applied sciences

MDPI

Article

Using Eye-Tracking Data to Examine Response Processes in Digital Competence Assessment for Validation Purposes

Juan Bartolomé [1,*], Pablo Garaizar [2], Erlantz Loizaga [1] and Leire Bastida [1]

[1] TECNALIA, Basque Research and Technology Alliance (BRTA), 48160 Derio, Bizkaia, Spain; erlantz.loizaga@tecnalia.com (E.L.); leire.bastida@tecnalia.com (L.B.)
[2] Faculty of Engineering, University of Deusto, 48007 Bilbao, Bizkaia, Spain; garaizar@deusto.es
* Correspondence: juan.bartolome@tecnalia.com

Abstract: Background: When measuring complex cognitive constructs, it is crucial to correctly design the evaluation items in order to trigger the intended knowledge and skills. Furthermore, assessing the validity of an assessment requires considering not only the content of the evaluation tasks, but also how examinees perform by engaging construct-relevant response processes. Objectives: We used eye-tracking techniques to examine item response processes in the assessment of digital competence. The eye-tracking observations helped to fill an 'explanatory gap' by providing data on the variation in response processes that cannot be captured by other common sources. Method: Specifically, we used eye movement data to validate the inferences made between claimed and observed behavior. This allowed us to interpret how participants processed the information in the items in terms of Area Of Interest (their size, placement, and order). Results and Conclusions: The gaze data provide detailed information about response strategies at the item level, profiling the examinees according to their engagement, response processes and performance/success rate. The presented evidence confirms that the response patterns of the participants who responded well do not represent an alternative to the interpretation of the results that would undermine the assessment criteria. Takeaways: Gaze-based evidence has great potential to provide complementary data about the response processes performed by examinees, thereby contributing to the validity argument.

Keywords: digital competence; response process validity; eye tracking; human computer interaction; computer-based assessment

Academic Editors: Rui Araújo and Lykourgos Magafas

Received: 9 December 2024
Revised: 20 January 2025
Accepted: 22 January 2025
Published: 24 January 2025

Citation: Bartolomé, J.; Garaizar, P.; Loizaga, E.; Bastida, L. Using Eye-Tracking Data to Examine Response Processes in Digital Competence Assessment for Validation Purposes. *Appl. Sci.* **2025**, *15*, 1215. https://doi.org/10.3390/app15031215

1. Introduction

Digital competence (DC), commonly referred to as the ability to use digital technologies effectively, safely and responsibly, is a key factor used to avoid social and labor exclusion, as it enables individuals to effectively navigate and utilize technology in various aspects of life [1]. Over the last few years, interest in the assessment of DC has grown significantly, and has been used to accredit the level of DC of individuals. Although several reviews have summarized the progress and shortcomings in this area, some issues are yet to be explored, with a specific focus on instrument validity and reliability, and the identifying of important gaps of sources of evidence to ensure the quality of available tools [2–4]. These limitations directly impact stakeholders in the following ways: individuals may receive inaccurate assessments of their DC, hindering their employability and social integration; educators struggle to design curricula that accurately reflect students' competencies; and policymakers face challenges in developing informed strategies to enhance digital literacy across populations. Consequently, the inadequacies of current DC assessment tools not only undermine the effectiveness of digital literacy initiatives, but also perpetuate the risks of social exclusion and technological

disparity. By addressing these gaps, we can ensure more reliable and valid assessments, thereby supporting individuals in achieving DC and fostering a more inclusive digital society.

A key challenge in assessing DC lies in the varied and context-dependent definitions of the construct. While experts have sought to identify essential DCs for all citizens [5], differing definitions [4,6] complicate the development of consistent assessment methods. The absence of a universally accepted definition of DC results in tools that may assess certain skills but overlook others [7], limiting comparability across studies and settings. Additionally, the context-specific nature of current assessment tools hinders their general applicability, as tools validated in one setting may not be reliable in another [6]. To address this, the European Commission introduced DigComp (Digital Competence Framework for Citizens) [7], which defines DC in terms of the knowledge, skills, and attitudes required to be digitally competent. While DigComp represents progress in standardizing DC, further research is needed to validate its use across diverse contexts and populations, and to address its limitations in capturing the evolving nature of DC. Continued efforts to refine and expand the framework will enable the development of more accurate, adaptable, and comprehensive assessment tools for DC.

Another significant issue is that, despite the availability of numerous DC assessment systems using various approaches, most rely on self-assessment, which inadequately measures practical skills and higher-level cognitive abilities [2,8]. Moreover, as noted by Law et al., many existing DC frameworks are developed by commercial enterprises [3]. This reliance on proprietary software, such as Microsoft Office or Windows, can shape curricular DCs around specific tools rather than broader competencies. While the introduction of DigComp has enabled more tailored implementations [2,9], current systems remain largely based on self-reporting, often using multiple-choice questions and Likert scales. This approach minimally assesses practical skills, and individuals often overestimate their competence [10], rendering such assessments unreliable and invalid for accurate competence accreditation.

Against this confusing and restricted background, we have identified several lines of action that could be envisaged to address the shortcomings identified.

The use of Technology-Enhanced Assessment (TEA) presents both opportunities and challenges in the evaluation of DC. TEA offers interactive and adaptive testing environments that better simulate real-world digital tasks, providing a more accurate measure of complex problem-solving and higher-order thinking skills [3,4,11]. It also enables the assessment of the skills dimension of DC, including high-level cognitive skills, through innovative item formats [12].

However, technical limitations, such as the digital divide and varying familiarity with platforms, may introduce biases that affect the validity of results [13]. Additionally, the design of assessment items is crucial, as performance outcomes are strongly influenced by item format [14,15]. Poorly designed items can misrepresent examinees' DC levels, especially when assessing complex cognitive skills. While innovative approaches like performance-based tasks (i.e., assessments where individuals demonstrate their skills or knowledge by completing a specific, real-world task rather than answering questions or selecting answers) offer a more authentic measure of DC [16,17], traditional self-report tools often lead to overestimations due to social desirability bias and the Dunning–Kruger effect [4].

Despite its potential, the high development costs and technical complexity of TEA limit its broader adoption and scalability [18]. Moreover, it is challenging to infer the response processes (RPs) examinees use to complete tasks, making it difficult to confirm if they approached problems as intended or employed alternative strategies. Thus, while TEA represents progress, its effectiveness depends on addressing these limitations and ensuring fair access and user competence.

Another key approach involves utilizing TEA to collect detailed data on examinee behavior and performance, such as log data, response time (RT), and click streams [19,20].

Although some researchers have used these data to enhance score interpretation and validate assessments [21], such studies are rare, particularly in the assessment of DC. Analyzing RP data from assessments can help validate item design by showing that tasks elicit the intended knowledge and skills. However, evaluating whether examinees engage with relevant RPs is often overlooked when considering test validity, and evidence of this is lacking in most DC evaluation tools [4,22].

The integration of ET data offers new opportunities to analyze cognitive processes during assessments [23]. ET data provide fine-grained insights into examinee behavior at the item level, revealing attention patterns, decision-making processes, and problem-solving strategies [21]. For example, in image-based questions, gaze data can identify the areas participants focus on most, offering a unique view into their problem-solving strategies and pinpointing differences in performance. Such data can help validate DC assessments by providing direct evidence of cognitive engagement and resource allocation [24].

However, the implementation of ET data presents challenges, such as the need for sophisticated technology, complex data interpretation, and potential intrusiveness, which could affect performance [25]. Additionally, the large-scale use of ET remains costly and logistically difficult, making it hard to scale [26]. While innovative item designs and ET data have the potential to improve the accuracy of DC assessments, practical and financial barriers must be overcome to reach widespread application.

In short, it is essential to design items that prompt more complex RPs from test-takers, while ensuring that these items trigger the intended knowledge and skills. Relying solely on RT data and item scores is insufficient. This study aims to use gaze data as additional evidence to support RP analysis and enhance the validity of the assessments. This paper builds on our previous research by incorporating gaze data into DC assessment to validate item design and differentiate participants based on their DC levels [27]. We focused on analyzing variations among participants with different DC levels using fixation-based metrics supported by visualization techniques.

The purpose of this exploratory study was to design items suitable for the assessment of DC focused on higher-order thinking skills and gather validation evidence in a custom TEA implementation based on DigComp. In alignment with the United Nations Educational, Scientific and Cultural Organization (UNESCO) [3] and the World Bank [28], we adopted the DigComp framework as our reference model, owing to its significant advantages, as follows: (1) it was developed following an in-depth examination of existing DC frameworks, (2) it underwent a rigorous process of consultation and refinement by specialists in the field of DC, and (3) consequently, it offers a holistic perspective grounded in DC and its related areas. We utilized data from RPS, including ET data, to aid in validating these items and consequently the test itself. We seek to contribute to further the understanding of participants' task-solving behaviors by examining the scan-paths of the participants, which may reveal important information about differences that may not be captured by the final responses, and to explore whether evidence from this analysis can contribute to the validity argument for inferences based on scores. Scan-path and fixation-based measures are well-suited for this study because they provide detailed insights into how participants interact with assessment tasks, offering a nuanced understanding of their cognitive processes and problem-solving strategies. Empirical research has demonstrated that fixation metrics, such as duration and sequence, correlate with information processing and cognitive load, making them valuable for assessing DC [24,29]. To date, we are not aware of a similar study having been carried out on the design of DC assessments.

Cronbach asserts that validation aims to support an interpretation by rigorously testing for potential errors [30]. A claim is only trustworthy if it has withstood serious attempts to disprove it. This perspective on validity is relevant to studies like this one, which do not offer

direct evidence that the cognitive processes involved in test item responses mirror those used in real-world tasks. However, they can help demonstrate their differences [31]. For example, if ET data show that test-takers completed simulation tasks without thoroughly reviewing instructions, it would challenge whether these tasks accurately reflect real-world cognitive processes. While ET data do not directly confirm cognitive processes, they provide valuable inferences [32]. So, we used ET data to validate the inferences made between claimed and observed behavior. We sought to respond to the following research questions:

(1) Is it possible to identify an alternative interpretation of how participants processed the item in terms of areas of interest (AOIs) examined (AOIs are usually defined by researchers to study eye movements within specific limited areas)?
(2) Is it possible to identify an alternative interpretation of how participants processed the item in terms of AOIs examined and the order followed for the different AOIs?

This paper begins by discussing the background of RP validation. Despite its importance in test validation, RP evidence has been largely underutilized. Given the benefits of RP evidence and the capabilities of TEA, we chose to explore RP data for DC assessment validation, specifically using ET data. Section 3 introduces ET technology and its ability to provide valuable insights into test-takers' performance, emphasizing its potential for assessment validation, particularly in DC evaluation. Section 5 outlines the experimental methodology, focusing on validating alternative interpretations of how participants processed selected items based on AOIs and the order in which they were examined. Metrics and visualization methods from the literature were used to analyze information acquisition and cognitive processing. Section 6 presents the results, followed by the conclusion and a discussion of future directions.

2. Validation, Evidence Based on Response Processes

Validation can be defined as the process of collecting evidence to support specific interpretations and uses of test scores. In the framework provided by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education [33], five sources of evidence were identified, as follows: (1) test content; (2) RPs; (3) internal structure; (4) relations to other variables; and (5) consequences of testing. Moreover, evidence based on test content is complemented with evidence based on RPs.

According to Ercikan and Pellegrino [34], "RPs refer to approaches and behaviors of examinees when they interpret assessment situations and formulate and generate solutions as revealed through verbalizations, eye movements, response times, or computer clicks. Such response process data can provide information about the extent to which items and tasks engage examinees in the intended ways". Hubley and Zumbo [35] defined RPs as "the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed score variation".

Validity evidence based on RPs involves understanding how examinees interact with a test and whether they engage the intended cognitive processes when answering items. Empirical data are needed to confirm this. RPs offer valuable insights into examinees' performance, helping test designers reduce the gap between observed responses and the constructs being measured. Although not a new area of research, the study of cognitive processes used by examinees in different tasks has a long history (e.g., [36]). Collecting this evidence involves various methods such as think-aloud interviews, cognitive interviews, keystroke and mouse click analysis, ET techniques, and RT analysis [33].

2.1. Applications of Evidence of RPs in Assessment

Small-scale process data have gained increasing attention in cognitive ability testing due to advances in digital measurement, which enhance data collection by allowing RPs

to be traced and recorded [29]. In large-scale assessments, such data support validation practices and test development [34,37], including the analysis of computer-generated log files, screen captures, facial expressions, eye movements, video-recorded behaviors, and physiological sensor data [38].

RP evidence is especially important for tests measuring higher-order thinking, such as DC assessments, or for developing new assessment formats [34]. Test-takers may use unintended pathways or strategies to solve items, potentially bypassing the cognitive skills being measured. For instance, Yaneva et al. used ET data to evaluate how the layout of options in multiple-choice questions influenced responses, offering an alternative interpretation of test scores [39]. Specian Junior et al. employed ET data to validate response processes in clinical vignette-type multiple-choice questions, improving question design and providing feedback for teaching [40]. Kraitzek and Förster explored whether ET data revealed emotional engagement in financial competence assessments [41]. RPs have also been studied extensively to assess the alignment between test-takers' responses and the intended construct [42–45], and to inform group comparisons by explaining score differences across examinee groups [34,46].

2.2. Challenges and Contradictory Findings

RT analysis is recommended in assessment design as it enhances and evaluates numerical models by reducing subjectivity [47,48]. For instance, RT analysis has been applied to examine test-takers' engagement and motivation, particularly in detecting rapid guessing [49]. Group comparisons often focus on final answers and responses, overlooking differences in RPs across groups. However, RT alone does not explain why users spent a certain amount of time on tasks. Therefore, for validation purposes, it is advisable to supplement RT data with other process data [50]. Araneda et al. emphasized that while RP data can vary, only relevant data linked to the intended constructs are crucial [51]. Not all processes reflect the targeted construct, and typically only some are used in scoring.

2.3. Advantages and Limitations of Using Evidence of RPs

Process data such us log data, ET and RT can contribute to validating the inferences made between claims and observed behavior in at least five possible ways, as follows [21]: (1) inference, to generate and test inferences about the construct of interest; (2) design, to improve task and item designs analyzing the patterns of behaviors shown by the participants; (3) contextual evidence, to offer a context to the inferences related to the construct of interest; (4) indicators of engagement, to provide an indicator of the level of engagement of the participants with the tests; and (5) data cleaning and filtering, to remove behaviors and responses not valid. For the purposes of this study, we will focus on the first two, since our aim is to validate the inferences made between stated and observed behavior, in order to confirm that the items were well designed and, as a result, showed the expected behaviors.

Although the importance of analyzing examinees' RPs is well-recognized [34,35], it is regrettable that evidence related to RPs has scarcely been integrated into validation research studies, probably because their analysis is difficult due complex sequences with many variables and nonstandard formats [52,53], and further, it is not properly understood [54]. To the best of our knowledge, no prior studies have explored RPs within the context of validating DC assessments using ET data.

3. Eye-Tracking Technology

ET is a method that records either the point of gaze (fixation) or eye movement (saccade) relative to the user's head and task-relevant objects [29]. Interest in ET has grown significantly in recent decades due to advancements in non-intrusive devices with faster

data capture. Eye movements play a key role in guiding cognitive processes by directing attention to stimuli processed by the brain [55] and reflecting cognitive activity during tasks like reading and information processing [25]. Thus, ET can provide valuable insights into test-takers' performance beyond traditional metrics like logs or response times.

3.1. Applications of ET

Recent years have seen the increased use of eye-tracking (ET) technology across various fields, driven by the strong link between cognitive behavior and eye movement patterns. ET has been applied to research on reading, visual search, scene perception, and usability [56]. Several systematic reviews have highlighted ET's value in understanding cognitive processes in multimedia learning environments [57,58], including animated or simulated settings [59], and virtual or mixed reality environments [60]. ET has also been integrated into assessments, showing its effectiveness in differentiating between skill levels using gaze-tracking metrics [61,62]. Studies have also used ET to design and validate assessments, such as those by Nisiforou and Laghos, Oranje et al., and Yaneva et al. [21,39,63], although only Yaneva et al. explored alternative interpretations of test scores, focusing on question design in multiple-choice formats, which primarily assess lower-order cognitive skills [39]. To assess higher-order DC, diverse response formats and more complex reasoning processes are required.

Moreover, numerous studies have examined how individuals interact with digital tools, crucial for understanding DC and the design of assessment items based on digital interfaces [5]. Eye fixation patterns have been shown to play a crucial role in recognizing and interacting with information [64–66]. Cognitive strategies have been studied by analyzing where and in what sequence participants look at specific areas [67]. ET research has also proven useful in understanding user behavior on search engines [68] and identifying which visual elements on web pages are fixated upon, creating common scan-paths in visual processing [69].

3.2. Advantages and Limitations of Using Evidence of ET

Tien et al. conducted a systematic review highlighting that gaze and fixation overlaps are commonly used as indicators of areas perceived as important by participants [70]. Eye movement modeling examples and observing another person's gaze have proven beneficial in various contexts. For instance, they guide novice radiographers' attention in medical diagnostics [71], improve students' interpretations of medical images [72], and aid in clinical reasoning [73]. Additionally, they enhance performance monitoring [74], facilitate the integrative processing of verbal and visual information [75], and support problem-solving in science assessments [70].

Despite these positive findings, some research presents conflicting results, suggesting that visual search effectiveness depends on the domain, task, and expertise level. For example, Eder et al. found no performance improvement among students interpreting dental panoramic radiographs after using eye movement modeling as a training tool [76]. This highlights the need for further investigation [77]. However, no studies have yet applied this technique in DC assessment, where it could be highly valuable.

3.3. Commonly Used Methods and Metrics

Most cognitive processing and information acquisition occur during fixations, which is why fixation-based metrics are commonly used in research [29,78,79]. Analyzing the locations and durations of longer fixations is a standard approach for assessing web processing difficulty [80], as fixation duration can serve as an indicator of cognitive load [24]. Researchers typically define AOIs to focus on specific regions of eye movement data and Time of Interest (TOI) to isolate relevant time intervals. The use of qualitative and exploratory analysis, especially through visualization techniques, has become more prominent [56]. These methods include string-edit and sequence-based representations [56,81,82], Markov models [83,84], and classifications of raw eye-tracking data [85,86]. Additionally, algorithms have been developed

to analyze individual scan-paths, identifying commonly followed visual patterns [82] of interest for the purpose of this study.

In brief, ET technology offers an innovative approach to assessment by capturing real-time data on visual attention and cognitive processes, providing insights that traditional methods often miss. While research underscores ET's potential in DC assessment, it is crucial to critically assess the methodologies and findings to fully understand their implications. Studies by Holmqvist et al. and Eivazi and Bednarik [87,88] reveal that ET offers detailed insights into users' attention and interaction patterns, enhancing the precision of DC assessments by evaluating both outcomes and cognitive processes. Additionally, studies like those by Oranje et al. and Yaneva et al. suggest ET can validate test score interpretations by analyzing how individuals engage with tasks [21,39,53].

However, significant challenges remain. The high cost and technical complexity of ET, as noted by Duchowski, restrict its accessibility and large-scale use [29]. ET data collection is resource-intensive, requiring specialized tools and expertise, and visual attention, measured through ET, does not always correspond to actual understanding [89]. Moreover, while ET may show if examinees engage with tasks as intended, it does not necessarily confirm that these cognitive processes mirror real-world actions. Variability in study outcomes, due to the lack of standardized protocols, further complicates its implementation [88,89].

Despite these obstacles, integrating ET into DC assessment presents a promising opportunity to enhance the validity and reliability of assessments. As research by Siddiq et al. highlights, ET can offer a more comprehensive evaluation of DC [90].

4. DigComp, Reference Framework for the Evaluation of DC

In 2013, the Institute for Prospective Technological Studies (IPTS) of the European Commission's Joint Research Centre introduced the Digital Competence Framework (DigComp), which consolidated existing models of DC [7]. This framework divides DC into five key domains, as follows: information and data literacy, communication and collaboration, digital content creation, safety, and problem-solving, encompassing a total of 21 distinct competencies. In 2017, DigComp 2.1 introduced major updates, expanding proficiency levels from three to eight, using Bloom's taxonomy to define DC descriptors [91].

DigComp serves as a comprehensive reference framework, structured around five dimensions—(1) competence areas, (2) descriptors for each competence, (3) proficiency levels, (4) expected knowledge, skills, and attitudes, and (5) areas of application. The World Bank recognized it as one of the most extensive and widely applied frameworks for general DC [28], which was key for this study, as its assessment tool targets the general public. Moreover, DigComp is technology-agnostic, describing DC independent of specific tools or devices.

In designing evaluation instruments, constrained response formats are commonly used due to their simplicity and ease of automatic correction. However, these formats are insufficient for assessing higher-order skills, particularly at intermediate and advanced levels of DigComp. More advanced methods, such as interactive simulations, are necessary for accurate DC evaluation [92].

5. Materials and Methods

5.1. Methodology

We carried out an exploratory study that used participants' eye-movements to provide insights into their performance in a custom TEA implementation based on DigComp. The details of the assessment tool used, such as the number and type of items administered and the level of validity and reliability, can be found in the previous work by the authors Bartolomé and Garaizar, and this information is available on www.evaluatucompetenciadigital.com (accessed on 23 January 2025) [93]. This tool provided 4 different tests targeting the following DCs according to DigComp version 2.1 [91]: browsing, searching, and filtering data, information, and digital

content; evaluating data, information, and digital content; managing data, information, and digital content; and netiquette. Each test assesses one DC independently, without considering the possible overlaps between the different DCs as stated in DigComp. This tool is also part of BAIT, the Basque Government's evaluation system, accessible at http://www.bait.eus (accessed on 23 January 2025).

In our study, several optimization measures were implemented to enhance the accuracy and reliability of ET data. First, an advanced equipment calibration protocol was followed to ensure the precise alignment of the ET system with the participant's gaze. This involved multi-point calibration, repeated at regular intervals throughout the data collection process to account for potential drift in tracking accuracy. Additionally, to mitigate interference factors, the experimental environment was carefully controlled, e.g., ambient lighting was standardized, reflective surfaces were minimized, and participants were instructed to maintain a consistent head position. These measures collectively aimed to minimize extraneous variability, thereby enhancing the validity of the ET data.

The interactions of participants with the tests were gathered to assess if the designed evaluation tasks elicited the targeted knowledge and skills. Additionally, these interactions were used to classify individuals into different expertise levels based on the assessment criteria established for each task. In doing so, we are evaluating the credibility of an alternative interpretation of the test results. That is, the test assesses the skills of participants using a strategy that may be only imprecisely related to the proficiency of the test of items. We examined the scan-paths of the participants' RPs during the resolution of the image/simulation-based items to evaluate an alternative interpretation of the test scores. To do so, we created the scan-paths in terms of the visual elements included.

Image/Simulation-based questions are well-suited for assessing intermediate and advanced levels of DC. In these tasks, participants evaluate scenarios presented in visual formats, selecting the correct option, similar to multiple-choice questions. This approach assesses various DCs, such as information literacy, practical abilities, and critical thinking. By interpreting visual data, such as infographics or multimedia, participants apply their knowledge to analyze and make informed decisions, reflecting real-world digital interactions.

Incorporating these questions into assessments offers a strong measure of one's ability to navigate digital content, aligning with the DigComp framework's focus on practical application and critical engagement. Considering that in the interactive simulations we can identify the scan-path followed by the participants thanks to the click log, we focused on this question format, since the gaze data are the unique source available for describing the strategy of the participants. This method is also cost-effective and applicable across various DCs.

In order to answer the first research question, we carry out a scan-path analysis evaluating alternative interpretations based on the AOIs examined, taking the following steps:

- We examined whether there were any differences between the scores and the type of question (if required a systematic approach to solve the question or not, i.e., 1 or 0). With this aim, we calculated the Pearson coefficient to analyze the relationship between the scores, the type of question, the time to solve the question, the scan-path length, and AOIs to be checked from the total number of AOIs defined. Pearson's correlation coefficient is a widely used measure of the strength and direction of the linear relationship between two variables. One important assumption for its proper application is that the data should ideally come from a bivariate normal distribution. However, Pearson's correlation is robust to deviations from normality, especially when sample sizes are sufficiently large. This robustness makes Pearson's correlation applicable and reliable in many practical situations where the normality assumption may not be perfectly met. Therefore, we opted to use Pearson's correlation to examine the relationships between variables in this study,

as it provides a straightforward and interpretable measure of linear association that is resilient to moderate departures from normality;

- Additionally, we analyzed the influence of each AOI with a double purpose. First, we wanted to detect the most relevant areas to assess each question correctly. To do so, we performed a feature correlation analysis based on the mutual information between each of the visited AOIs and the result of the interaction. First, we carried out a minimal study of the data variance to discard invariant data that would become impractical for the analytical procedure. Then, we performed a feature correlation analysis between each of the remaining AOIs and the task result. Considering the categorical nature of both features and target, we used the estimate mutual information for discrete variables as the correlation metric;
- Finally, besides the pure analytical analysis of the results, we decided to perform a categorical classification of the participants' behavior with a double purpose. We wanted to examine whether specific patters of AOIs visited to solve the task supposed higher success rates in an inconsistent way with expectations for each item. First, this kind of analysis allows us to identify the flexibility of a task, providing a direct insight into whether the task is somehow guided, with most of the participants following a single route of resolution, or whether it is a highly free task, with not obvious pattern to resolve it. Secondly, it is possible to link those possible patterns with the obtained results to verify if any of them is more effective than others in solving the problem. So, we used a clustering classification to identify the different behavioral patterns during the experiment, using the visits of each participant to each AOI as classification features. We employed the OPTICS algorithm [94] to determine appropriate clustering regarding the participants' behavior. This algorithm offers two significant advantages over other traditional algorithms such as k means. First, it is compatible with Boolean friendly distances, such as Hamming distance. Considering that all the features used during classification are of a Boolean nature, this is a key factor to select this approach. Second, unlike k means, this algorithm detects the optimum cluster amount based on the data distribution itself. Using this approach allows us to avoid predefining the number of clusters to discover during the process. In order to check if any of the identified pattern behaviors is more efficient in solving the task, we assigned each of the participants to the corresponding cluster, and then we performed a Kruskal–Wallis test to check for significant differences between the mean values of both the result (the correct answer for the specific task) and the overall performance (the global score considering a task related to the same competence of task in question). Additionally, we investigated the sensitivity of each item to the visitation of associated AOIs, aiming to determine if some items were more predictable than others based solely on AOI visits.

To answer the second research question, we evaluated alternative interpretations based on the AOIs examined and the processing order of the AOIs, by taking the following steps:

- We used the String-edit algorithm to investigate the variance within the groups, as it has been widely used in eye-tracking research [82]. The String-edit algorithm calculates the distance between two scan-paths by transforming one scan-path to another with the minimum number of editing operations (substitution, deletion, and addition). The similarity between the two scan-paths can be calculated as a percentage based on the String-edit distance. There is no variance for two identical scan-paths, and the variance is at the maximum level for two completely different scan-paths. For example, the String-edit algorithm calculates the distance between ABCD and ABCA as one because only one substitution operation between D and A is sufficient to transform one of them to another. We complemented the analysis with the use of the ScanGraph tool [95]. This tool is publicly

available at http://eyetracking.upol.cz/scangraph/ (accessed on 23 January 2025) and can be used to generate a visual graph based on the String-edit similarity, where similar scan-paths are connected to each other. Groups of similar sequences are displayed as cliques of this graph. In addition, we also examined the scan-paths of the participants who responded well to the item (onwards successful participants) to identify one scan-path for representing the entire group, which is typically known as a common scan-path. We wanted to know how the successful participants solved these questions, i.e., in which AOIs and in what order they realized fixations. This representation can be very helpful in determining, in a simple way, that the common scan-path does not represent an alternative to the interpretation of the results that would undermine the assessment criteria defined for a particular item. With this aim, we applied a position-based weighted model, just as Holsanova et al. applied, to analyze reading paths and priorities on newspaper spreads [96]. So, we firstly divide the images presented in the stimulus into their AOIs, and then rank them based on the first visits of the AOIs by participants. For instance, in Item4, we defined six AOIs for the different visual elements. Thus, we applied the position-based weighted model by giving 6 points to the first element visited, and no point to the non-visited visual elements, obtaining the following scan-path: DABCHG. Before applying the model, we decided to simplify the original AOIs by joining close and related AOIs. In Item4, we joined DEF into a unique AOI, D, representing the entire subject field of the email. In Item24, we joined AB into a unique AOI, A, representing the entire URL section of the browser. In Item32, we joined AB into a unique AOI, A, representing the entire URL section of the browser, and FG into a unique AOI, F, representing the headline section of the news. Although the same AOI could be visited several times, the repetitions were not considered. Despite the limitations indicated by Eraslan et al., we opted to use this model to gain initial insights into the successful participants' performance [69]. The method's simplicity yielded highly beneficial results that enabled us to check that participants who responded to the question successfully did not show an alternative behavior that would invalidate the question;

- Additionally, we examined the AOIs with the longest fixations to check if they were placed according to the assessment criteria defined for each item.

To conduct our analysis, we firstly created the scan-paths in terms of the visual elements of images included. For instance, if a participant fixated on the elements A, B and C, respectively, their scan-path was generated as ABC by keeping the fixation durations. We also decided to remove the data from participant 28 in Item24, and participants 5 and 31 in Item5, since we had problems recording their interactions, and their information was inconsistent.

5.2. Ethics Statement

This study was non-interventional observational research considered as having a minimal risk for the participants. This is why we sought the approval from the Data Protection Officer of Tecnalia (dpo@tecnalia.com), who ensured our study protocol was compliant with the GDPR, the H2020 ethics standards and the principles stated in the Declaration of Helsinki. The participants' right to refuse or withdraw from participating was fully maintained and the information provided by each participant was kept strictly confidential. To start participating in the study, participants had to read and sign the informed consent form, which explained the study's objective and conditions. After we determined the willingness and gained written consent from the participants, the required data were collected during the study.

5.3. Participants

Our lab-based study involved 30 participants (15 male, 15 female) from Tecnalia Research & Innovation, aged between 30 and 50, with varying levels of DC. All were

native Spanish speakers, and none encountered comprehension issues during the task. The majority were university graduates, with many holding master's degrees and four participants having doctoral qualifications.

While research in the literature has not extensively explored the optimal sample size for eye-movement studies [78], similar research has demonstrated that comparable sample sizes are adequate for identifying significant gaze patterns [77,97–99]. The sample size was determined based on methodological and practical considerations to ensure the identification of meaningful visual behavior patterns while maintaining resource efficiency. Our focus was to identify dominant patterns among high-performing participants, recognizing that a larger sample would be necessary to explore multiple patterns in depth.

In future research, we would recommend increasing the sample size following a stratified sampling, selecting a minimum sample for each of the 3 possible levels of DC according to DigComp (basic, medium, and advanced), i.e., at least $30 \times 3 = 90$ participants. In this way, we could explore multiple patterns in depth, not only of high-performing participants, but also of participants with a basic and medium level of DC.

Gender was not a selection criterion, and the age range (25 to 54) was chosen to reflect over 90% of BAIT service users. Participants self-reported their DC levels, with most rating themselves as "advanced" or "intermediate". However, in the netiquette domain, fewer participants identified as "advanced", and more placed themselves in the "basic" category, likely due to unfamiliarity with the term. Lastly, vision differences were not considered a significant source of variability in the study.

5.4. Materials

All participants individually completed the tests available on the online assessment tool. The 4 DCs selected were: (1) netiquette; (2) browsing, searching, and filtering data, information, and digital content; (3) evaluating data, information, and digital content; and (4) managing data, information, and digital content.

The details about the assessment tool, the dimensions selected for each DC and the type of items included in each test can be found in the previous article [27]. In order to measure not only low-order cognitive skills, the TEA provided different item formats to assess higher-order skills according to the medium and advance levels defined in DigComp—multiple-choice questions, interactive simulations, image/simulation-based questions, and open tasks. Furthermore, they were presented on one screen, necessitating a single-step response without the need for scrolling. The graphics consistently occupied the right half of the screen. Instructions and question statements were positioned in the upper right, with answer choices below. The "respond" button, situated in the lower-left corner of each item, facilitated saving results and progressing to the next question. Our aim was to minimize extraneous eye movements through a consistent layout of the screen. All items were displayed solely in Spanish and had to be responded to within the test application, in a controlled setting, without exiting the main interface. The participants received the items in a fixed sequence, and they were not permitted to delay addressing them or alter the order. This methodology was chosen due to the potential interrelation between certain questions. All interactions and attempts were monitored through the platform, with results automatically computed. Additionally, data on time spent per question and total test duration were recorded.

After gathering the data from the 30 participants, we opted to conduct a brief follow-up session, focusing on a selection of the most representative items to analyze the ET metrics. Among the various formats utilized in the tests, we chose the image/simulation-based tasks, wherein participants were asked to review and assess an illustration or simulation and then select the appropriate response. This format appears to be well-suited for assessing higher-order cognitive skills at the intermediate and advanced levels, as it challenges participants

to critically evaluate the presented scenarios. In addition, using an eye-tracker could aid in discerning whether users, in their responses, accurately assessed the predetermined areas outlined in the assessment criteria for each item. The items selected for this study and the assessment criteria defined for each item are shown in Table 1. Item4 was slightly different, as we asked users to click on the correct area after examining the image, rather than selecting the correct option from the list of possible choices.

Table 1. Items selected and assessment criteria.

Item	Lout	Format	Assessment Criteria
Item24	lout7	Image/Simulation-based	Review a bank account's ID page, noting any suspicious URLs to determine the site's reliability. Analyze visual cues to determine site credibility.
Item32	lout8	Image/Simulation-based	Review essential aspects of a news item on a website for reliability (URL, logo, author, date, etc.). By analyzing image visuals, determine if the news is fake, possibly accurate with anchor information, lacking anchor information, or clickbait.
Item4	lout1	Image/Simulation-based	Review the email fields, identify errors, and select the correct option. Verify all fields before choosing the incorrect one.

Figures 1–3 show the design of the 3 items selected (Item24, Item32 and Item4) and the AOIs defined.

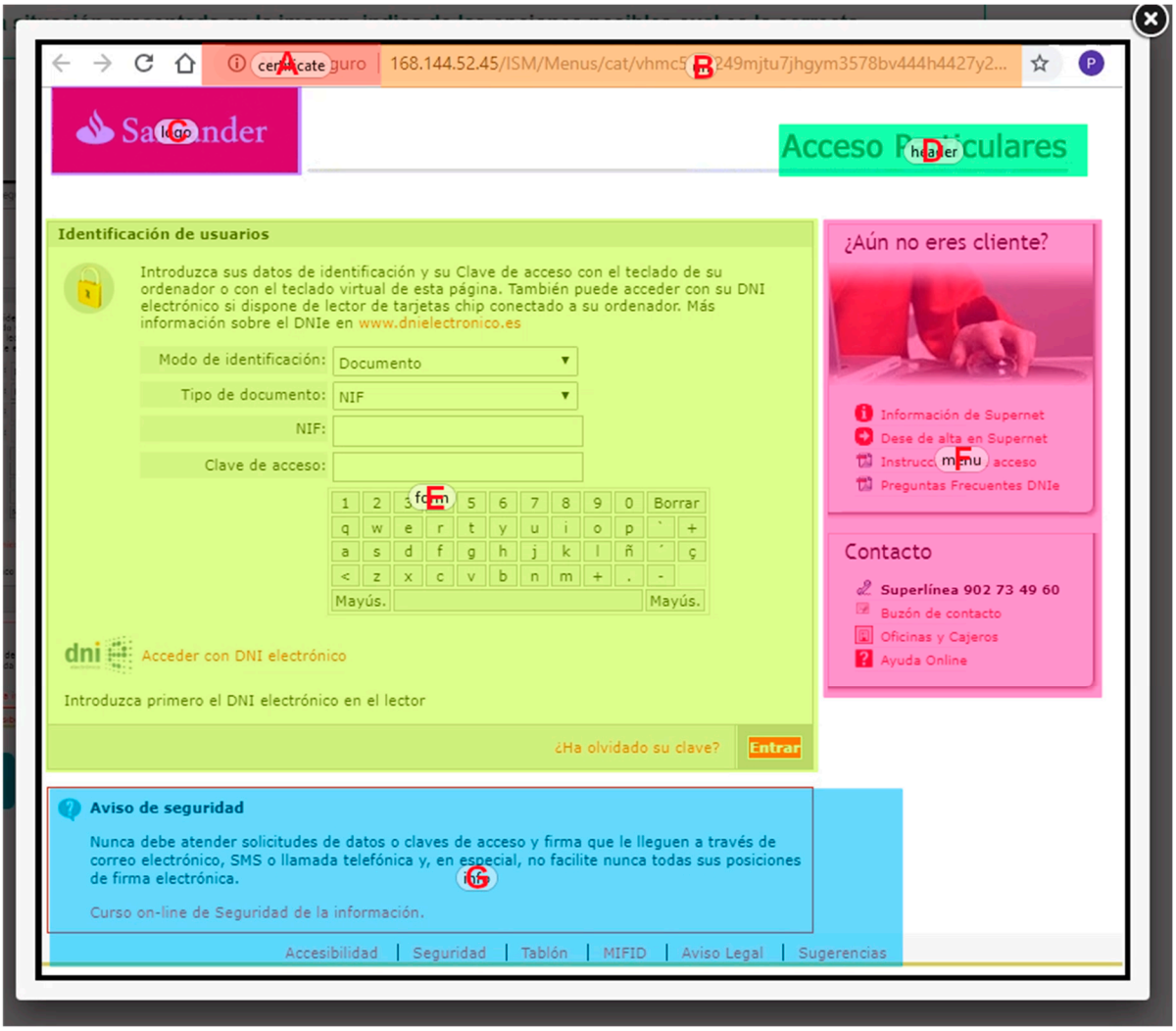

Figure 1. Design of Item24 showing an access page to an online bank in Spanish. The letters A–G correspond to the identifiers assigned to each defined area of interest.

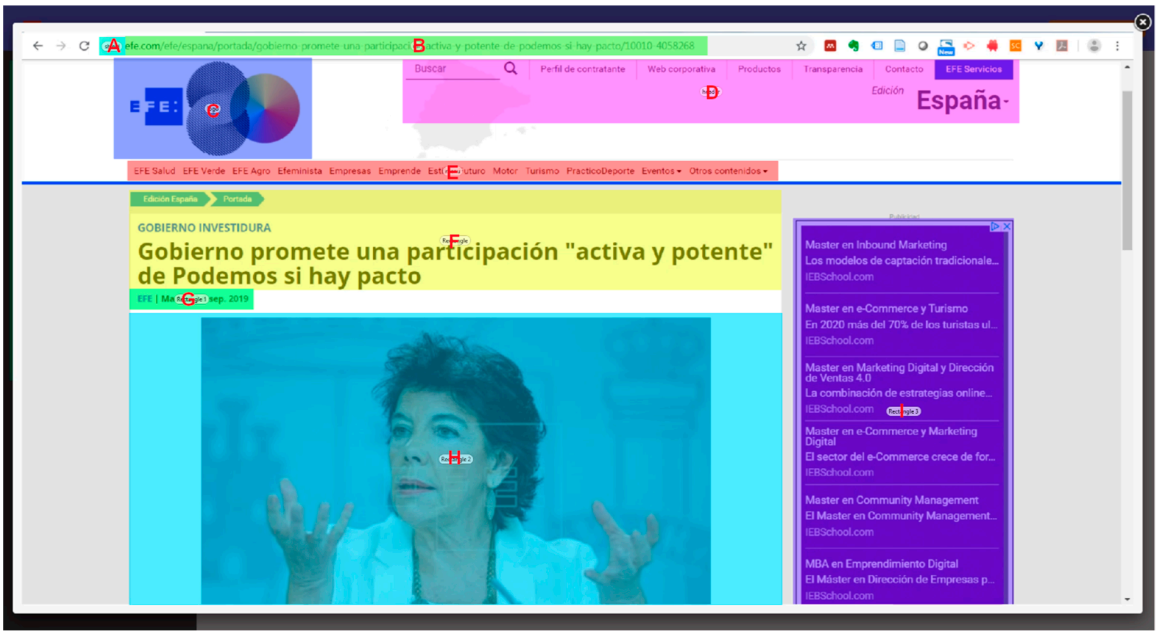

Figure 2. Design of Item32 showing a screenshot of a news article written in Spanish. The letters A–I correspond to the identifiers assigned to each defined area of interest.

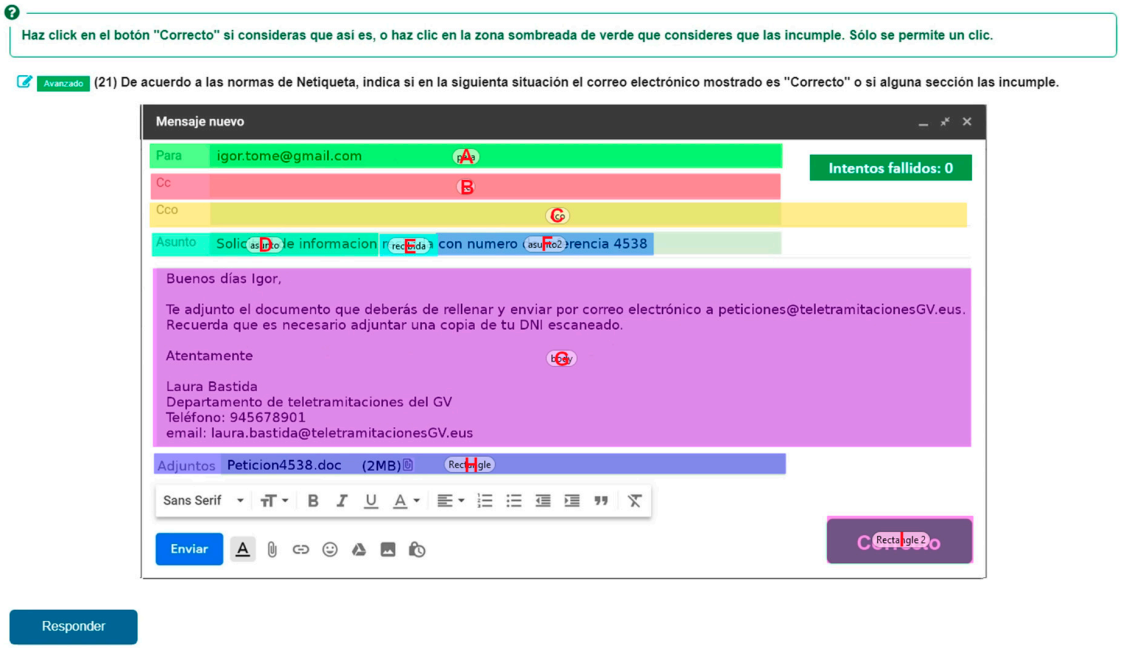

Figure 3. Design of Item4 showing a screenshot of an e-mail written in Spanish. The letters A–I correspond to the iden-tifiers assigned to each defined area of interest.

5.5. Experimental Equipment

The data were collected on a laptop in a Tecnalia meeting room throughout January 2020. The laptop was a MSI GS75 stealth i7 laptop and was located far enough away to avoid distractions caused by actions taken by the researcher. For the study, we employed Tobii Pro Lab (TPL) software version 1.130.24185, utilizing the integrated browser within the software to present the tests on a DELL e2310 23-inch monitor connected to a laptop. Eye movements of the participants were recorded with the Tobii X2-30 Eye Tracker, a discreet and standalone device designed for in-depth research on natural behavior. The eye-tracker, operating at a frequency of 30 Hz, was positioned at the base of a separate 17-inch monitor, which was used to monitor the participants' eye and head positions.

During tracking, the eye-tracker uses infrared illuminators to generate reflection patterns on the corneas of the participant's eyes. These reflection patterns, together with other visual data about the participant, are collected by image sensors. Sophisticated image processing algorithms identify relevant features, including the eyes and the corneal reflection patterns. Complex mathematics is used to calculate the 3D position of each eyeball and, finally, the gaze point (in other words, where the participant is looking). The timing and placement of mouse clicks were documented as part of the data collection.

Six distinct projects were created within the TPL framework for the purposes of this research. Each project commenced with a self-evaluation, during which participants indicated their perceived level of competence. Each stimulus was linked to an item previously selected in the tests. All tests and items were displayed in the same order. We selected two types of stimuli according to the item formats included, as follows: (1) Web stimulus, to display webpages to participants during a recording. TPL opens the website's URL in the built-in Lab browser, and automatically registers all mouse clicks, keystrokes, and webpages accessed. (2) Screen recording stimulus, to register all mouse clicks, keystrokes, programs, and webpages accessed, i.e., all activity displayed on the screen from the beginning to the end of the stimulus. In particular, the 3 items selected for this study were recorded using a Web stimulus.

For each stimulus, distinct AOIs and TOIs were established. The AOI is a concept used in TPL that allows the researcher to calculate quantitative eye movement measures. So, we drew a boundary encompassing the elements of the ET stimulus relevant to our study. We created AOIs in all the areas of the questions that we thought might attract their attention. TPL then calculates the metrics within the boundary over the time interval of interest. TOI is another TPL concept that provides a degree of analytical flexibility, allowing researchers to organize the recording data according to intervals of time during which meaningful behaviors and events take place. In simulations/image-based queries, we selected the intervals at which users viewed the image.

In summary, the main goal of the scan-path analysis was to understand how participants interacted with different stimuli and tasks, and how their attention was distributed across different areas and times. This was achieved through the use of ET and the TPL analysis tool.

5.6. Procedure

The participants were seated in front of the monitor for data collection, following which the ET system was individually calibrated, a process that took approximately three minutes. Subsequently, the items were presented in a consistent sequence for all participants. After completing each task, the researcher activated the next stimulus, advancing to the following item and ensuring that participants could not return to previous items. The eye-tracker recorded gaze data and click streams, which were supplemented by the information captured by the TEA. The I-VT (Fixation) filter was selected for exporting the data from the TPL with the velocity threshold by default, i.e., 30 degrees/s. The I-VT (Fixation) filter is based on the identification of fixations using the Velocity Threshold (VT) criterion. Fixations are defined as periods during which gaze velocity falls below a specified threshold, indicating stable fixation on a particular location. The I-VT filter provides a robust method for identifying fixations in ET data. By incorporating both velocity and spatial criteria, this method provides a robust and flexible approach to detecting fixations, enabling researchers to gain deeper insights into visual behavior.

6. Results

6.1. Evaluating Alternative Interpretations Based on the AOIs Examined

We began computing the Pearson coefficient (refer to Table 2), revealing a significant correlation between scores and question type r (87) = −0.389, $p < 0.01$, the RT r (87) = −0.313, $p < 0.01$, the scan-path length r (87) = −0.370, $p < 0.01$ and the number of AOIs to be checked

from the total number of AOIs defined and the scan-path length (AOIs rate) r (87) = 0.941, $p < 0.01$. That is, regarding the type of question, the item that did not require a systematic approach (all the AOIs needed to be examined) was more difficult (participants showed lower scores) than those that did require it, and participants also required longer scan-paths and RTs (i.e., the spent more time examining more AOIs, as if they lacked a sufficiently clear understanding of the key issues to be able to solve the task correctly). Furthermore, higher scores correlated with higher AOIs rates, meaning that participants showed a greater knowledge of the subject matter with more focus on the key AOIs than on the other AOIs (not relevant for the task resolution).

Table 2. Relationship between the scores and the type of question, RT, scan-path length and AOIs rate. Pearsons correlations (* $p < 0.05$ (bilateral); ** $p < 0.01$ (bilateral); r (87)).

	Item Type	AOIs Rate	Resp. Time	Score	Sc. Length
Item type	1	0.941 **	0.219 *	−0.389 **	0.317 **
(Sig)		0.000	0.042	0.000	0.003
AOIs rate	0.941 **	1	0.170	−0.378 **	0.211 *
(Sig)	0.000		0.116	0.000	0.050
Resp. time	0.219 *	0.170	1	−0.313 **	0.699 **
(Sig)	0.042	0.116		0.003	0.000
Score	−0.389 **	−0.378 **	−0.313 **	1	−0.370 **
(Sig)	0.000	0.000	0.003		0.000
Sc. length	0.317 **	0.211 *	0.699 **	−0.370 **	1
(Sig)	0.003	0.050	0.000	0.000	

As part of our analysis, we determined the critical AOIs necessary for successfully completing each task. Table 3 shows the proportion of participants that visited each of the AOI identified. Bolded values indicate invariant areas (visited either by none or by all the participants) that offer no information in the analytical process. A hyphen indicates that the AOI is missing in that task.

Table 3. Visit rate of the AOI within the target items. Bolded values indicate invariant areas (visited either by none or by all the participants) that offer no information in the analytical process.

	A	B	C	D	E	F	G	H	I
Item24	0.37	0.44	0.68	0.65	**1.00**	0.72	0.51	-	-
Item32	**0.00**	0.63	0.83	0.90	0.80	**1.00**	0.30	0.76	0.76
Item4	0.92	0.82	0.85	**1.00**	**1.00**	0.60	0.42	0.57	-

Each item has specific critical AOIs for resolution. Hence, the analysis must assess participants' visitation to these AOIs. If we had seen that these areas had hardly been considered in the resolution, that would have raised doubts regarding the validity the item. In Item24, all the AOIs were examined with different visit rates. In Item32, DC, B, C, E and F were critical, and all these AOIs had relevant visit rates. In this item, the B AOI was crucial, as one could not answer the question accurately without recognizing the invalidity of the URL. In Item4, D and E AOIs were critical, as the key to solving this item was a spelling mistake in the subject field of the email, and all the participants paid attention to these areas. Thus, data show that the critical AOI in each of the selected items was visited, and so, it is suitable to continue the analytical process.

Once we discarded the invariant features (bold values in Table 3), we performed a categorical classification of the participants' behavior, clustering the different behavioral patterns during the experiment, using the visits of each participant to each AOI as classification features.

We employed the OPTICS algorithm to determine an appropriate clustering regarding the participants' behavior. Tables 4 and 5 show the results of the clustering evaluation.

Table 4. Description of the clustering results for each item.

	Total Elements	Classified Elements	Unclassified Elements	Number of Clusters
Item24	29	15	14	5
Item32	30	16	14	5
Item4	28	23	5	6

Table 5. Element distribution along the identified clusters for each item.

	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5	Cluster #6
Item24	5	4	2	2	2	-
Item32	5	4	3	2	2	-
Item4	7	5	4	3	2	2

Even if clustering analysis is a valuable tool for identifying patterns, Tables 4 and 5 evidence the lack of clear clustering results. Table 4 shows the existence of several unclassified elements, especially for Items24 and 32. Table 5 shows that even the classified elements merge into very fragmented small size clusters. The absence of distinct clusters suggests a high degree of variability and complexity in participant responses. This variability indicates that participant behavior is multifaceted and may not conform to predefined patterns. This complexity could arise from various factors, such as individual differences or the variability of approaches engaged to overcome the tasks, consistent with the unguided nature of the experimental process.

Table 6 shows the results from the Kruskall–Wallis test regarding the efficiency of the behavioral pattern identified in the clustering process. The use of this test was preferred over the ANOVA test, as the normality of data cannot be assumed. The results indicate that the clustering process was not able to detect significant differences in behavioral patterns across clusters. However, considering that the process led to very fragmented small size clusters and many unclassified elements, there is no ground to establish that the participants' behavior was uniform among clusters.

Table 6. *p*-value of the Kruskall–Wallis test considering the partial result and the global performance.

	Partial Score	Overall Competence
Item24	0.616	0.602
Item32	0.844	0.530
Item4	0.685	0.542

The clustering analysis based solely on the visited AOI shows no significant ability to predict the outcome of each task. However, some items may be more predictable to these features than others, and show a greater correlation to certain AOIs. Looking to assess the predictability of the items based solely on the visited AOIs, we designed a predictive Decision Tree classifier and trained it over 15 random stratified splits. Figure 4 shows the results of this process. We can conclude that the distributions of the values for these items are relatively similar, with Item24 and Item32 exhibiting slightly higher central tendencies compared to Item4. Notably, the median values observed in the boxplot are relatively small, indicating that solely the visitation of AOIs is not a very efficient feature for use in predicting the outcome of the answer.

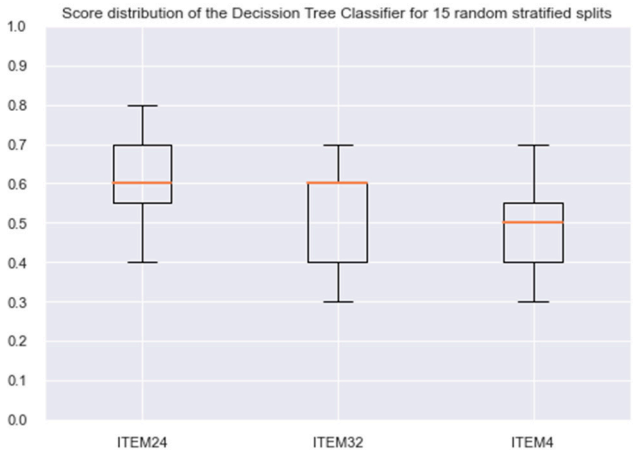

Figure 4. Accuracy distribution of the Decision Tree classifier among the stratified splits.

To further clarify the classifier's performance, we present Table 7, with detailed metrics including accuracy, precision, recall, and F1 score. These metrics provide a comprehensive overview of the classifier's effectiveness across different items. The performance across the items shows notable divergence. Item24 demonstrates relatively higher recall compared to Item32, indicating a stronger ability to identify positives. However, Item32 exhibits slightly higher precision and F1 score than Item24, suggesting better balance between precision and recall despite a slightly lower accuracy. In contrast, Item4 shows significantly lower precision and recall than both Item24 and Item32, resulting in a markedly lower F1 score and accuracy. These variations highlight distinct levels of classifier effectiveness and performance discrepancies across the different items.

Table 7. Mean performance metrics of the classifier among the studied items.

	Accuracy	Precision	Recall	F1
Item24	0.61	0.63	0.81	0.69
Item32	0.58	0.64	0.73	0.67
Item4	0.49	0.27	0.25	0.25

To determine if the differences among items are statistically significant, we performed a Kruskal–Wallis test. The test yielded a p-value of 0.0068, indicating that there is significant evidence to suggest differences between the items. The discovered statistical differences in predictability among tasks underscore the variability in the effectiveness of predictive models across different tasks. This suggests that certain tasks may exhibit a stronger association between participants' visitation AOIs and successful task performance compared to others. For instance, based on the score distribution obtained with the trained classifier, Item24 demonstrates higher predictability than the other tasks, with a median prediction accuracy of 0.60 and consistent performance across instances. These findings highlight the task-specific nuances in participant behavior, which reflect the complexity of the tasks themselves, including factors such as task difficulty, cognitive demands, and the clarity of instructions. Understanding these task-specific nuances can lead to the design of more effective items.

In summary, the visitation of AOIs by itself is not sufficient to predict the results of the interaction with each item. As we employed suitable statistical approaches, this result does not imply any issues with the procedure; rather, it points out that the behavioral patterns might be too complex or diverse for simple classification. Even so, there is enough evidence to show that the outcome of some items is more sensitive to the visited AOI than others.

6.2. Evaluating Alternative Interpretations Based on the Order of the AOIs Examined

We started analyzing the variance within the groups to reach some conclusions related to the type of resolution required for each question (systematically or not). According to our analysis, the means of changes necessary based on the Levenshtein method are lower for the successful group than for the unsuccessful group (see Table 8). In essence, successful individuals exhibit greater mean similarity than unsuccessful ones, indicating higher variance within the latter compared to the former.

Table 8. Mean (and standard deviation) of changes necessary according to the Levenshtein method for Image/simulation-based questions.

Item	Pass	Fail
Item24	9.5 (4.8) (n = 24)	10.3 (3.8) (n = 6)
Item32	15.2 (7.7) (n = 24)	30.3 (13.3) (n = 6)
Item4	16.9 (7.1) (n = 12)	18.6 (8.9) (n = 18)

To clearly illustrate the high variance within the unsuccessful participants group, we used the ScanGraph tool. As representative examples, Figures 5–7 illustrate the advised graph (a graph with 5% of the possible edges) for the successful group participants and the unsuccessful group participants for the different items selected. This graph illustrates that the successful group's scan-paths for Item24 and Item32 are more interconnected, indicating greater similarity compared to the unsuccessful group. For Item4, the unsuccessful group's scan-paths exhibit greater connectivity, possibly reflecting the need to scrutinize all email fields due to question characteristics. Participants who successfully answered the question had shorter scan-paths, which can be understood as they identified the wrong field earlier, while those who failed the question had a longer review of all the fields, performing a more similar strategy. For Item24 and Items32, participants had to examine the reliability of the identification page of a bank and a news item published on a website, respectively. In both cases, these types of items required us to check only some key parts in a systematic way. Participants did not have to examine all possible elements, resulting in greater similarities in the scan-paths of participants who successfully answered these items. The details of the scan-paths for each participant per item with the original, the adjacency matrix and the similarity groups can be found in the Supplementary Materials.

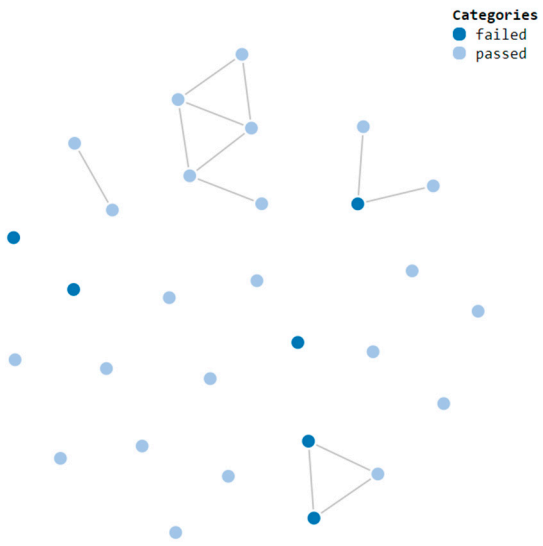

Figure 5. Item24 scan-path similarity graphs between successful and unsuccessful participants.

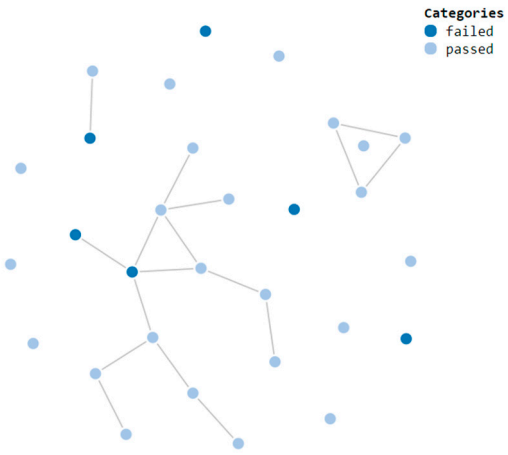

Figure 6. Item32 scan-path similarity graphs between successful and unsuccessful participants.

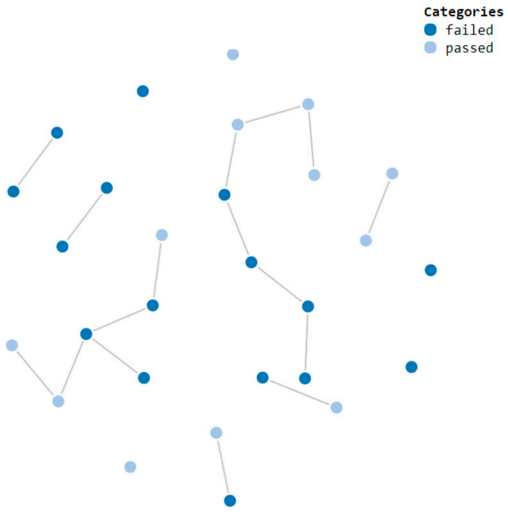

Figure 7. Item4 scan-path similarity graphs between successful and unsuccessful participants.

To examine the scan-paths of the successful participants so as to identify the common scan-path, we applied a position-based weighted model. Tables 9–11 show the sequences of the visual elements for all the scan-paths.

Table 9. Item4, results of applying the position-based weighted model.

SCAN-PATH	A	B	C	D	G	H
DABCHG	52	43	40	55	12	29

Table 10. Item24, results of applying the position-based weighted model.

SCAN-PATH	A	C	D	E	F	G
ECGDA	55	67	63	106	0	66

Table 11. Item32, results of applying the position-based weighted model.

SCAN-PATH	A	C	D	E	F	H	I
FCDAEHI	83	99	99	74	141	63	56

Figures 8–10 show the common scan-paths identified for each item. In Item4, participants made their first fixation in the email subject field. It seems that this AOI was key

for them to contextualize the email; for example, to find out whether it was a formal or informal email. Moreover, this is the field containing a spelling mistake, and it therefore does not comply with the rules of netiquette. Then participants continued examining the rest of the fields by taking a top-down approach, except for the AOI related to the content of the message, which was left for the last place. It is notable how participants deferred addressing this AOI until the end, placing minimal importance on the message content when applying the rules of netiquette. So, the approach followed by the participants was correct according to the assessment criteria defined for this item. In this sense, we could see how participants prioritized the forms more than the contents, both being equally relevant for solving the task. This could be since netiquette is something relatively new that requires more time to be assimilated. In Item24, participants made their first fixation in the main area of the webpage, where a user identification form is displayed. Then, participants continued examining the logo of the website, with the AOI showing a security warning and the header at the top right indicating that this website shows private access to a bank account. It seems that these AOIs were key for them to contextualize the website. In some ways, the behavior here was similar to with the previous item, but required examining more areas to better contextualize the task. This behavior was expected according to the evaluation criteria established for this question. Hereafter, participants continued examining the AOI related to the URL of the website. This AOI is crucial for task success. Participants must verify the website's valid web certificate and ensure the domain is not suspicious based on the visited bank. In Item32, participants made their first fixation in the header of the news. Then, participants continued examining the logo of the website and the header at the top right, including the menu of the journal. It seems that these AOIs were key for them to contextualize the website. The behavior closely resembled that of participants in Item24. Hereafter, participants continued examining the AOI related to the URL of the website. Participants should check if the website has a valid web certificate, and that the domain is valid according to the journal visited. Finally, participants continue examining the rest of the fields by taking a top-down approach. The behavior was similar to with the previous items, where participants started examining the key areas to contextualize the task. Then, they focused on the key areas to solve the task, as was defined in the assessment criteria for this item. In short, we could not identify an alternative interpretation of the results that would have underlined the evaluation criteria defined for these items.

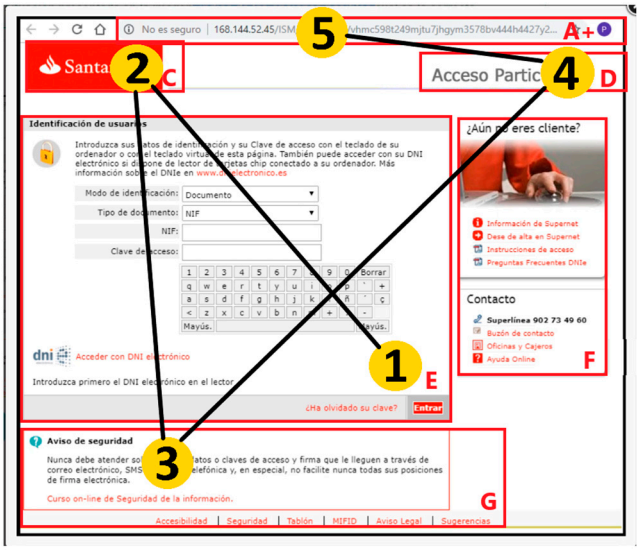

Figure 8. Common scan-paths identified for Item24 showing an access page to an online bank in Spanish. The letters A–G correspond to the identifiers assigned to each defined area of interest and the numbers 1–5 correspond to the visiting order followed.

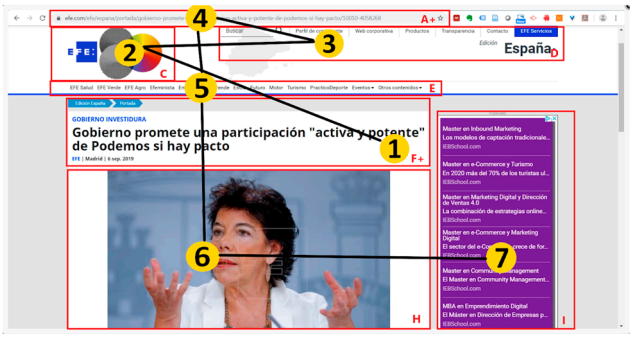

Figure 9. Common scan-paths identified for Item32 showing a screenshot of a news article written in Spanish. The letters A–I correspond to the identifiers assigned to each defined area of interest and the numbers 1–7 correspond to the visiting order followed.

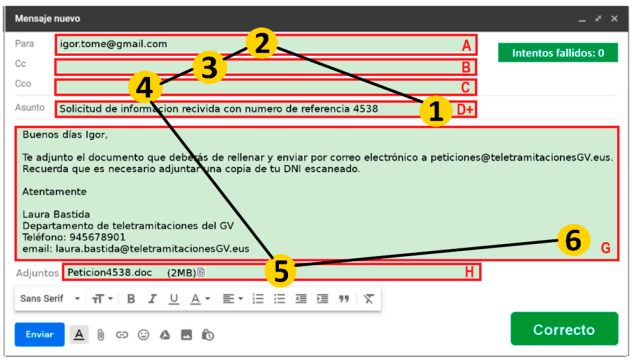

Figure 10. Common scan-paths identified for Item4 showing an e-mail written in Spanish. The letters A–H correspond to the identifiers assigned to each defined area of interest and the numbers 1–6 correspond to the visiting order followed.

Finally, we analyzed fixation lengths and their predominant locations. As shown in Table 12, Item32 had the shortest durations of fixations and Item24 had the longest.

Table 12. Mean (and standard deviation) of fixation durations for image/simulation-based questions.

Item	Fixation Durations
Item24	265.7 (195)
Item32	234.3 (155.7)
Item4	246.4 (177.1)

We examined which places attracted longer fixations to check if the behavior shown by the participants was expected according to the assessment criteria defined for each item (see Table 13). We thought that this information could be interesting to complement the information from the scan-paths followed, and to check if long fixations occurred in areas that were meaningless for the resolution of the task.

Table 13. Mean (and standard deviation) of fixation durations for each area of interest.

AOI	Item24	Item32	Item4
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)	

In Item24, participants exhibited lengthier fixations on AOIs associated with the website certificate and URL, with similar extended fixations observed on AOIs related to the entry form and security information area. In Item32, participants exhibited extended fixations on AOIs linked to the author and publication date, with similar prolonged fixations observed on AOIs associated with the website certificate and URL. In Item4, fixation durations were more comparable, as participants were required to evaluate all AOIs thoroughly to form conclusions. AOIs corresponding to the CC and attachment fields exhibited slightly longer fixations, while those associated with the subject line and email content field received slightly shorter fixations. To resolve Item24 and Item32, participants followed a systematic approach, while Item4 demanded a comprehensive examination across all domains.

The fixation durations distribution confirmed the established assessment criteria for the questions, with participants exhibiting expected behavior. Anomalies, such as prolonged durations in challenging areas, were noted. Longest fixations occurred at critical points. The fixation durations for Item24 and Item32 reflected a systematic approach, whereas Item4 exhibited a more uniform distribution, suggesting the comprehensive examination of all areas. These findings are consistent with those from previous scan-path analyses, and supported the research questions addressed.

7. Conclusions

Recent studies have sparked a lively debate regarding the roles various types of process data play in evaluating the performance and validity of test examinees [34,35]. Previous studies, such as those by D'Mello et al. and Azevedo [38,100], have emphasized the need to bring together different sources, such as log files, ET, emotion recognition, and think-aloud protocols, to better understand RPs. This multi-source approach is crucial for enhancing the robustness of validity arguments. For example, Ercikan and Pellegrino highlighted the importance of integrating process data to validate complex assessments [34], while Azevedo demonstrated how ET and log data can provide complementary insights into cognitive and emotional states during test-taking [38]. Despite its relevance, studies that have incorporated the analysis of RPs remain scarce in assessment domains focused on validating the design of the information structure, rather than the content being assessed, and are practically non-existent in DC assessment. By integrating findings from these studies, our work extends the understanding of how process data like log data, ET data, scores, and item response times can contribute to validating inferences made between claims and observed behaviors. The primary objective of our exploratory study was to illustrate specific applications of ET data in enhancing the validity of inferences drawn from scores. Our focus on evaluating participant interactions with assessed content aimed to determine if observed behaviors aligned with predefined assessment criteria for each item. This approach builds on the works of Greiff et al., who explored how response time analysis can reveal engagement and problem-solving strategies [47], and of Kroehne and Goldhammer, who demonstrated the benefits of combining response time with other process data to understand test-taker behavior [48]. Moreover, our findings extend the insights of Van Gog and Scheiter, who discussed the potential of ET to uncover cognitive processes in problem-solving tasks by applying similar methodologies to the domain of DC assessment [101]. This comparative analysis underscores the value of ET data in providing a nuanced understanding of test-taker behaviors, complementing traditional measures of assessment.

We explored specific patterns of participants' eye movement to make detailed observations of item RPs in an evaluation of DC. Specifically, we used ET observations to fill the 'explanatory gap' by providing data on the variation in item RPs that are not captured in

traditional TEAs. We focused on generating and testing inferences about the RPs performed by the participants, evaluating an alternative interpretation of the test scores in terms of AOIs examined and the order followed in examining the different AOIs.

In accordance with observations by other researchers, like Cronbach, stating that "A proposition deserves some degree of trust only when it has survived serious attempts to falsify it", we focused on evaluating an alternative interpretation of the test scores [30]. We tried to falsify the proposition that the items triggered the knowledge and skills required for solving the image/simulation-based items by testing an alternative interpretation: that test-takers do not pay attention to the key areas, or that they followed a meaningless order of fixations to answer the items correctly; that is, solving the tasks in a way that is inconsistent with expectations for each item according to the assessment criteria previously defined.

To evaluate this alternative interpretation of the test scores, we used ET technology to examine the scan-paths. We investigated how participants processed the different AOIs defined in order to evaluate the situation and choose the correct answer. We tried to answer two questions, as follows:

First, which AOIs within the items were examined, and whether specific patterns of AOIs visited might undermine the claim that the items required the same cognitive processes that are required in real-world tasks, according to the assessment criteria previously defined. We could not identify alternative patterns in terms of visit rates of AOIs that might undermine the assessment criteria defined. All the AOIs were examined with different visit rates, but we could not find any response pattern with an unexpected visit rate that was difficult to explain, which would have raised doubts as to whether the question was generating the expected RPs. Additionally, we carried out a clustering of the responses in terms of AOIs visited, applying an unsupervised OPTICS classification algorithm to examine whether specific patterns of AOIs visited to solve the task predicted higher success rates in the overall performance (the global score considering tasks related to the same DC of task in question), and in a way that is not inconsistent with expectations for each item. We could identify clusters with higher success rates, but nevertheless, the results were far from significant, and thus, it was not possible to conclude that any of the behavioral patterns were more efficient than the rest. It would be very interesting to carry out this analysis again, with a higher number of participants.

Secondly, we examined the scan-paths of the items in more depth, and considered the order of processing the different AOIs when solving them. We employed the Levenshtein method, facilitated by the ScanGraph tool, to ascertain that the unsuccessful group exhibits greater variance than the successful group. The significant variability and lack of clear patterns in participant behavior, as suggested by the clustering analysis and the variability within groups in the Levenshtein method analysis, might reflect the complex nature of DC assessment. In order to explain this variability, it would be interesting to incorporate additional variables not accounted for in this analysis (e.g., individual differences in cognitive strategies). Then, we went deeper into cognitive processes and problem-solving strategies, calculating the common scan-paths of only the successful participants whose variance was smaller. Despite the limitations of the position-based weighted model applied, we obtained interesting insights into the performances of the successful participants, which could not have been obtained in traditional TEAs. As a result, we acquired a visual representation enabling us to easily verify that participants who answered the question correctly did not exhibit any unexpected behavior that would invalidate the item. We consider this information helpful to examinees who incorrectly responded. If they wanted to review the question for the correct answer, we could show them the common scan-path as an extra value. To our knowledge, current DC assessments do not offer this functionality for such items during review. However, in the meta-analysis performed by Xie et al., it was

pointed out that the use of eye movement modeling examples was beneficial to learners' performance when non-procedural tasks were used instead of procedural tasks [102]. So, it should be further investigated to what extend the use of modeling examples in the review process of an assessment of DC might be appropriate, depending on the type of question. Finally, we also examined the fixation durations and the places where longer fixations occurred. We could confirm that successful participants undertook the longest fixations on the expected areas, and we could not identify an alternative interpretation that would undermine the assessment criteria defined.

Additionally, the fixation durations' distribution validated the item design, revealing two distinct approaches—systematic and non-systematic. While the systematic approach focuses on specific areas, the non-systematic approach uniformly examines all areas, as indicated by the homogeneous fixation durations distribution.

Previous studies used ET data in fields related to the DCs selected. For example, fixation patterns are relevant in the rapid analysis and recognition of information [64,67,103], and are helpful for understanding user behavior in search engines [68], or to show the differences in the strategies that participants used to answer the multiple-choice questions in the code comprehension and recall tasks [104]. These studies have yielded valuable insights into the cognitive mechanisms underlying rapid information analysis and recognition, highlighting the importance of context, task demands, and individual differences in shaping fixation behavior. Individual differences in eye movement behavior, such as age, expertise, and cognitive abilities, can influence fixation patterns, complicating the generalization of findings across diverse populations. Furthermore, some of these studies also mentioned some methodological limitations, such as the spatial and temporal resolution of ET devices, which can constrain the accuracy and precision of fixation data, potentially leading to misinterpretation or oversimplification of cognitive processes.

As far as we know, only Yaneva et al. used ET data for validation purposes to examine how the distributions of the options in multiple-choice questions influenced the way examinees responded to the questions [39]. Nevertheless, to our knowledge, prior research has not addressed the evaluation of DCs for validating RPs. We showed different ways of using data collected on fixation areas and density, which could be used to support validation practice and test development. Additionally, we have also shown that metrics from ET data were of great value for test designers in terms of understanding examinees' interactions with the tests and helping them to adjust the level of difficulty.

When interpreting the findings of this study, it is important to acknowledge certain limitations, including the relatively small participant sample and the limited number of items used. Future research would benefit from examining whether these outcomes are consistent with a larger sample that also includes individuals with a fundamental level of DC, or by employing additional items with a similar structure. Furthermore, subsequent studies should explore items with varied response formats, which might incorporate more intricate RPs. Zumbo et al. recently highlighted the need to be more rigorous when validating the way that RPs and their data are processed as measurement opportunities [105]. ET studies can assist the process of constructing a validity argument. However, RP data can more clearly challenge an interpretation than directly support it [31]. In addition, further research would be recommended. For example, with a larger participant sample, we may discover that a single common scan-path is inadequate, and multiple common scan-paths may prove valid. This could indicate different participant cohorts, such as those with advanced levels of DC.

We are also aware that while scan-paths and fixation durations offer valuable insights into cognitive processing, they may not capture the full complexity of examinees' cognitive processes. Several factors can influence these metrics and affect their interpretation, such

as task complexity, examinee characteristics, environmental distractions, or test anxiety. Although during the study we tried to minimize these factors, such as reducing distractions, we should be aware of their existence and influence.

Furthermore, additional research is recommended to better understand the efficacy of displaying the common scan-path of successful participants during the review. Implementing this suggestion poses challenges, such as determining which common scan-path to display if multiple valid paths are identified for an item.

While our study is exploratory in nature and primarily focuses on utilizing ET data to enhance the validity of inferences drawn from test scores, it is essential to situate our findings within the broader context of existing research on DC assessment and the use of process data. Previous studies, such as those by Law et al. and Siddiq et al. [3,4], have highlighted significant gaps in the validity and reliability of DC assessment tools, particularly those relying heavily on self-reported measures. Our findings extend this body of work by demonstrating that ET data can provide detailed insights into the RPs of examinees, thereby addressing some of these validity concerns. Furthermore, our results align with the conclusions of Papamitsiou and Economides, who emphasized the need for more granular analyses of assessment and learning design strategies [106]. By integrating ET data, we contribute to a more nuanced understanding of how examinees interact with assessment items, corroborating the findings of Scherer et al. on the benefits of TEA environments [12]. Additionally, our study contrasts with the work of Rienties and Toetenel, who primarily focused on log data and response times [15]; we show that ET data can uncover variations in response strategies that are not captured by these traditional metrics. Overall, our research underscores the potential use of ET as a complementary tool in DC assessment, providing richer and more actionable data compared to conventional methods, and paving the way for future studies to further explore this promising avenue. To effectively integrate and analyze ET data with other process data such as log data and RTs, it would be advisable to use a multi-modal data fusion framework that leverages machine learning algorithms and statistical models. The integration process would begin with data synchronization, where timestamps from different data sources should be aligned to ensure temporal consistency. ET data would be processed alongside log data capturing user interactions and RTs that indicate cognitive load and decision-making speed. For the analysis, certain approaches could be employed, e.g., a combination of Dynamic Time Warping (DTW) for aligning temporal sequences and Principal Component Analysis (PCA) to reduce dimensionality, while preserving key patterns in the data. To explore relationships and predictive capabilities, machine learning models could be utilized, such as Random Forests and Support Vector Machines (SVM) for classification and regression tasks, respectively. Additionally, Hidden Markov Models (HMM) could be used to identify latent states in user behavior. The integration of these data streams could help to provide a comprehensive view of user behavior, improving the understanding of cognitive processes and enhancing predictive analytics in human–computer interaction studies.

Our study's findings can be further situated within established theoretical frameworks of cognitive processing and test validity. According to theories of cognitive processing, such as those posited by Ercikan and Pellegrino and Kane and Mislevy [31,34], ET data provide a robust method to infer cognitive strategies and processes involved in test-taking. These theories suggest that RPs are crucial for validating the cognitive constructs that assessments aim to measure. The alignment of fixation patterns and scan paths with the hypothesized cognitive processes supports the validity of the test items and the inferences drawn from test scores. Additionally, the use of ET in our study aligns with Cronbach's validation theory [30], which emphasizes the necessity of rigorous validation methods to substantiate the intended interpretations of test scores. By examining the detailed eye movement data, we have

demonstrated that participants engage with test items in ways that reveal underlying cognitive strategies, thus reinforcing the construct validity of our DC assessment. This integration of ET data into the validation framework not only provides empirical support for the intended constructs, but also extends the applicability of cognitive processing theories in the context of digital assessments. Although we think that the results presented are useful and can make an important contribution to the process of constructing a validity argument in a DC assessment, they are limited because ET studies are time-consuming, and for this reason, the participant samples tend to be limited. Even more so, there are more metrics available that should be investigated and which might contribute to furthering the understanding of participants' task-solving behaviors. The results of this research could be considered when thinking about how to use ET data to validate the design of items when measuring complex cognitive constructs such as DC. Studies like this can be part of a complete validity argument.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app15031215/s1, Table S1: Description of everyone's self-report of their DC level, as well as the results obtained in the tests and the descriptors for each proficiency level; SCANGRAPH_Item4.xlsx, SCANGRAPH_Item24.xlsx and SCANGRAPH_Item32.xlsx: Details of the scan-paths for each participant for the items Item4, Intem24 and Item32, with the original, the adjacency matrix and the similarity groups.

Author Contributions: Conceptualization, J.B.; Methodology, J.B. and P.G.; Software, J.B., E.L. and L.B.; Validation, J.B. and P.G.; Investigation, J.B., E.L. and L.B.; Resources, J.B.; Data curation, J.B., P.G. and E.L.; Writing—original draft, J.B.; Writing—review & editing, P.G.; Visualization, E.L.; Supervision, P.G. and L.B.; Project administration, L.B. All authors equally contributed to writing and reviewing this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The research did not involve biomedical interventions, biological samples, or procedures that posed physical or psychological risks to participants. According to Spanish Law 14/2007 on Biomedical Research, ethical approval from a committee was not legally required for this type of study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The research team thanks the individuals who generously shared their time, experience, and materials for the purposes of this project.

Conflicts of Interest: Authors Juan Bartolomé, Erlantz Loizaga and Leire Bastida were employed by the company TECNALIA. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Iñiguez-Berrozpe, T.; Boeren, E. Twenty-first century skills for all: Adults and problem solving in technology rich environments. *Technol. Knowl. Learn.* **2020**, *25*, 929–951. [CrossRef]
2. Kluzer, S.; Priego, L.P. *Digcomp Into Action: Get Inspired, Make it Happen. A User Guide to the European Digital Competence Framework (No. JRC110624)*; Joint Research Centre: Seville, Spain, 2018.
3. Law, N.; Woo, D.; Wong, G. *A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2*; UNESCO: Paris, France, 2018; p. 146.
4. Siddiq, F.; Hatlevik, O.E.; Olsen, R.V.; Throndsen, I.; Scherer, R. 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. *Educ. Res. Rev.* **2016**, *19*, 58–84. [CrossRef]

5. Ala-Mutka, K. *Mapping Digital Competence: Towards a Conceptual Understanding*; Institute for Prospective Technological Studies: Seville, Spain, 2011; pp. 7–60.
6. Spante, M.; Hashemi, S.S.; Lundin, M.; Algers, A. Digital competence and digital literacy in higher education research: Systematic review of concept use. *Cogent Educ.* **2018**, *5*, 1519143. [CrossRef]
7. Ferrari, A.; Punie, Y. *DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe*; Publications Office of the European Union: Luxembourg, 2013.
8. Stödberg, U. A research review of e-assessment. *Assess. Eval. High. Educ.* **2012**, *37*, 591–604. [CrossRef]
9. Bartolomé, J.; Garaizar, P.; Larrucea, X. A Pragmatic Approach for Evaluating and Accrediting Digital Competence of Digital Profiles: A Case Study of Entrepreneurs and Remote Workers. *Technol. Knowl. Learn.* **2022**, *27*, 843–878. [CrossRef]
10. Kruger, J.; Dunning, D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personal. Soc. Psychol.* **1999**, *77*, 1121. [CrossRef]
11. Binkley, M.; Erstad, O.; Herman, J.; Raizen, S.; Ripley, M.; Miller-Ricci, M.; Rumble, M. Defining twenty-first century skills. In *Assessment and Teaching of 21st Century Skills*; Springer: Dordrecht, The Netherlands, 2012; pp. 17–66. [CrossRef]
12. Scherer, R.; Meßinger-Koppelt, J.; Tiemann, R. Developing a computer-based assessment of complex problem solving in Chemistry. *Int. J. STEM Educ.* **2014**, *1*, 2. [CrossRef]
13. Redecker, C.; Johannessen, Ø. Changing assessment—Towards a new assessment paradigm using ICT. *Eur. J. Educ.* **2013**, *48*, 79–96. [CrossRef]
14. Nguyen, Q.; Rienties, B.; Toetenel, L.; Ferguson, R.; Whitelock, D. Examining the designs of computer-based assessment and its impact on student engagement, satisfaction, and pass rates. *Comput. Hum. Behav.* **2017**, *76*, 703–714. [CrossRef]
15. Rienties, B.; Toetenel, L. The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Comput. Hum. Behav.* **2016**, *60*, 333–341. [CrossRef]
16. Hämäläinen, R.; De Wever, B.; Nissinen, K.; Cincinnato, S. What makes the difference–PIAAC as a resource for understanding the problem-solving skills of Europe's higher-education adults. *Comput. Educ.* **2019**, *129*, 27–36. [CrossRef]
17. Von Davier, A.A.; Zhu, M.; Kyllonen, P.C. (Eds.) *Innovative Assessment of Collaboration*; Springer: Cham, Switzerland, 2017.
18. Bennett, R.E. Formative assessment: A critical review. *Assess. Educ. Princ. Policy Pract.* **2011**, *18*, 5–25. [CrossRef]
19. Greiff, S.; Wüstenberg, S.; Avvisati, F. Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* **2015**, *91*, 92–105. [CrossRef]
20. Timmis, S.; Broadfoot, P.; Sutherland, R.; Oldfield, A. Rethinking assessment in a digital age: Opportunities, challenges and risks. *Br. Educ. Res. J.* **2016**, *42*, 454–476. [CrossRef]
21. Oranje, A.; Gorin, J.; Jia, Y.; Kerr, D. Collecting, Analyzing, and Interpreting Response Time, Eye-Tracking, and Log Data. In *Validation of Score Meaning for the Next Generation of Assessments*, 1st ed.; Routledge: Abingdon, UK, 2017; pp. 39–51. [CrossRef]
22. Saltos-Rivas, R.; Novoa-Hernández, P.; Serrano Rodríguez, R. On the quality of quantitative instruments to measure digital competence in higher education: A systematic mapping study. *PLoS ONE* **2021**, *16*, e0257344. [CrossRef]
23. Gegenfurtner, A.; Lehtinen, E.; Säljö, R. Expertise and the Sequential Organization of Video Game Play: Insights from Eye-Tracking. *Front. Psychol.* **2020**, *11*, 1684. [CrossRef]
24. Just, M.A.; Carpenter, P.A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **1980**, *87*, 329. [CrossRef]
25. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372. [CrossRef]
26. Greiff, S.; Wüstenberg, S.; Csapó, B.; Demetriou, A.; Hautamäki, J.; Graesser, A.C.; Martin, R. Domaingeneral problem solving skills and education in the 21st century. *Educ. Res. Rev.* **2014**, *13*, 74–83. [CrossRef]
27. Bartolomé, J.; Garaizar, P.; Bastida, L. Validating item response processes in digital competence assessment through eye-tracking techniques. In Proceedings of the Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain, 21–23 October 2020; pp. 738–746. [CrossRef]
28. Bashir, S.; Miyamoto, K. *Digital Skills: Frameworks and Programs*; World Bank: Washington, DC, USA, 2020; Available online: https://openknowledge.worldbank.org/handle/10986/35080 (accessed on 3 March 2022).
29. Duchowski, A.T. *Eye Tracking Methodology: Theory and Practice*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2017. [CrossRef]
30. Cronbach, L.J. Validity on parole: How can we go straight. *New Dir. Test. Meas.* **1980**, *5*, 99–108.
31. Kane, M.; Mislevy, R. Validating score interpretations based on response processes. In *Validation of Score Meaning for the Next Generation of Assessments*; Routledge: Abingdon, UK, 2017; pp. 11–24.
32. Fitts, P.M.; Jones, R.E.; Milton, J.L. Eye movements of aircraft pilots during instrument-landing approaches. *Aeronaut. Eng. Rev.* **1950**, *3*, 56.
33. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.
34. Ercikan, K.; Pellegrino, J.W. (Eds.) *Validation of Score Meaning for The Next Generation of Assessments: The Use of Response Processes*; Taylor & Francis: Abingdon, UK, 2017. [CrossRef]

35. Hubley, A.M.; Zumbo, B.D. Response processes in the context of validity: Setting the stage. In *Understanding and Investigating Response Processes in Validation Research*; Springer: Cham, Switzerland, 2017; pp. 1–12. [CrossRef]
36. Cronbach, L.J. *Essentials of Psychological Testing*, 2nd ed.; Harper & Brothers: New York, NY, USA, 1949.
37. Zumbo, B.D.; Hubley, A.M. *Understanding and Investigating Response Processes in Validation Research*; Springer International Publishing: Cham, Switzerland, 2017; Volume 26.
38. Azevedo, R. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educ. Psychol.* **2015**, *50*, 84–94. [CrossRef]
39. Yaneva, V.; Clauser, B.E.; Morales, A.; Paniagua, M. Using Eye-Tracking Data as Part of the Validity Argument for Multiple-Choice Questions: A Demonstration. *J. Educ. Meas.* **2021**, *54*, 515–537. [CrossRef]
40. Specian Junior, F.C.; Santos, T.M.; Sandars, J.; Amaral, E.M.; Cecilio-Fernandes, D. Identifying the response process validity of clinical vignette-type multiple choice questions: An eye-tracking study. *Med. Teach.* **2023**, *45*, 845–851. [CrossRef] [PubMed]
41. Kraitzek, A.; Förster, M. Measurement of Financial Competence—Designing a Complex Framework Model for a Complex Assessment Instrument. *J. Risk Financ. Manag.* **2023**, *16*, 223. [CrossRef]
42. Baxter, G.P.; Glaser, R. Investigating the cognitive complexity of science assessments. *Educ. Meas. Issues Pr.* **1998**, *17*, 37–45. [CrossRef]
43. De Boeck, P.; Jeon, M. An overview of models for response times and processes in cognitive tests. *Front. Psychol.* **2019**, *10*, 102. [CrossRef] [PubMed]
44. Lee, Y.H.; Hao, J.; Man, K.; Ou, L. How do test takers interact with simulation-based tasks? A response-time perspective. *Front. Psychol.* **2019**, *10*, 906. [CrossRef]
45. Messick, S. Validity. In *ETS Research Report Series*; Wiley: Hoboken, NJ, USA, 1987; p. i-208.
46. Ercikan, K.; Guo, H.; He, Q. Use of response process data to inform group comparisons and fairness research. *Educ. Assess.* **2020**, *25*, 179–197. [CrossRef]
47. Greiff, S.; Niepel, C.; Scherer, R.; Martin, R. Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Comput. Hum. Behav.* **2016**, *61*, 36–46. [CrossRef]
48. Kroehne, U.; Goldhammer, F. How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika* **2018**, *45*, 527–563. [CrossRef]
49. Van Der Linden, W.J. Conceptual issues in response—Time modeling. *J. Educ. Meas.* **2009**, *46*, 247–272. [CrossRef]
50. Li, Z.; Banerjee, J.; Zumbo, B.D. 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*; Springer: Cham, Switzerland, 2017; pp. 159–177. [CrossRef]
51. Araneda, S.; Lee, D.; Lewis, J.; Sireci, S.G.; Moon, J.A.; Lehman, B.; Keehner, M. Exploring Relationships among Test Takers' Behaviors and Performance Using Response Process Data. *Educ. Sci.* **2022**, *12*, 104. [CrossRef]
52. Newton, P.E. What is response process validation evidence and how important is it? An essay reviewing Ercikan and Pellegrino (2017) and Zumbo and Hubley (2017). *Assess. Educ. Princ. Policy Pract.* **2018**, *26*, 245–253. [CrossRef]
53. Yaneva, V.; Clauser, B.E.; Morales, A.; Paniagua, M. Assessing the validity of test scores using response process data from an eye-tracking study: A new approach. *Adv. Health Sci. Educ.* **2022**, *27*, 1401–1422. [CrossRef]
54. Zumbo, B.D.; Chan, E.K. *Validity and Validation in Social, Behavioral, and Health Sciences*; Springer International Publishing: New York, NY, USA, 2014; Volume 54.
55. Just, M.A.; Carpenter, P.A. Eye fixations and cognitive processes. *Cogn. Psychol.* **1976**, *8*, 441–480. [CrossRef]
56. Blascheck, T.; Kurzhals, K.; Raschke, M.; Burch, M.; Weiskopf, D.; Ertl, T. Visualization of eye tracking data: A taxonomy and survey. *Comput. Graph. Forum* **2017**, *36*, 260–284. [CrossRef]
57. Alemdag, E.; Cagiltay, K. A systematic review of eye tracking research on multimedia learning. *Comput. Educ.* **2018**, *125*, 413–428. [CrossRef]
58. Mutlu-Bayraktar, D.; Cosgun, V.; Altan, T. Cognitive load in multimedia learning environments: A systematic review. *Comput. Educ.* **2019**, *141*, 103618. [CrossRef]
59. Coskun, A.; Cagiltay, K. A systematic review of eye-tracking-based research on animated multimedia learning. *J. Comput. Assist. Learn.* **2022**, *38*, 581–598. [CrossRef]
60. Rappa, N.A.; Ledger, S.; Teo, T.; Wai Wong, K.; Power, B.; Hilliard, B. The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: A scoping review. *Interact. Learn. Environ.* **2019**, *30*, 1338–1350. [CrossRef]
61. Just, M.A.; Carpenter, P.A. Using eye fixations to study reading comprehension. In *New Methods in Reading Comprehension Research*; Routledge: Abingdon, UK, 2018; pp. 151–182.
62. Halszka, J.; Holmqvist, K.; Gruber, H. Eye tracking in Educational Science: Theoretical frameworks and research agendas. *J. Eye Mov. Res.* **2017**, *10*, 1–18. [CrossRef] [PubMed]

63. Nisiforou, E.A.; Laghos, A. Do the eyes have it? Using eye tracking to assess students cognitive dimensions. *Educ. Media Int.* **2013**, *50*, 247–265. [CrossRef]
64. Ashraf, H.; Sodergren, M.H.; Merali, N.; Mylonas, G.; Singh, H.; Darzi, A. Eye-tracking technology in medical education: A systematic review. *Med. Teach.* **2018**, *40*, 62–69. [CrossRef]
65. Hu, Y.; Wu, B.; Gu, X. An eye tracking study of high-and low-performing students in solving interactive and analytical problems. *J. Educ. Technol. Soc.* **2017**, *20*, 300–311. [CrossRef]
66. Tsai, M.J.; Hou, H.T.; Lai, M.L.; Liu, W.Y.; Yang, F.Y. Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Comput. Educ.* **2012**, *58*, 375–385. [CrossRef]
67. Brunyé, T.T.; Drew, T.; Weaver, D.L.; Elmore, J.G. A review of eye tracking for understanding and improving diagnostic interpretation. *Cogn. Res. Princ. Implic.* **2019**, *4*, 7. [CrossRef]
68. Lewandowski, D.; Kammerer, Y. Factors influencing viewing behaviour on search engine results pages: A review of eye-tracking research. *Behav. Inf. Technol.* **2021**, *40*, 1485–1515. [CrossRef]
69. Eraslan, S.; Yesilada, Y.; Harper, S. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *J. Eye Mov. Res.* **2016**, *9*, 1–19. [CrossRef]
70. Tien, T.; Pucher, P.H.; Sodergren, M.H.; Sriskandarajah, K.; Yang, G.Z.; Darzi, A. Eye tracking for skills assessment and training: A systematic review. *J. Surg. Res.* **2014**, *191*, 169–178. [CrossRef]
71. Litchfield, D.; Ball, L.J.; Donovan, T.; Manning, D.J.; Crawford, T. Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *J. Exp. Psychol. Appl.* **2010**, *16*, 251. [CrossRef] [PubMed]
72. Seppänen, M.; Gegenfurtner, A. Seeing through a teacher's eyes improves students' imaging interpretation. *Med. Educ.* **2012**, *46*, 1113–1114. [CrossRef]
73. Jarodzka, H.; Balslev, T.; Holmqvist, K.; Nyström, M.; Scheiter, K.; Gerjets, P.; Eika, B. Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instr. Sci.* **2012**, *40*, 813–827. [CrossRef]
74. Kok, E.; Hormann, O.; Rou, J.; van Saase, E.; van der Schaaf, M.; Kester, L.; van Gog, T. Re-viewing performance: Showing eye-tracking data as feedback to improve performance monitoring in a complex visual task. *J. Comput. Assist. Learn.* **2022**, *38*, 1087–1101. [CrossRef]
75. Mason, L.; Scheiter, K.; Tornatora, M.C. Using eye movements to model the sequence of text–picture processing for multimedia comprehension. *J. Comput. Assist. Learn.* **2017**, *33*, 443–460. [CrossRef]
76. Eder, T.F.; Scheiter, K.; Richter, J.; Keutel, C.; Hüttig, F. I see something you do not: Eye movement modelling examples do not improve anomaly detection in interpreting medical images. *J. Comput. Assist. Learn.* **2022**, *38*, 379–391. [CrossRef]
77. 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. How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. *Adv. Health Sci. Educ.* **2017**, *22*, 765–787. [CrossRef]
78. King, A.J.; Bol, N.; Cummins, R.G.; John, K.K. Improving visual behavior research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. *Commun. Methods Meas.* **2019**, *13*, 149–177. [CrossRef]
79. Privitera, C.M.; Stark, L.W. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 970–982. [CrossRef]
80. Goldberg, J.H.; Stimson, M.J.; Lewenstein, M.; Scott, N.; Wichansky, A.M. Eye tracking in web search tasks: Design im-plications. In Proceedings of the 2002 Symposium on Eye Tracking Research & Applications, New Orleans, LA, USA, 25–27 March 2002; pp. 51–58. [CrossRef]
81. Kübler, T.C.; Rothe, C.; Schiefer, U.; Rosenstiel, W.; Kasneci, E. SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behav. Res. Methods* **2017**, *49*, 1048–1064. [CrossRef]
82. Tai, R.H.; Loehr, J.F.; Brigham, F.J. An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *Int. J. Res. Method Educ.* **2006**, *29*, 185–208. [CrossRef]
83. Coutrot, A.; Hsiao, J.H.; Chan, A.B. Scanpath modeling and classification with hidden Markov models. *Behav. Res. Methods* **2018**, *50*, 362–379. [CrossRef] [PubMed]
84. Groner, R.; Walder, F.; Groner, M. Looking at faces: Local and global aspects of scanpaths. In *Advances in Psychology*; North-Holland: Amsterdam, The Netherlands, 1984; Volume 22, pp. 523–533. [CrossRef]
85. Boisvert, J.F.; Bruce, N.D. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing* **2016**, *207*, 653–668. [CrossRef]
86. Kanan, C.; Ray, N.A.; Bseiso, D.N.; Hsiao, J.H.; Cottrell, G.W. Predicting an observer's task using multi-fixation pattern analysis. In Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, FL, USA, 26–28 March 2014; pp. 287–290. [CrossRef]
87. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; van de Weijer, J. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; Oxford University Press: Oxford, UK, 2011.

88. Eivazi, S.; Bednarik, R. Predicting problem-solving behavior and performance levels from visual attention data. In Proceedings of the Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI, Palo Alto, CA, USA, 13 February 2011; University of Eastern Finland: Kuopio, Finland; pp. 9–16.
89. Gegenfurtner, A.; Lehtinen, E.; Säljö, R. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educ. Psychol. Rev.* **2011**, *23*, 523–552. [CrossRef]
90. Siddiq, F.; Gochyyev, P.; Wilson, M. Learning in digital networks—ICT literacy: A novel assessment of students' 21st century skills. *Comput. Educ.* **2017**, *109*, 11–37. [CrossRef]
91. Carretero, S.; Vuorikari, R.; Punie, Y. *DigComp 2.1: The Digital Competence Framework for Citizens*; Publications Office of the European Union: Luxembourg, 2017.
92. Messick, S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* **1995**, *50*, 741. [CrossRef]
93. Bartolomé, J.; Garaizar, P. Design and Validation of a Novel Tool to Assess Citizens' Netiquette and Information and Data Literacy Using Interactive Simulations. *Sustainability* **2022**, *14*, 3392. [CrossRef]
94. Ankerst, M.; Breunig, M.M.; Kriegel, H.P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Rec.* **1999**, *28*, 49–60. [CrossRef]
95. Dolezalova, J.; Popelka, S. Scangraph: A novel scanpath comparison method using visualisation of graph cliques. *J. Eye Mov. Res.* **2016**, *9*. [CrossRef]
96. Holsanova, J.; Rahm, H.; Holmqvist, K. Entry points and reading paths on newspaper spreads: Comparing a semiotic analysis with eye-tracking measurements. *Vis. Commun.* **2006**, *5*, 65–93. [CrossRef]
97. Eraslan, S.; Yaneva, V.; Yesilada, Y.; Harper, S. Web users with autism: Eye tracking evidence for differences. *Behav. Inf. Technol.* **2019**, *38*, 678–700. [CrossRef]
98. Inal, Y. User-friendly locations of error messages in web forms: An eye tracking study. *J. Eye Mov. Res.* **2016**, *9*. [CrossRef]
99. León, J.A.; Moreno, J.D.; Escudero, I.; Kaakinen, J.K. Selective attention to question-relevant text information precedes high-quality summaries: Evidence from eye movements. *J. Eye Mov. Res.* **2019**, *12*, 1–16. [CrossRef]
100. D'Mello, S.; Dieterle, E.; Duckworth, A. Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educ. Psychol.* **2017**, *52*, 104–123. [CrossRef] [PubMed]
101. Van Gog, T.; Scheiter, K. Eye tracking as a tool to study and enhance multimedia learning. *Learn. Instr.* **2010**, *20*, 95–99. [CrossRef]
102. Xie, H.; Zhao, T.; Deng, S.; Peng, J.; Wang, F.; Zhou, Z. Using eye movement modelling examples to guide visual attention and foster cognitive performance: A meta-analysis. *J. Comput. Assist. Learn.* **2021**, *37*, 1194–1206. [CrossRef]
103. Manning, D.; Ethell, S.; Donovan, T.; Crawford, T. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography* **2006**, *12*, 134–142. [CrossRef]
104. Sharafi, Z.; Soh, Z.; Guéhéneuc, Y.G.; Antoniol, G. Women and men—Different but equal: On the impact of identifier style on source code reading. In Proceedings of the 2012 20th IEEE International Conference on Program Comprehension (ICPC), Passau, Germany, 11–13 June 2012; IEEE: New York, NY, USA; pp. 27–36. [CrossRef]
105. Zumbo, B.D.; Maddox, B.; Care, N.M. Process and Product in Computer-Based Assessments. *Eur. J. Psychol. Assess.* **2023**, *39*, 252–262. [CrossRef]
106. Papamitsiou, Z.; Economides, A.A. An Assessment Analytics Framework (AAF) for enhancing students' progress. In *Formative Assessment, Learning Data Analytics and Gamification*; Academic Press: Cambridge, MA, USA, 2016; pp. 117–133. [CrossRef]