional item response theory models. In Multidimensional item response theory (pp. 79-112). Springer, New York, NY.

Reeves, T. (2006). Design research from a technology perspective. In Educational design research (pp. 64-78). Routledge.

Reich, J. (2022). Learning Analytics and Learning at Scale. by Charles Lang, Alyssa Friend Wise, Agathe Merceron, Dragan Gaševic, and George Siemens. 2nd ed. Vancouver, Canada: SOLAR.

Reichert, F., Zhang, D. J., Law, N. W., Wong, G. K., & de la Torre, J. (2020). Exploring the structure of digital literacy competence assessed using authentic software applications. Educational Technology Research and Development, 68(6), 2991-3013.

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. Computers in Human Behavior, 60, 333-341. https://doi.org/10.1016/j.chb.2016.02.074

Revuelta, J. y Ponsoda, V. (1998). "A comparison of item exposure control methods in computerized adaptive testing." Journal of Educational Measurement(35): 311-327.

Revuelta, J., Ximénez, M. C. y Olea, J. (2003). "Psychometric and psychological effects of item selection and review on computerized testing." Educational and Psychological Measurement 63(5): 791-808.

Rezaie, M., & Golshan, M. (2015). Computer adaptive test (CAT): Advantages and limitations. International Journal of Educational Investigations, 2(5), 128-137.

Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in Zoombinis puzzle gameplay. Computers in Human Behavior, 120, 106707.

Sá, M. J., Santos, A. I., Serpa, S., & Miguel Ferreira, C. (2021). Digitainability—Digital competences post-COVID-19 for a sustainable society. Sustainability, 13(17), 9564.

Sahut, J. M., Iandoli, L., & Teulon, F. (2019). The age of digital entrepreneurship. Small Business Economics 1–11.

Sandoval, W. (2014). Conjecture mapping: An approach to systematic educational design research. Journal of the learning sciences, 23(1), 18-36.

Saltos-Rivas, R., Novoa-Hernández, P., & Serrano Rodríguez, R. (2021). On the quality of quantitative instruments to measure digital competence in higher education: A systematic mapping study. Plos one, 16(9), e0257344.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement.

Santos, A. I., & Serpa, S. (2017). The importance of promoting digital literacy in higher education. Int'l J. Soc. Sci. Stud., 5, 90.

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform.

Schallenmueller, S. (2016). Smart workplace technology buzz. In the impact of ICT on work (pp. 127-150). Springer, Singapore.

Scherer, R. (2015). Is it time for a new measurement approach? A closer look at the assessment of cognitive adaptability in complex problem solving. Frontiers in Psychology, 6. http://dx.doi.org/10.3389/fpsyg.2015.01664.

Scherer, R., Greiff, S., & Kirschner, P. A. (2017). Editorial to the special issue: current innovations in computer-based assessments. Computers in Human Behavior.

Scherer, R., Meßinger-Koppelt, J., & Tiemann, R. (2014). Developing a computer-based assessment of complex problem solving in Chemistry. International Journal of STEM Education, 1(1), 1-15.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. Computer-based testing: Building the foundation for future assessments, 34, 237-266.

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable?. Injury, 42(3), 236-240.

Schwab, K. (2018). The Future of Jobs Report. (pp. 1–133). World Economic Forum.

Scuotto, V., & Morellato, M. (2013). Entrepreneurial knowledge and digital competence: Keys for a success of student entrepreneurship. Journal of the Knowledge Economy, 4(3), 293-303.

Selwyn, N. (2020). Re-imagining 'Learning Analytics'... a case for starting again?. The Internet and Higher Education, 46, 100745.

Seo, D. G. (2017). Overview and current management of computerized adaptive testing in licensing/certification examinations. Journal of educational evaluation for health professions, 14.

Seppänen, M., & Gegenfurtner, A. (2012). Seeing through a teacher's eyes improves students' imaging interpretation. Medical Education, 46(11), 1113-1114. https://doi.org/10.1111/medu.12041

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D., & Chen, G. (2021, June). Assessing algorithmic fairness in automatic classifiers of

educational forum posts. In International conference on artificial intelligence in education (pp. 381-394). Springer, Cham.

Sharafi, Z., Soh, Z., Guéhéneuc, Y. G., & Antoniol, G. (2012, June). Women and men—different but equal: On the impact of identifier style on source code reading. In 2012 20th IEEE International Conference on Program Comprehension (ICPC) (pp. 27-36). IEEE. https://doi.org/10.1109/ICPC.2012.6240505

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning?. British Journal of Educational Technology, 51(5), 1450-1484.

Shaw, A. (2022). Creative minecrafters: Cognitive and personality determinants of creativity, novelty, and usefulness in minecraft. Psychology of Aesthetics, Creativity, and the Arts.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. Educational Assessment, 21, 34–59.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. Journal of Computer Assisted Learning, 33(1), 1-19.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. Computers in Human Behavior, 116, 106647.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. Educational Research Review, 19, 58-84.

Siegel, K., & Hadi, K. (2017). Articulate Storyline 3 & 360: Beyond the Essentials. IconLogic.

Siemens, G. (2012, April). Learning analytics: envisioning a research discipline and a domain of practice. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 4-8).

Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10), 1380-1400.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254).

Siiman, L. A., Mäeots, M., & Pedaste, M. (2016, October). A review of interactive computer-based tasks in large-scale studies: Can they guide the development of an instrument to assess students' digital competence?. In International Computer Assisted Assessment Conference (pp. 148-158). Springer, Cham.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema.

Sireci, S. G., & Zenisky, A. L. (2015). Computerized innovative item formats: Achievement and credentialing. In Handbook of test development (pp. 329-350). Routledge.

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In The History of Educational Measurement (pp. 111-135). Routledge.

Soler-Costa, R., Lafarga-Ostáriz, P., Mauri-Medrano, M., & Moreno-Guerrero, A. J. (2021). Netiquette: Ethic, education, and behavior on internet—a systematic literature review. International journal of environmental research and public health, 18(3), 1212.

Sostero, M., Milasi, S., Hurley, J., Fernandez-Macías, E., & Bisello, M. (2020). Teleworkability and the COVID-19 crisis: a new digital divide? (No. 2020/05). JRC working papers series on labour, education and technology.

Sparks, J. R., Katz, I. R., & Beile, P. M. (2016). Assessing digital information literacy in higher education: A review of existing frameworks and assessments with recommendations for next-generation assessment. ETS Research Report Series, 2016(2), 1-33.

Srinivasan, A., & Venkatraman, N. (2018). Entrepreneurship in digital platforms: A network-centric view. Strategic Entrepreneurship Journal, 12(1), 54-71.

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. Computers in Human Behavior, 111, 106442.

Steinfeld, J., & Robitzsch, A. (2021). Item parameter estimation in multistage designs: A comparison of different estimation approaches for the Rasch model. Psych, 3(3), 279-307.

Steingroever, H., Jepma, M., Lee, M. D., Jansen, B. R., & Huizenga, H. M. (2019). Detecting strategies in developmental psychology. Computational Brain & Behavior, 2(2), 128-140. https://doi.org/10.1007/s42113-019-0024-x

Steininger, D. M. (2019). Linking information systems and entrepreneurship: A review and agenda for IT-associated and digital entrepreneurship research. Information Systems Journal, 29(2), 363-407.

Carretero, S., Vuorikari, R., & Punie, Y. (2017). The digital competence framework for citizens. Publications Office of the European Union.

Stocking, M. L., y Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. Journal of Educational and Behavioral Statistics, 23(1), 57-75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In Computerized adaptive testing: Theory and practice (pp. 163-182). Springer, Dordrecht.

Stödberg, U. (2012). A research review of e-assessment. Assessment & Evaluation in Higher Education, 37(5), 591-604.

Sumbling, M., Sanz, P., Viladrich, M. C., Doval, E. y Riera, L. (2007). Development of a multiplecomponent CAT for measuring foreign language proficiency (SIMTEST). GMAC Conference on Computerized Adaptive Testing, Minnesota, Minneapolis (USA).

Sun, Z., & Theussen, A. (2022). Assessing negotiation skill and its development in an online collaborative simulation game: A social network analysis study. British Journal of Educational Technology.

Sussan, F., & Acs, Z. J. (2017). The digital entrepreneurial ecosystem. Small Business Economics, 49(1), 55-73.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50(3), 349-364.

Sweeney, T., West, D., Groessler, A., Haynie, A., Higgs, B. M., Macaulay, J., ... & Yeo, M. (2017). Where's the transformation? Unlocking the potential of technology-enhanced assessment. Teaching and Learning Inquiry, 5(1), 1-16.

Sympson, J. B. y Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive tests. 27th Annual meeting of the Military Testing Association, San Diego, CA (USA).

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. International journal of research & method in education, 29(2), 185-208. https://doi.org/10.1080/17437270600891614

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. Computers in Human Behavior, 76, 641-655.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. Computers in Human Behavior, 47, 157-167. https://doi.org/10.1016/j.chb.2014.05.038

The Digital Skills Gap in Europe. (2017). Factsheets. Digital skills in Europe. http://ec.europa.eu/ newsroom/ dae/document.cfm?doc_id= 47880. Accessed October 28, 2019.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47(2), 175-186.

Thissen, D. M. y Mislevy, R. J. (2000). Testing algorithms. Computerized adaptive testing: a primer (second edition). H. Wainer. Mahwah, New Jersey (USA), Lawrence Erlbaum Associates: 101-132.

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. Practical Assessment, Research, and Evaluation, 16(1), 1.

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G. Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: a systematic review. journal of surgical research, 191(1), 169-178. https://doi.org/10.1016/j.jss.2014.04.032

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. British Educational Research Journal, 42(3), 454-476.

Troussas, C., Krouska, A., & Virvou, M. (2020). Using a Multi Module Model for Learning Analytics to Predict Learners' Cognitive States and Provide Tailored

Learning Pathways and Assessment. In M. Virvou, E. Alepis, G. Tsihrintzis, & L. Jain (Eds.), Machine learning paradigms. Springer (158) https://doi.org/10.1007/978-3-030-13743-4_2.

Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. Computers & Education, 58(1), 375-385. https://doi.org/10.1016/j.compedu.2011.07.012

Tsai, Y. S., Moreno-Marcos, P. M., Jivet, I., Scheffel, M., Tammets, K., Kollom, K., & Gašević, D. (2018). The SHEILA framework: Informing institutional strategies and policy processes of learning analytics. Journal of Learning Analytics, 5(3), 5-20.

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. Computers & Education, 97, 69–85.

Uttamchandani, S., & Quick, J. (2022). An Introduction to fairness, absence of bias, and equity in learning analytics. Handbook of Learning Analytics.

Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. Applied Psychological Measurement, 25(3), 273-282.

van der Gijp, A., Ravesloot, C. J., Jarodzka, H., van der Schaaf, M. F., van der Schaaf, I. C., van Schaik, J. P. J., & ten Cate, T. J. (2017). How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education, 22, 765–787. https://doi.org/10.1007/s10459-016-9698-1

Van der Kleij, F. M., Eggen, T. J., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. Computers & Education, 58(1), 263-272.

Van der Linden, W. J. (1995). Bayesian item selection in adaptive testing. Annual meeting of the Psychometric Society, Minneapolis, Minnesota (USA).

Van der Linden, W. J., y Hambleton, R. K. (1997). Handbook of item response theory. Taylor & Francis Group. Citado na pág, 1(7), 8.

Van der Linden, W. J. (1998). "Bayesian item selection criteria for adaptive testing." Psychometrika 63(2): 201-216.

Van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. Computerized adaptive testing: theory and practice. W. J. van der Linden y C. A. W. Glas. Dordrecht (The Netherlands), Kluwer Academic Press: 25-52.

Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive Testing with shadow tests. Journal of Educational and Behavioral Statistics, 29(3), 273-291.

Van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. Journal of Educational and Behavioral Statistics, 32(4), 398-418.

Van Der Linden, W. J. (2009). Conceptual issues in response-time modeling. Journal of Educational Measurement, 46(3), 247-272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Van der Linden WJ y Glas CA (2010) Elements of adaptive testing. Springer

Van der Linden, W. J. (Ed.). (2018). Handbook of item response theory: Three volume set. CRC Press.

Van der Linden, W. J. y Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. Computerized adaptive testing: theory and practice. W. J. van der Linden y C. A. W. Glas. Dordrecht (The Netherlands), Kluwer Academic Publishers: 1-25.

Van Deursen, A. J., & Helsper, E. J. (2015). The third-level digital divide: Who benefits most from being online?. In Communication and information technologies annual. Emerald Group Publishing Limited.

Van Deursen, A. J., Helsper, E. J., & Eynon, R. (2016). Development and validation of the Internet Skills Scale (ISS). Information, Communication & Society, 19(6), 804-823.

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. Learning and instruction, 20(2), 95-99. https://doi.org/10.1016/j.learninstruc.2009.02.009.

Van Laar, E., Van Deursen, A. J., Van Dijk, J. A., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. Computers in human behavior, 72, 577-588.

Van Voorhis, V., & Paris, B. (2019). Simulations and serious games: Higher order thinking skills assessment. Journal of Applied Testing Technology, 20(S1), 35-42.

Vargas Llave, O., Mandl, I., Weber, T., & Wilkens, M. (2020). Telework and ICT-based mobile work: Flexible working in the digital age.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 1-8.

Vaterlaus, J. M., Aylward, A., Tarabochia, D., & Martin, J. D. (2021). "A smartphone made my life easier": An exploratory study on age of adolescent smartphone acquisition and well-being. Computers in Human Behavior, 114, 106563.

Veerkamp, W. J. J. y Berger, M. P. F. (1997). "Some new item selection criteria for adaptive testing." Journal of Educational and Behavioral Statistics 22(2): 203-226.

Veldkamp, B. P. (2015). Computerized test construction. In The International Encyclopedia of Social and Behavioral Sciences, 2nd Edition (pp. 510-514). Elsevier.

Veldkamp, B. P., Avetisyan, M.,Weissman, A., & Fox, J.-P. (2017). Stochastic programming for individualized test assembly with mixture response time models. Computers in Human Behavior. http://dx.doi.org/10.1016/j.chb.2017.04.060.

Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. Ensaio: Avaliação e Políticas Públicas em Educação, 21(78), 57–82. http://doi.org/10.1590/S0104-40362013005000001.

Ventura-León, J. L., & Caycho-Rodríguez, T. (2017). El coeficiente Omega: un método alternativo para la estimación de la confiabilidad. Revista Latinoamericana de Ciencias Sociales, niñez y juventud, 15(1), 625-627.

Verschoor, A. J. y Straetmans, G. J. J. M. (1999). MATHCAT: A flexible testing system in mathematics education for adults. Tests Informatizados: Fundamentos y Aplicaciones. G. Prieto. Madrid (España), Pirámide: 101-116.

Vie, J. J., Popineau, F., Bruillard, É., & Bourda, Y. (2017). A review of recent advances in adaptive assessment. Learning analytics: fundaments, applications, and trends, 113-142.

Vie, J. J., Popineau, F., Tort, F., Marteau, B., & Denos, N. (2017, April). A heuristic method for large-scale cognitive-diagnostic computerized adaptive testing. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale (pp. 323-326).

Viner, K. (2016) How Technology Disrupted the Truth. The Guardian. 12 July 2016. Available online: https://www.theguardian.com/media/2016/jul/12/how-technology-disrupted-the-truth (accedido el 13 de octubre 2022).

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. Computers in Human Behavior, 76, 656-671.

von Davier, A. A., Hao, J., Liu, L., & Kyllonen, P. (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: Lessons learned from developing a collaborative science assessment prototype. Computers in Human Behavior, 76, 631-640.

Voogt, J., Erstad, O., Dede, C., & Mishra, P. (2013). Challenges to learning and schooling in the digital networked world of the 21st century. Journal of computer assisted learning, 29(5), 403-413.

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. Journal of curriculum studies, 44(3), 299-321.

Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. Political Communication, 37(1), 136-144.

Vuorikari, R., Punie, Y., Gomez, S. C., & Van Den Brande, G. (2016). DigComp 2.0: The digital competence framework for citizens. Update phase 1: The conceptual reference model (No. JRC101254). Joint Research Centre (Seville site).

Vuorikari Rina, R., Kluzer, S., & Punie, Y. (2022). DigComp 2.2: The Digital Competence Framework for Citizens-With new examples of knowledge, skills and attitudes (No. JRC128415). Joint Research Centre (Seville site).

Wainer, H., Kaplan, B. y Lewis, C. (1992). "A comparison of the performance of simulated hierarchical and linear testlets." Journal of Educational Measurement(29): 243-251.

Wainer, H. (2000b). Computerized adaptive testing: a primer (second edition). Mahwah, New Jersey (USA), Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T. y Du, Z. (2000). Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. Computerized adaptive

testing: theory and practice. W. J. van der Linden y C. A. W. Glas. Dordrecht (The Netherlands), Kluwer Academic Publishers: 245-269.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). Computerized adaptive testing: A primer. Routledge.

Wainer, H., Bradlow, E. T., y Wang, X. (2007). Testlet response theory and its applications. Cambridge University Press.

Walker, J., Böhnke, J. R., Cerny, T., y Strasser, F. (2010). Development of symptom assessments utilising item response theory and computer-adaptive testing—a practical method based on a systematic review. Critical reviews in oncology/hematology, 73(1), 47-67.

Walsh, A. (2009). Information literacy assessment: where do we start?. Journal of librarianship and information science, 41(1), 19-28.

Wang, T. y Vispoel, W. P. (1998). "Properties of ability estimation methods in computerized adaptive testing." Journal of Educational Measurement(35): 109-135.

Wang S, Fellouris G, Chang HH (2015) Sequential design for computerized adaptive testing that allows for response revision. arXiv preprint arXiv:150101366

Wang S, Lin H, Chang HH, Douglas J (2016) Hybrid computerized adaptive testing: From group sequential design to fully sequential design. Journal of Educational Measurement 53(1):45–62

Wang, X., Zhang, Y., Yu, S., Liu, X., Wang, F.Y. (2017). Computerized Adaptive English Ability Assessment Based on Deep Learning. Image and Video Technology, Satoh S. (eds.), Lecture Notes in Computer Science, 10799, Springer, Cham, 158–171.

Weise, J. J., Greiff, S., & Sparfeldt, J. R. (2022). Focusing on eigendynamic effects promotes students' performance in complex problem solving: A log-file analysis of strategic behavior. Computers & Education, 189, 104579.

Weiss, D. J. (1982). "Improving measurement quality and efficency with adaptive testing." Applied Psychological Measurement(6): 473-492.

Weiss, D. J. (1985). Adaptive testing by computer. Journal of consulting and clinical psychology, 53(6), 774.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. Measurement and Evaluation in Counseling and Development, 37(2), 70-84.

Weiss, D. J. y Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of educational measurement, 21(4), 361-375.

Wietsma, T.: OSCATS: Open Source Computerized Adaptive Testing System, March 2016. https://github.com/tristanwietsma/oscats

Wilson, M. (2004). Constructing Measures: An Item Response Modeling Approach: An Item Response Modeling Approach. Routledge.

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. Computers in Human Behavior, 112, 106457.

Wise, S. L., y Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. Psicológica, 21(1), 135-155.

Wise, A. F., y Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2(2), 5-13.

Wise, A. F., Sarmiento, J. P., & Boothe Jr, M. (2021, April). Subversive learning analytics. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 639-645).

World Economic Forum (2017). Accelerating workforce reskilling for the fourth industrial revolution: An agenda for leaders to shape the future of education, gender and work. Geneva, Switzerland: World Economic Forum.

Wright, B. D., & Stone, M. H. (1979). Best test design.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. Journal of Applied Measurement.

Wu, H. M., Kuo, B. C., & Wang, S. C. (2017). Computerized dynamic adaptive tests with immediately individualized feedback for primary school mathematics learning. Journal of Educational Technology & Society, 20(1), 61-72.

Xie, H., Zhao, T., Deng, S., Peng, J., Wang, F., & Zhou, Z. (2021). Using eye movement modelling examples to guide visual attention and foster cognitive performance: A meta-analysis. Journal of Computer Assisted Learning, 37(4), 1194-1206.

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. British Journal of Mathematical and Statistical Psychology, 69(3), 291-315.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. Educational Measurement: Issues and Practice, 37(4), 16-27.

Yan D, von Davier AA, Lewis C (2014) Computerized Multistage Testing. CRC Press

Yan, D., Von Davier, A. A., y Lewis, C. (Eds.). (2016). Computerized multistage testing: Theory and applications. CRC Press.

Yaneva, V., Clauser, B. E., Morales, A., & Paniagua, M. (2021). Using Eye-Tracking Data as Part of the Validity Argument for Multiple-Choice Questions: A Demonstration. Journal of Educational Measurement, 58(4), 515-537.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. Applied psychological measurement, 43(5), 388-401.

Yunis, M., Tarhini, A., & Kassar, A. (2018). The role of ICT and innovation in enhancing organizational performance: the catalysing effect of corporate entrepreneurship. Journal of Business Research, 88, 344–356

Zamfir, A. M., & Aldea, A. B. (2020). Digital Skills and Labour Market Resilience. Postmodern Openings/Deschideri Postmoderne, 11.

Zechner, K., Yoon, S. Y., Bhat, S., & Leong, C. W. (2017). Comparative evaluation of automated scoring of syntactic competence of non-native speakers. Computers in Human Behavior, 76, 672-682.

Zenisky, A. L., & Luecht, R. M. (2016). The future of computer-based testing. In C. Wells, & M. Faulkner-Bond (Eds.), Educational Measurement: From foundations to future (pp. 221e238). New York, NY: Guilford Press.

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. British Journal of Mathematical and Statistical Psychology, 71(2), 262-286.

Zhang, J., Ober, T. M., Jiang, Y., Plass, J., & Homer, B. D. (2021). Predicting Executive Functions in a Learning Game: Accuracy and Reaction Time. In EDM.

Zhang, S., Bergner, Y., DiTrapani, J., & Jeon, M. (2021). Modeling the interaction between resilience and ability in assessments with allowances for multiple attempts. Computers in Human Behavior, 122, 106847.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H. H. (2014). Overview of test assembly methods in multistage testing. Computerized multistage testing: Theory and applications, 87-99.

Zhao, Y., Llorente, A. M. P., & Gómez, M. C. S. (2021). Digital competence in higher education research: A systematic literature review. Computers & Education, 168, 104212.

Zickar, M. J. (2020). Measurement development and evaluation. Annual Review of Organizational Psychology and Organizational Behavior, 7, 213-232.

Zumbo, B. D. (2014). What role does, and should, the test standards play outside of the United States of America?.

Zumbo, B. D., & Chan, E. K. (2014) (Eds.) Validity and validation in social, behavioral, and health sciences. (Vol. 54). New York (US): Springer International Publishing.

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). Understanding and investigating response processes in validation research (Vol. 26). Cham, Switzerland: Springer International Publishing.