Exploring the Integration of Serious Games in High-Stakes Assessments: Evaluating the Role of the Multidimensionality of Engagement on Performance Outcomes

Emine İ. Hotamış S2268736 e.i.hotamis@student.utwente.nl

Dr. Erik C. Roelofs e.c.roelofs@utwente.nl

Dr. Johannes Steinrücke

Faculty of Behavioural, Management and Social Sciences (BMS), University of Twente

Romy Noordhof

Cito

5 march, 2025

Keywords: Serious Games, High-Stakes Assessment, Engagement, Student Performance, Log Data, Game Analytics, Game-Based Assessment

Summary

Traditional high-stakes assessments play a crucial role in determining academic and professional opportunities but face increasing criticism for their lack of ecological validity, failure to assess higher-order cognitive skills, and limited engagement with test-takers. Serious games (SGs) within game-based assessments (GBAs) offer a promising alternative by integrating assessment directly into interactive, immersive gameplay. Engagement is a critical factor in GBAs, influencing assessment validity, motivation, and performance outcomes. However, the multidimensional nature of engagement, consisting out of behavioural, cognitive, emotional, and social dimensions, remains underexplored in high-stakes contexts.

This study investigates the role of engagement in game-based high-stakes assessments by analysing log data and self-reported engagement measures from final-year vocational students. Two serious game environments, Crossroads and Crusade, were used to evaluate engagement-driven performance outcomes. Log-based engagement metrics, including time on task, action sequences, and decision-making patterns, were integrated with subjective questionnaire responses to examine their impact on test performance.

Findings suggest that higher engagement levels, particularly behavioural and cognitive engagement, correlate with improved decision-making accuracy and performance. Regression analyses highlight the significant predictive value of interaction frequency and action diversity, while emotional engagement plays a secondary role in performance outcomes. Additionally, engagement varies between game environments, indicating that game design elements influence assessment effectiveness. These results contribute to the ongoing discourse on GBAs' validity, highlighting the potential for serious games to enhance engagement, reduce test anxiety, and provide more authentic competency assessments in high-stakes settings.

Contents

Introduction

1.1 Problem Statement

Traditional high-stakes assessments play a decisive role in determining academic and professional opportunities, yet they face growing criticism for their lack of ecological validity, limited capacity to assess higher-order cognitive skills, and failure to engage test-takers effectively (Messick, 1994; von der Embse & Witmer, 2014). Conventional assessment formats, such as multiple-choice exams and essay-based tests, often measure static knowledge in controlled conditions rather than evaluating how individuals apply problem-solving, decision-making, and adaptive reasoning in complex, real-world scenarios ((Shute et al., 2016; Roelofs et al., 2019). Moreover, these traditional assessments can induce anxiety, further distorting performance outcomes (Bijl et al., 2024; Steele et al., 2010)

A promising alternative is game-based assessment (GBA), mainly through serious games (SGs), which offer interactive and immersive environments designed to evaluate competencies in more dynamic settings (Checa & Bustillo, 2020). By integrating assessment directly into gameplay, SGs can enhance engagement, reduce test anxiety, and provide deeper insights into cognitive processes(Mislevy et al., 2012; Shute et al., 2016). However, while engagement is often named a significant advantage of GBA, its actual impact on test performance in high-stakes assessments remains insufficiently explored (Gorbanev et al., 2018).

Engagement is a multifaceted construct consisting of behavioural, cognitive, emotional, and social dimensions, each of which may influence performance differently (Fredricks et al., 2004; Henrie et al., 2015). While some research suggests that higher engagement correlates with improved learning outcomes and test performance (Shute et al., 2016b; Wang et al., 2016), it remains unclear whether engagement directly enhances assessment validity or alters test-taker behaviour. Additionally, individual differences in prior gaming experience and digital familiarity may moderate engagement levels, raising concerns about fairness and standardisation in GBA (Mislevy et al., 2016; Roelofs et al., 2019).

Given the increasing rise of digital assessments in education and professional certification, it is important to understand how engagement interacts with test performance in high-stakes GBA. This study aims to fill this research gap by examining the relationship between engagement and performance in serious game-based assessments, contributing to the broader discussion on the effectiveness and validity of GBAs as an alternative to traditional high-stakes testing.

1.2. Theoretical framework

1.2.1. Comparing Hands-Off, Simulation/Game-Based and Hands-On Assessments

Assessment methods differ in their ability to measure competencies, engagement levels, and assessment validity. Each type has its strengths and limitations depending on the context of evaluation (Mislevy, 2013; Wools et al., 2019a). Hands-off assessments rely on indirect, structured tasks that test knowledge through pre-defined response formats. In contrast, simulation/game-based assessments provide interactive and immersive experiences that more closely replicate real-world challenges. On the other hand, hands-on assessments offer the most authentic assessment of practical skills, but face challenges related to standardisation and scoring objectivity (Roelofs et al., 2019).

The choice of assessment method depends on several factors, including the nature of the skills being measured and the extent to which assessments should reflect real-world conditions. Table 1 presents a comparison between these three basic forms of assessment, showing key considerations for each approach (Mislevy et al., 2015; Wools et al., 2019). While hands-off assessments maintain high levels of standardisation, they often lack realism and fail to capture complex cognitive and behavioural processes. Hands-on assessments, on the other hand, provide highly authentic evaluations but may be limited by logistical constraints and subjective scoring challenges. Game-based assessments (GBAs) provide a structured yet more interactive format that falls in between these two methods as they offer standardisation with elements of real-world problem-solving, adaptability, and decision-making.

GBAs are particularly used for evaluating complex skills such as strategic thinking and adaptability by engaging participants in interactive tasks that reflect real-world applications. The advantages of having both authentic and interactive tasks and collecting structured and comparable data make GBAs an appealing option for assessing engagement and performance in educational and professional settings.

Table 1

Comparison of Three Main Types of Assessment

	Hands-Off Assessment	Simulation/Game-Based Assessment	Hands-On Assessment
Nature of the task	It involves indirect tasks like selecting the best response in hypothetical scenarios (Kane et al., 1999; Whetzel & McDaniel, 2009).	It involves performing tasks in a simulated or virtual environment, often goal-oriented (Shute et al., 2016; Mislevy et al., 2016).	Involves real-world tasks performed in authentic environments (Messick, 1994; Straetmans, 2006).
Nature of the task environment	Controlled, simplified, and often standardised (von der Embse & Witmer, 2014; Whetzel & McDaniel, 2009)	Designed to mimic real-world scenarios without real consequences (Checa & Bustillo, 2020; Slater, 1999).	Highly dynamic, with authentic settings and real-world consequences (Straetmans, 2006; Steele et al., 2010).
Nature of evidence obtained	Responses and decision-making processes are often selected from predefined options (Serrano-Laguna et al., 2017; Whetzel & McDaniel, 2009)	Behavioural and process data collected from task execution in the simulated environment (Mislevy et al., 2016; Shute et al., 2016).	Observed task performance directly (Messick, 1994; Roelofs et al., 2019).
Warrants and threats to validity	Limited by a lack of real-world engagement and task-specific results (Kane et al., 1999; Whetzel & McDaniel, 2009).	Validity depends on the realism of the simulation and players' engagement (Mislevy et al., 2016; Wools et al., 2019).	Prone to observer biases but often considered highly authentic and valid (Straetmans, 2006; Messick, 1994).
Accessibility	Highly accessible due to standardisation and less dependent on physical or contextual factors (Roelofs et al., 2019; Shute et al., 2016).	Moderately accessible; requires technological infrastructure but adaptable for different users (Slater, 1999; Mislevy et al., 2016).	It may be limited by physical, contextual, and environmental constraints (Straetmans, 2006).

Table 1

Comparison of Three Main Types of Assessment

Engagement and motivation	Lower engagement; tasks may not feel relevant to participants (Whetzel & McDaniel, 2009; Kane et al., 1999).	Higher engagement due to interactive and immersive elements (Mislevy et al., 2016; Shute et al., 2016).	Engagement depends on real-world stakes and task relevance (Straetmans, 2006; Roelofs et al., 2006).
Reliability of scoring	High, as scoring is objective and standardised (Whetzel & McDaniel, 2009).	It varies based on task complexity and scoring mechanisms and requires validation (Shute et al., 2016; Wools et al., 2019).	Lower reliability due to potential observer variability (Messick, 1994).
Generalisation to test domain	Limited generalisation due to simplified scenarios (Whetzel & McDaniel, 2009; Mislevy et al., 2016).	Moderate generalisation when tasks closely resemble real-world situations (Mislevy et al., 2016).	High generalisation, as tasks replicate the domain directly (Straetmans, 2006; Roelofs et al., 2019).
Extrapolation to target domain	Limited tasks lack fidelity (Whetzel & McDaniel, 2009; Kane et al., 1999).	Improved when simulations are realistic (Mislevy et al., 2016; Shute et al., 2016).	Substantial extrapolation due to direct observation of real-world tasks (Straetmans, 2006; Messick, 1994).

1.2.2. Game-Based Assessment

Game-based assessment is often an integral part of GBL, which integrates assessment directly into gameplay. Game-based learning (GBL) is an instructional approach that uses game mechanics in educational contexts, encouraging active learning and engagement (Gee, 2016; Griffin & Care, 2015). Unlike traditional hands-off assessment methods, which primarily focus on passive information absorption, Game-Based Assessment as part of GBL enables the development of 21st-century skills such as problem-solving, creativity, collaboration, and communication (Shute et al., 2016b). By embedding educational content within these interactive digital environments, the learners can apply theoretical knowledge to practical situations. This enhances cognitive processing and skill retention and, at the same time, gives information about the achieved level of competence (Steinrücke et al., 2020).

While hands-off assessment methods, such as multiple-choice and written essays, primarily measure static cognitive knowledge through predefined questions and answers (Kane et al., 1999), they often fail to assess students' ability to apply knowledge in dynamic, real-world scenarios (Messick, 1994). GBA addresses this limitation by integrating assessment components into gameplay, allowing learners to show their competencies in interactive environments (Ifenthaler et al., 2012; Shute et al., 2016b).

Assessments in serious games have evolved beyond basic simulations, incorporating multiple integrated methods to evaluate cognitive and behavioural skills (Shute et al., 2016b). Game mechanics enable the measurement of real-time decision-making, adaptability, and problem-solving strategies, which traditional assessments often neglect (Roelofs et al., 2019).

A key advantage of GBA is its ability to collect continuous performance data, providing deeper insights into individual learning progress and skill development (Ifenthaler et al., 2012). Unlike conventional tests, SG-based assessments can dynamically adjust to the learner's performance and engagement levels, ensuring personalised challenges that foster meaningful learning experiences (Mislevy et al., 2015).

When compared with live performance assessments (hands-on assessment), GBA offers several distinct advantages, particularly in addressing validity concerns. Live performance assessments, despite their realism (Kane et al., 1999), face challenges such as variability between candidates, rater inconsistency, and limited capacity to collect process data (Wools et al., 2019b). These issues can cause low reliability and limit generalisability, as inconsistent conditions across candidates make it harder to ensure consistent measurement of skills (Roelofs, 2019).

In contrast, GBAs provide a standardised and controlled environment, ensuring that all participants are assessed under the same conditions while automatically capturing process data, such as decision-making sequences and engagement patterns (Mislevy et al., 2015). This reduces the subjective biases of the observers and enhances reliability.

By incorporating these elements, GBAs provide an innovative and possibly more valid alternative to traditional assessment methods, ensuring that assessments are engaging, valid, and reflective of real-world applications. This structured and immersive approach to evaluation makes GBA particularly suited for high-stakes assessment environments where engagement and practical application are critical to performance outcomes.

1.2.3. Validity in GBA

Validity is an important aspect of any assessment method, determining whether an assessment accurately measures what it intends to measure. (Kane et al., 1999; Messick, 1994). In GBA, ensuring validity is complex due to the interactive nature of gameplay, the integration of game mechanics into assessment tasks, and the individual differences in how players engage with digital environments (Mislevy et al., 2015).

1.2.3.1. Design features that affect GBA Validity

As in many other types of assessment, construct validity is a central issue in the context of GBA. Construct validity refers to the extent to which an assessment accurately captures the intended skill or competency rather than measuring extraneous factors such as gaming ability, reaction speed, or familiarity with digital environments (Messick, 1994; Wools et al., 2019b). In the context of GBAs, construct validity requires that game mechanics and assessment tasks align with measured skills as they would take place within the target practice. For example, suppose a GBA is designed to assess problem-solving abilities. In that case, the game must ensure that players engage in critical thinking and logical reasoning rather than relying on trial-and-error gameplay. On the other hand, a well-designed GBA should align task difficulty with cognitive effort, allowing participants to reach flow states where they are fully engaged (Csikszentmihalyi, 2014). Overall, if the game's mechanics do not align with the intended construct, engagement levels may drop, affecting the inferences that can be made about the level of competence of the test-taker (Shute et al., 2016b).

The degree of construct validity highly depends on the design of the GBA. The game tasks that are to be conducted and scored on relevant target skills should represent the domain of relevant target situations. In addition, these tasks need to elicit the target skills. Put differently, the assessment should accurately represent the full range of a given skill or competency in its design (Kane et al., 1999).

Test takers' engagement level may vary in GBA depending on how well the game design reflects real-world tasks (Roelofs & Visser, 2019). A lack of diverse and meaningful tasks can lead to boredom or disengagement, while overly complex tasks may cause cognitive overload (Wools et al., 2019b). The engagement also tends to be higher in realistic, meaningful game scenarios that align with test-takers experiences (Mislevy et al., 2015). If the game design is weak in terms of its representation of tasks, the participants may disengage, perceiving the assessment as irrelevant to their skills and not likely to show their typical performance (Shute et al., 2016b).

Accessibility is a critical attribute of assessment that impacts construct validity, especially when participating in the assessment requires skills that come with innovative technology that cannot be taken for granted. The question is whether performance failures in assessments are due to a lack of target skills or due to a lack of access skills, needed to deal with the chosen form of assessment presentation. Related to this, a key issue in GBA is whether prior gaming experience influences engagement levels, leading to discrepancies in performance (Straetmans, 2006). Test-takers who frequently play video games may engage more easily, as they feel more comfortable navigating the game mechanics. In contrast, those unfamiliar with gaming interfaces may experience higher cognitive load, reduced engagement, and frustration (Mislevy et al., 2015; Roelofs, 2019). In the design and use of GBAs, the lack of accessibility needs to be mitigated to enable valid inferences about test-takers' levels of targeted skills as measured through game participation (Mislevy et al., 2015; Roelofs & Visser (2019)). If this is not done, the GBA may yield unfair decisions about those test takers that have less access to the assessment, due to their lack of experience with games.

1.2.4. The Role of Engagement in GBA

Returning to the issue of construct coverage, we assume that engagement impacts performance in game-based assessments and therefore is a possible threat to validity when part of the test takers is disengaged while others are highly engaged. So, in this study engagement in GBAs is measured and related to the game performance, to judge the possible impact on the validity of GBA.

To judge the possible impact of engagement on GBA performance, the construct of engagement itself needs further elaboration.

Engagement is a multidimensional construct critical in learning, performance and assessment (Fredricks et al., 2004). In GBA, engagement is particularly important because interactive digital environments rely on continuous participation to generate meaningful performance data (Shute et al., 2016b). Unlike traditional tests where test-takers passively respond to questions, GBA requires active decision-making, problem-solving, and interaction with environments, making engagement a key factor in assessment validity and performance outcomes (Wools et al., 2019b).

Engagement is usually defined by four interrelated dimensions: behavioural engagement, cognitive engagement, emotional engagement, and social engagement.

Behavioural engagement refers to observable student behaviour, such as persistence, task completion, and active involvement in the game, to measure the extent to which players interact with the world (Appleton et al., 2008) For example, indicators like the revisit frequency to tasks illustrate long-term focus and provide deeper insights into the level of engagement (Serrano-Laguna et al., 2017). Behavioural engagement engages the player and helps in his overall learning experience and performance.

Cognitive engagement reflects the intellectual commitment necessary for problem-solving, strategic planning, and deep learning. Players must utilize critical thinking and adaptive approaches to be effective in activities that often replicate real-world challenges (Henrie et al., 2015). Serious games applied for logistics or planning, for example, simulation games, challenge participants actively to explore and evaluate multiple solutions.

Emotional engagement encompasses affective responses, such as satisfaction, frustration, and enjoyment, which influence motivation and sustained effort (Fredricks et al., 2004). Positive emotions can enhance engagement using intrinsic motivation, while negative emotions can disempower progress and performance. Research by Csikszentmihalyi (2014)# has shown that emotional engagement is closely associated with flow states, where students experience heightened concentration and intrinsic enjoyment.

The fourth dimension, social interaction, exists in collaborative serious games. Coordination, collaboration, and shared decision-making are required to finish game tasks and to improve the overall engagement experience (Wang et al., 2016).

By promoting high rates of engagement, game-based tests encourage players to display behaviours and strategies typical of professional or academic settings, making assessment outcomes more reflective of true skills (Hookham & Nesbitt, 2019). This is particularly important in high stakes testing situations, where engaged participants will tend to better show their abilities, reducing test anxiety and maximising the validity of their performance evaluations (von der Embse & Witmer, 2014).

Engagement in games is found to contribute to learning by creating deeper immersion in the assessment environment. When players are highly engaged, they stay focused and actively process information from the material. They are also more likely to apply knowledge effectively to overcome problems (Baker & Yacef, 2009). This continued involvement enhances cognitive processing and enables individuals to adapt better to dynamic assessment tasks, leading to improved performance outcomes (Baker & Yacef, 2009; Shute et al., 2016b). Furthermore, engagement is found to influence test-taking behaviour, as participants who are more immersed in the assessment tend to make more calculated decisions, have a higher persistence in solving complex problems, and minimise errors, resulting in a more reliable evaluation of their skills and competencies (Shute et al., 2016b).

However, while engagement is desirable, it is not consistent across all test-takers in GBA, as it can be influenced by factors such as game design, cognitive demands, and prior experience with digital environments (Mosiane & Brown, 2020). While some players may find interactive elements such as adaptive feedback, challenge-based tasks, and immersive simulations more engaging, others may experience cognitive overload or disengagement due to unfamiliarity with game mechanics (Mislevy et al., 2015). As long as these variations in engagement levels are unrelated to the competencies being measured, they may act as a disturbing factor, potentially compromising the validity of inferences drawn about test-takers true competencies.

Overall, engagement is a driver of performance and an important prerequisite for assessment validity. Fostering engagement in GBA design by encouraging continuous attention, cognitive effort, and authentic test-taker responses, enhances the quality of performance data collected in these assessments.

1.3. Log Data

In the context of GBAs, log data is an objective method for capturing engagement and behavioural patterns. Log data refers all the steps users have taken as they interact with a digital environment, such as time spent on tasks, action sequences, task completion rates, and decision-making processes (Baker & Yacef, 2009; Ifenthaler et al., 2012).

A key advantage of log data is its ability to quantify cognitive, behavioural, and emotional engagement (Fredricks et al., 2004). Unlike self-reported engagement measures, which are subject to biases such as social desirability or recall errors, log data provides real-time insights into participant interactions (Shute et al., 2016).

Behavioural engagement is reflected in task persistence, revisiting frequency, and completion rates. Cognitive engagement is captured through decision-making strategies, problem-solving attempts, and logical sequencing (Henrie et al., 2015). Emotional engagement can be inferred from interaction patterns, such as hesitation before responding or repeated retries, which may indicate frustration or determination (Reschly & Christenson, 2012).

In high-stakes GBAs, log data plays a critical role in distinguishing engagement from ability. Performance variability can often be attributed to differences in engagement rather than actual skill levels (Shute et al., 2016a). Additionally, log data can capture flow states, cognitive load, and sustained attention (Arnab et al., 2015). For instance, frequent retries on a challenging task may indicate perseverance and cognitive effort, while rapid task completion with minimal interaction and high errors may suggest disengagement (Shute et al., 2016a).

Despite these advantages, interpreting log data presents challenges. While log data provides objective behavioural measures, it does not fully reveal user motivations or cognitive processes (Henrie et al., 2015). For example, extended time on a task could indicate deep cognitive engagement for one participant but confusion or difficulty for another. Furthermore, distinguishing between engagement-driven and ability-driven performance remains complex (Fredricks et al., 2004).

To enhance interpretability, this study combines log data with self-reported engagement measures, providing a comprehensive analysis of how engagement impacts performance in serious game-based assessments.

1.4. Research Question

In this thesis, the impact of engagement in high-stakes game-based assessments on game performance is explored. For that purpose, two GBA environments developed by Cito in 2022, were focused on in this study. Engagement is considered a prerequisite for performing in game-based assessment, in which performance in turn is a basis for yielding valid assessment scores. Therefore, it is important to study how levels of engagement impact the game performance, by tracking how individuals interact with game-based environments.

Thus, posing the research question: *How does engagement in serious games for high-stakes assessment affect the performance of test takers?*

The following sub-questions are posed:

1) What is the relationship between engagement, as measured through log-based metrics and self-reported measures, and performance in serious games for high-stakes assessment?

2) How do variations in game environments affect engagement levels, and how does this interaction influence test-taker performance in high-stakes assessments?
3) How is prior experience with gaming and digital devices related to engagement levels and performance in high-stakes game-based environments?

By examining these relationships, this study provides empirical insights into the role of engagement in GBA, contributing to the broader understanding of engagement-driven assessment performance in high-stakes contexts.

2. Methods

2.1. Research design

This study uses a mixed-methods research design to investigate the role of engagement in serious games for high-stakes assessments. The data were collected using two GBA environments developed by Cito, specifically designed to evaluate problem-solving and decision-making skills in realistic, high-stakes scenarios. Unlike hands-off assessments, which primarily evaluate theoretical legal knowledge, these serious games were developed to assess Dutch municipal enforcement officers (BOA) students. The aim was to explore how game-based elements can enhance test engagement and improve assessment validity.

The role of a municipal enforcement officer (BOA) in the Netherlands includes maintaining public order, enforcing municipal regulations, and addressing minor legal violations. They do not have the same authority as the police, but they are authorised to enforce certain laws. BOAs operate in public spaces, transportation hubs, and local communities, ensuring compliance with laws related to environmental regulations, public disturbances, parking enforcement, and administrative fines. Their work requires critical thinking, situational judgment, and decision-making under uncertainty.

This study uses process data, captured through log-based metrics, to quantify engagement behaviours and problem-solving strategies from the environments. Additionally, questionnaire responses were collected to measure participants' perceived engagement levels, providing subjective insights that complement the behavioural data.

2.2. Participants

The initial sample included 116 final-year students who were enrolled in the BOA track. The students were recruited from thirteen vocational schools and nine municipalities. All participants provided written informed consent to participate in the study. Participants who did not use the assigned codes could not be linked and were subsequently excluded from the final analysis (n = 16).

2.3. Data collection

Data collection took place related to and within two game environments. The log data were collected in the context of two serious games, designed to assess students' problem-solving skills in an environment that reflects their usual tasks. The games are described below.

2.3.1. Serious Game 1: Crossroads

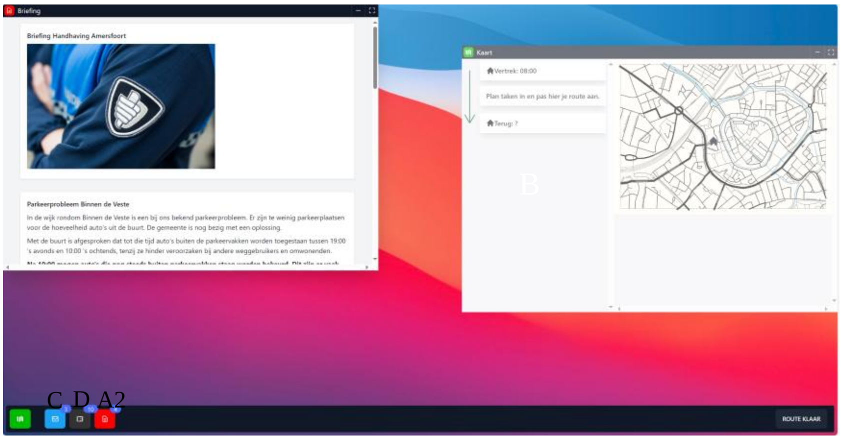
Crossroads aligns the most with their actual tasks as a BOA. The player assumes the role of a BOA and is tasked with developing a surveillance plan for a city. This game provides a realistic environment where players interact with scenarios similar to those they might encounter in real life. Although elements, such as specific knowledge of the city, are irrelevant, the main interactions closely resemble real-world tasks. The player first encounters three slides providing practical information about the assignment. This information is also available in the instruction file (Appendix A). These slides include a video that explains the available actions, how each button functions, the sequence of actions, and where the player can click. The introduction is not skippable, so all information must be viewed.

As shown in Figure 1, the player sees two main fields: the briefing and the map. In the bottom left corner, three icons give access to three actions: emails, tickets, and the briefing. The briefing outlines the tasks to be completed. There are three emails, and participants can either forward them to another authority (e.g., the police), transfer the task for reasons such as lack of time or low priority, or plan the task, explaining their choice (Figure 2). A similar system applies to the tickets, where the player receives ten tickets, each detailing complaints or problems.

The briefing also contains important information for players, such as traffic restrictions (e.g., no biking allowed on a street between 8:00 and 18:00). This information must be considered when planning routes. Once the player has planned their route, they receive two additional notifications: alerts that need to be scheduled after the route is finalised.

Figure 1

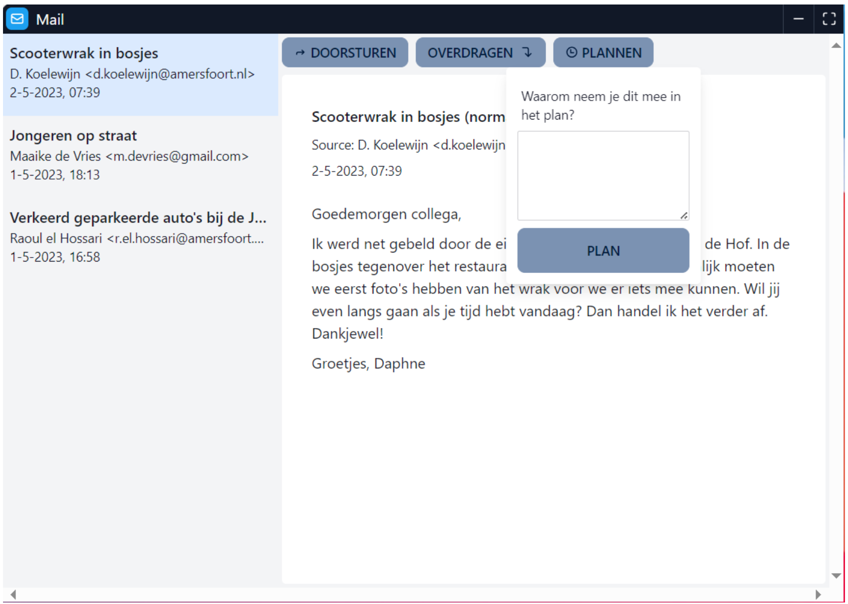
The main page of Crossroads



Note. On the left side the tab “Briefing” is visible (A1) and on the right tab the Map is visible (B). At the bottom, three icons are visible: Emails (C), Tickets (D), and Briefing (A2). From: Crossroads, 2024, Crossroads.web.app

Figure 2

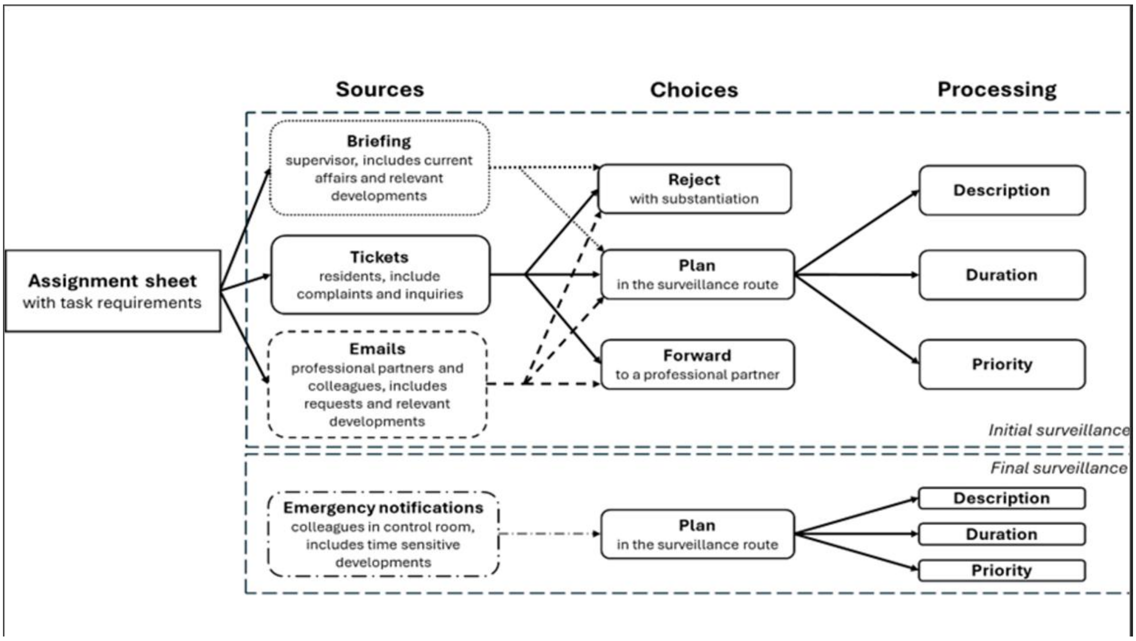
The mailbox in Crossroad



Note. The mailbox that can be accessed in the game. On the left side, the previews of the three emails are shown and, on the right, the email details are shown. On top there are three buttons “doorsturen” (forward), “overdragen” (transfer), “plannen” (plan). From: Crossroads, 2024, Crossroads.web.app

Figure 3

Flowchart Crossroads



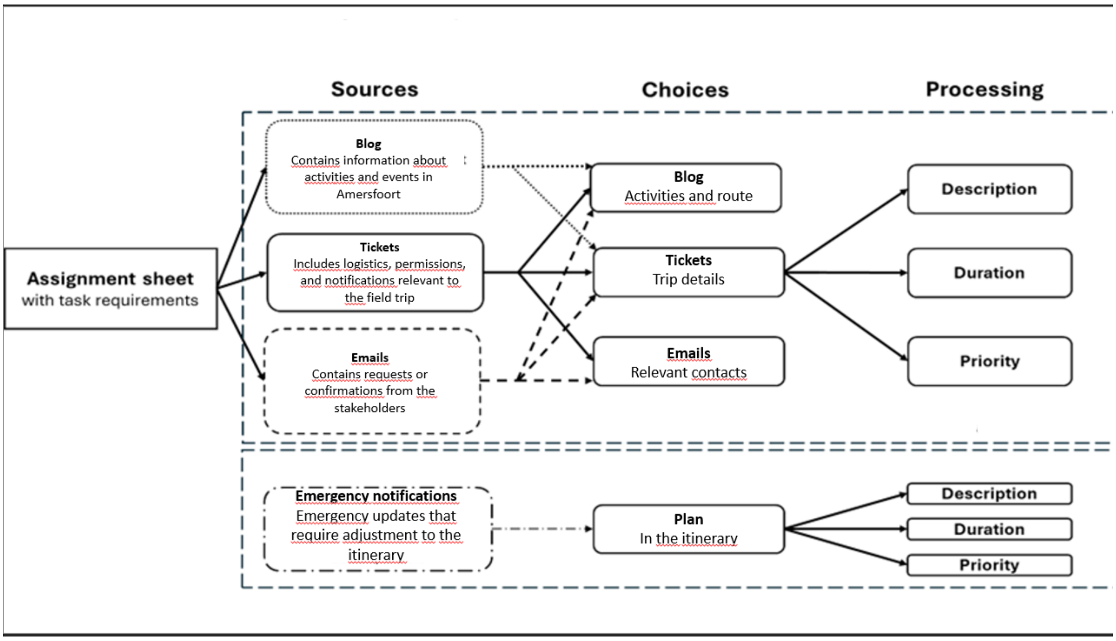
Note. Environment Crossroads, 2024, Crossroads.web.app

2.3.2 Serious Game 2: Crusade

In the second environment, Crusade, the players plan a field trip, focusing on logistics and event organisation within a city (Figure 4). The gameplay elements are parallel to Crossroads, with introductory slides and instructions provided in the manual (Appendix C). Instead of a briefing, players access activity information via a blog, selecting appropriate activities for the group. Tasks include managing schedules and ensuring all required activities are addressed.

Figure 4

Flowchart Crusade



Note. Environment Crusade, 2024, Crossroads.web.app

2.3.3 Log-data

Log data was collected during participant interactions with the Crossroads and Crusade serious game environments. These data include real-time behavioural records such as time spent on tasks, action sequences, interaction frequency, and decision-making strategies. The log data was extracted from backend systems and stored in structured datasets for analysis.

To quantify engagement, the following log-based indicators were used:

1. Time Spent on Task: Measures the total duration participants engaged with a task, indicating persistence and cognitive investment (Henrie et al., 2015). The mean and median task durations were analysed to differentiate between engaged and disengaged participants (Csikszentmihalyi, 2014).
2. Interaction Frequency: Captures the total number of actions taken within the game environment (e.g., clicks, text entries, and object interactions), reflecting active engagement levels (Fredricks et al., 2004). A higher interaction frequency suggests increased engagement and sustained attention (Shute et al., 2016b).

3. Action Types: Examines the variety of actions performed within the game, providing insights into engagement diversity and problem-solving approaches. Greater action diversity suggests higher cognitive engagement.
4. Decision-Making Sequences: Tracks the order of actions performed, revealing strategic thinking and problem-solving behaviours (Serrano-Laguna et al., 2017). This metric helps distinguish between trial-and-error approaches and deliberate, structured decision-making.
5. Average Time per Action: Measures the efficiency of decision-making by dividing total time spent by number of actions performed. A longer average time per action may indicate hesitation or cognitive overload, whereas a shorter time may suggest impulsivity or disengagement.

Table 2 presents the engagement rubric, detailing the metrics used to assess engagement levels.

Table 2

Engagement Rubric

	Metrics	Indicators
Behavioural Engagement	Task completion rates, action diversity, frequency of task revisits	High action diversity, consistent task completion, frequent revisits for problem refinement
Cognitive Engagement	Frequency of strategy adjustments, problem-solving attempts, logical sequence in tasks	Active exploration of strategies, logical decision-making, and task revisits for optimisation.
Emotional Engagement	Time spent on tasks, self-reported enjoyment/frustration, motivation indicators	Balanced time distribution, high motivation, low frustration interruptions

2.3.4. Questionnaire

To complement the log data, **s**elf-report measures, such as questionnaires and surveys, provide valuable insights into test-takers subjective engagement experiences as they capture cognitive and emotional feelings that log data alone cannot measure (Fredricks et al., 2004; Wools et al., 2019b). These measures assess factors such as intrinsic motivation, task difficulty, frustration, and interest levels, contributing to a better understanding of how engagement influences performance (Henrie et al., 2015; Shute et al., 2016b). Additionally, self-reported engagement allows for a direct comparison between participants' perceived engagement and their actual behavioural patterns.

However, self-reports have limitations, including response bias and retrospective inaccuracies. Participants may struggle to accurately recall and assess their engagement, particularly when self-reporting takes place after the assessment rather than during it (Wools et al., 2019). Additionally, social desirability bias may lead test-takers to provide responses that align with expected norms rather than their true experience (Fredricks et al., 2004). By combining these two self-reported measures the study aimed to assess whether prior digital experience influenced engagement levels, and, if so, to what extent it impacted test-takers interactions with the game environments.

The first questionnaire was the environment questionnaire, adapted from the System Usability Scale (SUS) (Bangor et al., 2008; de Klerk et al., 2018). It was designed to measure participants' perceptions of each serious game environment, focusing on engagement, usability, and immersion. Each version of the questionnaire consisted of eight items, with questions tailored to include the specific environment's name. Items included statements such as "The environment felt like a game" and "I think that the environment hurt my performance," prompting participants to reflect on the environment, game-like qualities, and any perceived impact on their assessment performance. These responses provide subjective insights into the participants' levels of presence and engagement within each environment. This questionnaire can be found in Appendix B.

The second questionnaire was the computer use questionnaire, containing eight items, designed to capture participants' familiarity with digital devices and their usage habits, which may influence their engagement levels and comfort within digital game-based environments (Kuhlemeier & Hemker, 2007). It includes items like "How long do you use your digital devices on an average day during the weekend (outside of school/work)." This questionnaire provided additional context on participants' general digital proficiency, which was useful in examining potential moderating effects on engagement and presence in the game environments. This questionnaire can be found in Appendix D.

2.3.5. Integrating Log Data and Self-Reports in Engagement Measurement

While self-report data alone is subjective, integrating it with log data analysis enhances engagement measurement by combining objective behavioural indicators with subjective perceptions (Baker & Yacef, 2009; Ifenthaler et al., 2012). This mixed-method approach allows for more nuanced interpretations of engagement patterns.

For instance, a participant who spends a long time on a task (log data) but reports low motivation (self-report) may be experiencing cognitive overload rather than deep engagement (Henrie et al., 2015). Conversely, a participant who completes tasks quickly but reports high engagement may demonstrate efficient problem-solving rather than disengagement (Shute et al., 2016a).

By integrating both methods, this study bridges the gap between objective engagement behaviours and subjective participant experiences, offering a richer and more accurate assessment of how different factors shape engagement in game-based assessments (Wools et al., 2019b).

2.4. Data collection procedure

The data used in this research were collected from a previous experiment that was run from May 23 to July 20, 2023. During the experiment, participants engaged with two GBA environments developed by Cito: "Crossroads," and "Crusade." All participants played both Crossroads and Crusade. The sequence of game environments was randomised for each participant, who could experience the following orders: A-B or B-A.

Ethical approval was initially obtained for the data collection from the Ethics Committee of the Behavioural, Management and Social Sciences Faculty of the University of Twente under request number 230152. Subsequently, approval for using this data for the current study was granted under case number 240702. Participants were fully informed about the study objectives and asked for consent to use their data. Participants who did not give consent to further use of their data were excluded. In addition, all personal identifying information was anonymised to ensure privacy.

Participants received detailed instructions before starting each game-based task (see Appendices A and C). Following each game session, they completed a general questionnaire to capture their opinion about the environment. A secondary questionnaire gathered data on participants' gaming and computer experience, providing additional context for interpreting engagement levels.

The study's process data, gathered from participants' interactions within the SG environments, recorded engagement behaviours such as time on task, revisits, and decision-making patterns.

2.5. Data Analysis

The participants' log data were analysed to quantify engagement within the game environments. These data captured detailed records of user interactions, including time spent on tasks, entries, removed inputs, and revisits within the game environment. To assess the distribution of engagement scores, Shapiro-Wilk tests were conducted. Given that all engagement variables violated the assumption of normality ($p < 0.001$ across all variables), non-parametric Wilcoxon Rank-Sum Tests were applied to compare the two environments. Additionally, Chi-Square Tests were used to assess the relationship between missing responses and environment type. Correlation analysis examined the relationship between the participant's computer usage and their engagement scores.

2.5.1. Questionnaire Data on Engagement

The questionnaire responses complemented log data by identifying subjective engagement components, namely perceived presence, cognitive involvement, and emotional responses. Some of the important questionnaire items pointed towards insights into the perceived presence, emotional engagement, and any prior digital experience of the participants. These data were instrumental in examining how participants' self-reported experiences aligned with objective log data metrics (Appleton et al., 2008).

2.5.2. Integrated Analysis of Log Data and Questionnaire Responses

To provide a comprehensive analysis of engagement in serious game-based assessments, both log-based engagement metrics and self-reported engagement measures were examined. This mixed-methods approach enabled a more thorough understanding of the relationship between engagement and performance outcomes.

2.5.3. Correlation Analysis

To explore the associations between engagement dimensions and performance metrics, a Pearson correlation analysis was conducted. This analysis examined how log-based engagement indicators, including time spent on task, interaction frequency, action types, and average time per action, correlated with self-reported engagement scores (e.g., perceived presence, cognitive engagement) and performance measures (decision-making accuracy and timing performance). This approach facilitated an assessment of whether self-reported engagement perceptions aligned with behavioural engagement patterns and how these factors contributed to task performance.

2.5.4. Regression Analysis

To evaluate the predictive value of engagement metrics on performance, multiple regression analyses were conducted for both Crossroads and Crusade. The dependent variable (DV) was Performance score (decision-making accuracy) and the independent variables (IVs) were: time spent on task (total duration of engagement), interaction frequency (number of actions performed), action types (variety of distinct actions performed), and average time per action (time taken per interaction, measuring response efficiency).

Separate regression analyses were conducted for each game environment to examine potential differences in engagement effects.

2.5.5. Interaction Effects

To assess whether prior digital experience influenced the relationship between engagement and performance, an interaction term was included to the regression mode. Specifically, interaction effects were tested between prior digital experience and engagement metrics to determine whether participants with high digital experience showed different engagement patterns and performance outcomes compared to those with less experience.

2.5.6. Data Preprocessing and Assumptions

Prior to conducting these analyses, the raw log data were pre-processed to remove redundant data such as repeated clicks, and when the participant did not do anything for a longer period. The descriptive statistics were calculated to examine the distribution of engagement scores and assess normality assumptions. As the engagement variables violated normality assumptions (Shapiro-Wilk test, $p < 0.001$), non-parametric Wilcoxon Rank-Sum Tests were conducted for comparisons between game environments. Additionally, a Chi-Square Test was conducted to determine whether the presence of missing responses was related to the type of game environment.

3. Results

In this section, the findings about engagement and relationships are presented. The players' decision-making, engagement, and task performance were measured in these environments. The results include performance metrics and comparative analyses.

3.1. Descriptive statistics

The descriptive statistics for behavioural, cognitive, and emotional engagement, as well as decision-making accuracy and timing performance, are listed in Table 3. The findings indicated that cognitive engagement had the highest mean score, while behavioural engagement exhibited the greatest variability among participants scores. The cognitive engagement had a mean score of 3.78 (SD = 3.93), with a range of 0.00 to 14.00. The unexpectedly high upper range indicates that some participants engaged in frequent decision-making adjustments and strategic planning. Emotional engagement had a mean score of 5.68 (SD = 0.80), ranging from 3.50 to 7.00. The decision-making performance had a mean score of 56.25 (SD = 12.34), with a range of 30.00 to 78.00. Some participants demonstrated high accuracy in task completion, while others struggled to select optimal responses and timing performance, measured in seconds per task, had a mean of 77.78 (SD = 10.22), with completion times ranging from 50.00 to 95.00 seconds. Results indicate variation in task completion strategies, with some participants completing tasks efficiently and others taking longer to process information.

Table 3

Descriptive statistics for behavioural, cognitive and emotional engagement and decision-making and timing performance.

Variable	M	SD	Min	Max
Behavioural engagement	1.50	1.97	0.38	7.47
Cognitive engagement	3.78	3.93	0.00	14.00
Emotional Engagement	5.68	0.80	3.50	7.00
Decision-Making Performance	56.25	12.34	30.00	78.00
Timing performance	77.78	10.22	50.00	95.00

3.2. Relationship between Engagement and Performance

A Pearson correlation analysis was conducted to examine the relationships between engagement dimensions (behavioural, cognitive, and emotional) and performance metrics (decision-making performance and timing efficiency). The correlation coefficients are presented in Table 4.

Table 4

Correlation Analysis of Engagement and Performance Metrics

Variable	Total Actions (r)	Performance Scores (r)
Total Time Spent	0.54	0.42
Action Types	0.51	0.39
Average Time per Action	-0.30	-0.45

The findings indicated that total time spent was moderately positively correlated with total actions completed ($r = 0.54, p < .01$) and decision-making accuracy ($r = 0.42, p < .01$). The number of action types used was also positively correlated with performance ($r = 0.39, p < .05$). A significant negative correlation was found between average time per action and both total actions ($r = -0.30, p < .05$) and decision-making accuracy ($r = -0.45, p < .01$).

3.3. Regression Analysis

Multiple regression analyses were performed for both Crossroads and Crusade to assess the predictive value of engagement on performance. The dependent variable in both models was total actions, which served as a proxy for overall task completion and engagement in the GBA. The independent variables included total time spent, average time per action, and action types to assess their impact on performance. the results were presented in table 5 and table 6.

For Crossroads, the regression model explained 85.6% of the variance ($R^2 = 0.856$) in total actions (Table 5). Total time spent was a significant positive predictor, while average time per action had a significant negative effect. Action types did not significantly predict performance.

For Crusade, the regression model explained 48.9% of the variance ($R^2 = 0.489$) in total actions (table 6). Total time spent significantly predicted performance. Action types were also a significant positive predictor. Average time per action had a significant negative effect.

Table 5

Regression Analysis Crossroads

Variable	B	SE	t	p	95% Cl	
					LL	UL
Constant	42.340	11.210	3.68	0.001**	18.730	63.72
Total Time Spent	0.0396	0.003	13.74	<.001***	0.034	0.045
Average Time per Action	-1.726	0.158	-10.96	<.001***	-2.042	-1.41
Action Types	1.766	2.253	0.75	0.456	-2.959	6.49

Table 6

Prediction model of game performance on the game Crusade, using engagement indicators as predictors

Variable	B	SE	t	p	95% Cl	
					LL	UL
Constant	-15.72	2.19	-7.18	<.001***	-20.02	-11.43
Total Time Spent	0.02	0.00	54.66	<.001***	0.02	0.02
Average Time per Action	-0.29	0.01	-31.63	<.001***	-0.30	-0.27
Action Types	14.94	0.49	30.29	<.001***	13.98	15.91

3.4. Engagement Measures Across Environments

Table 7

Wilcoxon Rank-Sum Test

Variable	W	p-value
Ease of Use	1556.5	.659
Need for Assistance	1797.0	.275
Enjoyment of the Experience	1560.5	.600
Perceived Responsiveness of Game	1390.5	.666
Perceived Alignment with Skills	1854.5	.030
Effectiveness of the Assessment	1928.5	.007*

Note. n = 116. Wilcoxon Rank Sum Test results are reported, $p < .05$*, p < .01**

A Wilcoxon Rank-Sum Test was conducted to compare engagement levels between Crossroads and Crusade. Results indicated no significant differences in ease of use need for help, fun experience, or game feel.

However, there were significant differences in perceived skill matching and skill assessment effectiveness, with participants rating Crossroads higher than Crusade on both measures.

A Chi-Square Test revealed no significant association between missing responses and game environment type ($\chi^2(1) =$, $p > 0.05$).

3.5. Correlation Between Computer Usage and Engagement

Table 8

Correlation Between Computer Usage and Engagement

Variable	Correlation with Total Usage
Ease of Use	0.2242
Need for Assistance	-0.2235
Enjoyment of the Experience	0.0963
Perceived Responsiveness of Game	0.1936
Perceived Alignment with Skills	-0.0004
Effectiveness of the Assessment	0.0068

A correlation analysis examined the relationship between participants' overall computer usage and engagement levels. Results showed a weak positive correlation between total computer usage and ease of use and game feel.

A weak negative correlation was found between computer usage and need for help. Additionally, perceived skill alignment and skill assessment effectiveness were strongly correlated.

4. Discussion

In GBA, engagement plays a key role. It enhances decision-making and task completion by promoting behavioural interaction, cognitive effort, and emotional investment. Given its multidimensional nature, understanding how engagement operates within serious games is important to refine its use in high-stakes testing, ensuring proper competency measurement. By combining log data with self-reported engagement measures, we provide insights into how cognitive, behavioural, and emotional engagement contribute to test-taking behaviours and performance outcomes.

In conclusion. This study highlights the critical role of engagement in serious game-based assessments, influencing both performance and hence validity. Higher levels of behavioural, cognitive, and emotional engagement consistently led to better decision-making and task completion. However, emotional engagement showed a more variable relationship with performance, suggesting that motivation and enjoyment are not always predictive of improved outcomes.

Additionally, prior digital experience raises concerns about fairness. Participants with more digital familiarity tended to engage more actively and perform better, indicating that accessibility and digital proficiency must be considered in the design of GBA. This would ensure that engagement differences do not impact the performance outcomes, which is essential for maintaining validity in high-stakes assessments.

These findings align with previous literature, which identifies engagement as an important motivator of learning and assessment validity (Fredricks et al., 2004; Shute et al., 2016b). The positive correlation between engagement and performance reinforces the role of serious gaming in creating an interactive and cognitively demanding assessment environment (Mislevy et al., 2015). However, the observed variability in engagement across different game environments suggests that engagement is context-dependent, which would suggest that game design choices influence the performance (Serrano-Laguna et al., 2017).

Furthermore, the impact of prior digital experience on engagement highlights an important challenge for fairness and accessibility in game-based assessments. As

4.1. Addressing the Research Questions

The following question was asked:

How does engagement in serious games for high-stakes assessment affect the performance of test takers?

To systematically address this research question, three more specific questions were posed:

1. *What is the relationship between engagement, as measured through log-based metrics and self-reported measures, and performance in serious games for high-stakes assessment?*

The findings indicate a positive correlation between engagement and performance, with higher engagement levels consistently associated with better decision-making and task completion rates. As captured through log data metrics (e.g., task revisits, interaction frequency, and action sequences), behavioural engagement was particularly significant in predicting performance, confirming that participants who engaged more actively with the game environments achieved better outcomes. This aligns with existing research suggesting that engagement is a key factor in educational effectiveness, particularly in digital learning environments (Fredricks et al., 2004; Henrie et al., 2015).

The correlation analysis (Table 4) supports this relationship, showing that total time spent within the game was moderately positively correlated with both total actions completed (r = 0.54, p < .01) and performance scores (r = 0.42, p < .01). These findings align with prior research indicating that sustained interaction enhances task familiarity, leading to improved decision-making accuracy.

Cognitive engagement, measured through problem-solving behaviours and strategic planning, was also a strong predictor of performance. Participants who frequently revised their decisions, adjusted strategies, and explored multiple solutions demonstrated higher test results. This is further supported by the positive correlation between the number of action types used and performance (r = 0.39, p < .05), suggesting that engaging with a greater variety of actions allowed participants to navigate the game more effectively, leading to better outcomes. However, the regression analyses (Table 5 & Table 6) suggest that this effect varied depending on the game environment.

In contrast, the negative correlation between average time per action and performance (r = -0.45, p < .01) suggests that taking longer per action did not improve accuracy. Instead, this may indicate cognitive overload or hesitation, where participants struggled to process information efficiently, leading to reduced performance outcomes. This finding highlights the importance of balancing task complexity and response efficiency in game-based assessments, ensuring that longer response times do not interfere with effective decision-making.

all in all, these findings suggest that serious games facilitate a more engaging and cognitively demanding assessment experience, leading to stronger performance outcomes and potentially greater test validity in high-stakes settings.

2. *How do variations in game environments affect engagement levels, and how does this interaction influence test-taker performance in high-stakes assessments?*

The relationship between engagement and performance varied across the Crossroads and Crusade environments, highlighting how game mechanics influence the way engagement translates into performance outcomes.

The regression analysis for Crossroads (Table 5) indicates that total time spent was a significant positive predictor of performance, while average time per action had a significant negative effect. However, action types did not significantly predict performance, suggesting that engagement diversity was less relevant in this environment, where task sequences were more predefined. This implies that Crossroads may have required a more linear approach, where efficiency in decision-making was prioritised over exploration or diverse actions.

In contrast, the regression analysis for Crusade (Table 6) revealed that both total time spent, and action types were significant positive predictors of performance. This suggests that participants who engaged in a broader range of actions performed better, likely due to the open-ended, problem-solving nature of the game. In this context, exploration and flexibility in engagement strategies contributed more directly to successful task completion.

The Wilcoxon Rank-Sum Test (Table 5) further highlights differences in engagement perceptions across game environments. While there were no significant differences in ease of use, need for help, fun experience, or game feel, participants rated Crossroads significantly higher for perceived skill matching ($p = 0.0295$) and skill assessment effectiveness ($p = 0.0066$). This suggests that participants felt Crossroads better aligned with their abilities, possibly due to its structured nature, which may have enhanced their perception of assessment fairness and relevance.

These findings indicate that game-based assessments must be carefully designed to ensure that engagement does not disproportionately affect assessment validity. While open-ended environments like Crusade may encourage cognitive engagement, structured environments like Crossroads may provide a clearer alignment between task requirements and participant skill levels. Understanding these design-related differences is critical for ensuring that game-based assessments accurately measure intended competencies rather than engagement-driven behaviours..

3. How is prior experience with gaming and digital devices related to engagement levels and performance in high-stakes game-based environments?

A tertiary finding of the study was the impact of prior digital experience on engagement levels and performance outcomes. Participants with more extensive gaming and digital device usage exhibited higher engagement and better performance, confirming previous research suggesting that digital proficiency enhances adaptability to game-based environments (Mislevy et al., 2016; Roelofs et al., 2019).

The correlation analysis (Table 6) supports this, showing that total computer usage was positively correlated with ease of use ($r = 0.22$) and game feel ($r = 0.19$), while negatively correlated with need for help ($r = -0.22$). This suggests that participants with higher digital proficiency found the game environments easier to navigate and required less support, which likely contributed to their higher engagement and performance. Additionally, perceived skill alignment and skill assessment effectiveness were strongly correlated ($r = 0.56$), indicating that participants who felt the game reflected their skills also perceived it as an effective assessment tool.

However, this finding raises concerns regarding the standardisation and fairness of game-based assessments (GBAs), particularly in relation to prior digital experience. The results suggest that participants with more gaming familiarity demonstrated higher engagement and better performance, implying that those with less digital experience may be at a disadvantage. This aligns with concerns in the literature about GBAs introducing unintended biases related to digital literacy (Mislevy et al., 2016; Roelofs et al., 2019).

Accessibility plays a critical role in ensuring fairness, as differences in prior exposure to digital environments should not influence the validity of assessment outcomes. If test-takers struggle with navigation, controls, or game mechanics due to unfamiliarity rather than lacking the competencies being measured, this could compromise the reliability of the assessment. This highlights the need for GBAs to be designed with inclusivity in mind, incorporating features such as adaptive interfaces, tutorial stages, and user-friendly mechanics to accommodate varying levels of digital proficiency.

Ensuring that GBAs are both engaging and accessible is essential to maintaining fairness, particularly in high-stakes contexts where assessment outcomes have significant consequences. If accessibility is not carefully considered, these assessments risk favouring test-takers with prior digital or gaming experience, ultimately undermining the validity of GBAs as an evaluation tool. To address this issue, future GBA designs should integrate user-friendly interfaces that accommodate test-takers with varying levels of digital proficiency, ensuring that assessment results reflect ability rather than prior technological familiarity.

Overall, games that demand strategic thinking are more likely to foster cognitive engagement, while those that require continuous interaction and responsiveness encourage higher behavioural engagement. This is in line with previous studies on game-based learning, which suggest that engagement is largely context-dependent, influenced by game design and task complexity (Loh et al., 2015; Serrano-Laguna et al., 2017).

From a validity perspective, engagement serves as a prerequisite for test-takers to fully demonstrate the competencies that are being assessed. If a game fails to engage participants, there is a risk that the assessment will primarily measure engagement itself rather than the intended competencies. This issue is especially relevant in serious game-based assessments, where performance outcomes depend not only on a test-takers ability but also on their willingness and motivation to interact in a meaningful way with the game environment. In this study, differences in engagement patterns between game environments suggest that task design plays a critical role in ensuring that performance outcomes reflect actual ability rather than variations in motivation, familiarity, or interaction style.

To maintain construct validity, game-based assessments must balance engagement elements with the competencies they aim to measure. Excessive reliance on engagement-driven mechanics, such as interactions that must be done within a certain period or even gamified rewards, may inadvertently disadvantage test-takers who approach tasks with a more analytical or strategic mindset. On the other hand, assessments that fail to engage the player may lead to underperformance that does not accurately reflect a test-taker's true competency level.

4.2. Strengths, Limitations and Future Research Directions

This study provides valuable insights into the role of engagement in serious game-based assessments, contributing to both educational research and practical applications in high-stakes assessment contexts. One of the key strengths of this study is its integration of log data and self-reported engagement measures, offering a multi-dimensional approach to understanding engagement. By combining objective behavioural metrics with subjective user experiences, a comprehensive assessment of how engagement influences performance in serious gaming environments is provided.

Another strength lies in the use of two distinct serious game environments, Crossroads and Crusade, designed to assess decision-making and problem-solving. The comparative analysis of these environments demonstrates how different game structures and environments impact engagement and performance. These findings contribute to the broader field of digital assessment, offering insights into how interactive and immersive environments influence cognitive and behavioural engagement.

Additionally, this study contributes to the discussion on fairness and accessibility in GBAs, highlighting potential biases introduced by digital literacy differences. By identifying the impact of prior experience on engagement levels, the importance of inclusive design in ensuring that GBAs are valid and accessible for diverse test-taker populations are higlighted.

Despite these strengths, several limitations must be acknowledged. The generalisation of findings is limited due to the specific sample used in this study, which consisted of BOA students. Although relevant to the context of the assessment, this restricts the applicability of results to broader populations. Future research should aim to expand the sample to include professionals from diverse fields, age groups, and cultural backgrounds to ensure greater external validity.

Additionally, the study was confined to two serious game environments, Crossroads and Crusade, designed to assess decision-making and problem-solving in controlled contexts. While these environments provided important insights into engagement patterns, they only captured a subset of the competencies required in professional practice. Both games focused on non-interactive skills, such as task management, logistical planning, and situational judgment. However, in real-world settings, BOAs must also demonstrate interpersonal skills, including communication and conflict resolution, which were not assessed in these game environments.

From a validity perspective, this raises concerns about whether the games fully represent the real-world skills of the profession. While cognitive and behavioural engagement could be measured using log data, the dynamic and unpredictable nature of human interactions in public spaces was not captured. If serious games are to be used as valid assessment tools, they must ensure that the skills measured align with those required in professional contexts, rather than focusing on isolated elements.

To address these validity limitations, future game-based assessments should integrate interactive elements that simulate professional encounters in realistic environments. For example, AI-driven characters could require players to engage in verbal communication, de-escalation, and ethical decision-making during simulated encounters. Similarly, virtual reality (VR) or augmented reality (AR) simulations could create more immersive experiences in realistic public settings. Test-takers could be tasked with navigating crowded areas, interacting with diverse characters, and making situational judgments in different scenarios. Such enhancements would better reflect the social complexity of real-world professional roles, improving the construct validity of GBAs.

Given the impact of engagement on performance, we recommend that future research explores how interactive and more immersive elements affect engagement levels and whether test-takers respond differently to static decision-making tasks versus dynamic, human-centred interactions.

Another limitation is the cross-sectional nature of this study, which captures only a single point of engagement and performance rather than examining how engagement evolves throughout the assessment. Future research should incorporate longitudinal designs to assess long-term skill retention, engagement stability, and behavioural adaptation across multiple testing sessions. Such studies could explore whether repeated exposure to game-based assessments improves test performance and comfort in using alternative assessment environments.

In terms of engagement measurement, this study relied on a combination of log data and self-reported questionnaires, which, while effective, do not capture the full complexity of engagement states. The log data used in this study were originally collected for a different research purpose, meaning that engagement-specific questions and predetermined metrics were not initially included. Future research should integrate more advanced data collection techniques, such as eye-tracking, to assess real-time engagement indicators more precisely (Hookham & Nesbitt, 2019).

Finally, fairness and accessibility remain critical concerns in adopting serious game-based assessments. While GBAs offer many advantages, they also introduce new biases, particularly related to gender and digital literacy. Individuals with limited prior experience in digital environments may face challenges in adapting to game-based assessments, potentially affecting performance and perceived engagement. Future research should investigate how GBAs can be adjusted to mitigate these biases using inclusive and accessible game mechanics.

Overall, the findings highlight the potential of GBAs as a high-stakes assessment tool while also emphasizing the need for inclusive game design to foster engagement and ensure accessibility. Despite the valuable insights gained, limitations in sample diversity, game environments, and engagement measurement call for further research. Future studies should aim to expand participant demographics, explore diverse game structures, and refine engagement tracking methodologies. As GBAs evolve, they offer a promising pathway towards more engaging, competency-driven, and high-stakes performance assessment.

References

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 369–386. https://doi.org/10.1002/PITS.20303

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., De Freitas, S., Louchart, S., Suttie, N., Berta, R., & De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, *46*(2), 391–411. https://doi.org/10.1111/BJET.12113

Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–16. www.educationaldatamining.org,

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bijl, A., Veldkamp, B. P., Wools, S., & de Klerk, S. (2024). Serious games in high-stakes assessment contexts: a systematic literature review into the game design principles for valid game-based performance assessment. *Educational Technology Research and Development*, *72*(4), 2041–2064. https://doi.org/10.1007/S11423-024-10362-0/TABLES/9

Checa, D., & Bustillo, A. (2020). A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications*, *79*(9–10), 5501–5527. https://doi.org/10.1007/S11042-019-08348-9/FIGURES/13

Csikszentmihalyi, M. (2014). Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, 1–298. https://doi.org/10.1007/978-94-017-9088-8

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2018). The design, development, and validation of a multimedia-based performance assessment for credentialing confined space guards. *Behaviormetrika*, *45*(2), 565–596. https://doi.org/10.1007/S41237-018-0064-X/TABLES/6

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Gee, J. P. (2016). Good Video Games and Good Learning. *Good Video Games and Good Learning*. https://doi.org/10.3726/978-1-4539-1162-4

Gorbanev, I., Agudelo-Londoño, S., González, R. A., Cortes, A., Pomares, A., Delgadillo, V., Yepes, F. J., & Muñoz, Ó. (2018). A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. *Medical Education Online*, *23*(1). https://doi.org/10.1080/10872981.2018.1438718

Griffin, P., & Care, E. (2015). Assessment and Teaching of 21st Century Skills. *Assessment and Teaching of 21st Century Skills*. https://doi.org/10.1007/978-94-017-9395-7

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education, 90*, 36–53. https://doi.org/10.1016/J.COMPEDU.2015.09.005

Hookham, G., & Nesbitt, K. (2019). A Systematic Review of the Definition and Measurement of Engagement in Serious Games. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3290688.3290747

Ifenthaler, D., Eseryel, D., & Ge, X. (2012). Assessment for Game-Based Learning. *Assessment in Game-Based Learning*, 1–8. https://www.academia.edu/2763694/Assessment_for_Game_Based_Learning

Kane, M., Crooks, T., & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. https://doi.org/10.1111/J.1745-3992.1999.TB00010.X

Kuhlemeier, H., & Hemker, B. (2007). The impact of computer use at home on students' Internet skills. *Computers & Education*, *49*(2), 460–480. https://doi.org/10.1016/J.COMPEDU.2005.10.004

Messick, S. (1994). ALTERNATIVE MODES OF ASSESSMENT, UNIFORM STANDARDS OF VALIDITY1. *ETS Research Report Series, 1994*(2). https://doi.org/10.1002/J.2333-8504.1994.TB01634.X

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, *178*(10 Suppl), 107–114. https://doi.org/10.7205/MILMED-D-13-00213

Mislevy, R. J., Behrens, J. T., Dicerbo Pearson, K. E., & Levy, R. (2012). Design and Discovery in Educational Assessment: Evidence-Centered Design, Psychometrics, and Educational Data Mining. *Journal of Educational Data Mining*, *4*(1), 11–48. https://doi.org/10.5281/ZENODO.3554641

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2015). Psychometrics and Game-Based Assessment. *Technology and Testing: Improving Educational and Psychological Measurement*, 23–48. https://doi.org/10.4324/9781315871493-2

Mosiane, S., & Brown, I. (2020). Factors Influencing Online Game-Based Learning Effectiveness. *Electronic Journal of Information Systems Evaluation*, *23*(1), pp79-95-pp79-95. https://doi.org/10.34190/EJISE.20.23.1.006

Reschly, A. L., & Christenson, S. L. (2012). Jingle, Jangle, and Conceptual Haziness: Evolution and Future Directions of the Engagement Construct. *Handbook of Research on Student Engagement*, 3–19. https://doi.org/10.1007/978-1-4614-2018-7_1

Roelofs, E. (2019). A Framework for Improving the Accessibility of Assessment Tasks. *Methodology of Educational Measurement and Assessment*, 21–45. https://doi.org/10.1007/978-3-030-18480-3_2/FIGURES/7

Roelofs, E. C., & Visser, J. (2019). De inhoud van toetsen. In P. Sanders (Ed.), *Toetsen op school: Over de inhoud, betrouwbaarheid en validiteit van toetsscores* (pp. 22–45). Cito.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards and Interfaces*, *50*, 116–123. https://doi.org/10.1016/J.CSI.2016.09.014

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016a). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/J.CHB.2016.05.047

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016b). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/J.CHB.2016.05.047

Slater, M. (1999). Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, *8*(5), 560–565. https://doi.org/10.1162/105474699566477

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (n.d.). *Incorporating Student Performance Measures into Teacher Evaluation Systems*. Retrieved 11 September 2024, from www.rand.org

Steinrücke, J., Veldkamp, B. P., & de Jong, T. (2020). Information Literacy Skills Assessment in Digital Crisis Management Training for the Safety Domain: Developing an Unobtrusive Method. *Frontiers in Education*, *5*, 549453. https://doi.org/10.3389/FEDUC.2020.00140/BIBTEX

von der Embse, N. P., & Witmer, S. E. (2014). High-Stakes Accountability: Student Anxiety and Large-Scale Testing. *Journal of Applied School Psychology*, *30*(2), 132–156. https://doi.org/10.1080/15377903.2014.888529

Wang, R., DeMaria, S., Goldberg, A., & Katz, D. (2016). A Systematic Review of Serious Games in Training Health Care Professionals. *Simulation in Healthcare : Journal of the Society for Simulation in Healthcare*, *11*(1), 41–51. https://doi.org/10.1097/SIH.0000000000000118

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*(3), 188–202. https://doi.org/10.1016/J.HRMR.2009.03.007

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019a). The Validity of Technology Enhanced Assessments—Threats and Opportunities. *Methodology of Educational Measurement and Assessment*, 3–19. https://doi.org/10.1007/978-3-030-18480-3_1/TABLES/1

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019b). The Validity of Technology Enhanced Assessments—Threats and Opportunities. *Methodology of Educational Measurement and Assessment*, 3–19. https://doi.org/10.1007/978-3-030-18480-3_1/TABLES/1

Appendix A
Crossroads

De digitale omgeving *Crossroads*

- Ga naar: https://Crossroads.web.app/
- Kies voor Crossroads
- Log in met de code: **PZZCD**

Het is vandaag dinsdag 2 mei. Je gaat tijdens je dienst surveilleren in het centrum van Amersfoort met een collega. De surveillance bereid je individueel voor.

Een aantal praktische zaken om rekening mee te houden:

- In de gemeente Amersfoort is de Model APV van toepassing.
- De gemeente Amersfoort maakt gebruik van de bestuurlijke strafbeschikking.
- Je hebt opsporingsbevoegdheid in domein 1, maar je mag geen geweld toepassen.
- De surveillance duurt van 8:00 – 15:00 (in ieder geval tussen 14:45 en 15:15).
- Jullie zijn op de fiets.
- Jullie zijn het enige koppel op dienst. Meldingen die je overdraagt worden in de volgende dienst meegenomen.

De hoofdtaak voor vandaag is het aanpakken van parkeeroverlast. Ook wordt er binnen de gemeente Amersfoort hard gewerkt aan het verminderen van overlast door afval. Er komen namelijk steeds meer meldingen binnen met klachten over afval (bv. bijplaatsen van afval naast containers).

Voor het plannen van de surveillance krijg je een aantal bronnen toegereikt:

De briefing;

Tickets uit het meldingssysteem van de gemeente Amersfoort;

Mails van o.a. collega's en ketenpartners.

Taakomschrijving

De taak bestaat uit 2 delen:

1. Maak een surveillanceplan op basis van de informatie uit de bronnen.
   - Voor elke taak kies je of je het zelf gaat **PLANNEN**, het gaat **OVERDRAGEN** aan een collega, en/of het wilt **DOORSTUREN** naar een ketenpartner (zie Bijlage 1).
   - Geef per taak aan wat de **PRIO** is, de verwachte **DUUR** en **wat je daar gaat doen**.
   - Zorg ervoor dat de surveillance tussen **14:45 – 15:15** klaar is.
   - Houdt bij het plannen de dagtaak en het gemeentebeleid in gedachten.
   - Als je route klaar is, klik je op **ROUTE KLAAR.**
2. De surveillance is begonnen.
   - Tijdens het lopen krijg je een aantal spoedmeldingen om te verwerken. **Let op, de surveillance moet nog steeds tussen 14:45 – 15:15 klaar zijn**.
   - Als je dat gedaan hebt klik je op **INLEVEREN.** Je bent dan klaar met de opdracht.

Bijlage 1. Ketenpartners

De gemeente Amersfoort heeft een aantal ketenpartners. Wanneer een binnengekomen melding beter past bij een van de ketenpartners, kan je deze doorsturen. De ketenpartners in Amersfoort zijn:

- **Afdeling stedelijk beheer**. Deze afdeling is verantwoordelijk voor de afvalinzameling, riolering en groenvoorzieningen binnen de gemeente. Onder groenvoorziening valt de aanleg van groen (bv. parken, plantsoenen en bermen), het onderhoud van groen, en het afhandelen van meldingen over groen in de gemeente.
- **Afdeling vergunningen**. De afdeling vergunningen is verantwoordelijk voor het beoordelen van aanvragen en het verlenen van vergunningen aan burgers en bedrijven.
- **Politie.** De politie is een belangrijke ketenpartner met andere bevoegdheden dan de handhaver. De politie moet optreden op het moment dat een casus niet (meer) binnen de bevoegdheden ligt van een Boa domein I maar wel binnen die van de politie (bv. bij gevaarzetting).
- **Boa domein II.** Een Boa domein II (Milieu, Welzijn en Infrastructuur) zijn werkzaam als bijvoorbeeld boswachter of bij bouw- en woningtoezicht, milieu inspectie, Rijkswaterstaat of de voedsel- en warenautoriteiten.

Appendix B

General Questionnaire Environments

1. Ik vond De omgeving makkelijk te gebruiken.

☐ Helemaal mee oneens

☐ Oneens

☐ Neutraal

☐ Eens

☐ Helemaal mee eens

2. Ik heb meer hulp nodig om De omgeving te kunnen gebruiken.

☐ Helemaal mee oneens

☐ Oneens

☐ Neutraal

☐ Eens

☐ Helemaal mee eens

3. Ik zou De omgeving vaker willen gebruiken.

☐ Helemaal mee oneens

☐ Oneens

☐ Neutraal

☐ Eens

☐ Helemaal mee eens

4. Ik denk dat het gebruik van de omgeving een negatief effect had op mijn prestatie.

- ☐ Helemaal mee oneens
- ☐ Oneens
- ☐ Neutraal
- ☐ Eens
- ☐ Helemaal mee eens

5. Ik vond de taak in de omgeving leuk om te doen.

- ☐ Helemaal mee oneens
- ☐ Oneens
- ☐ Neutraal
- ☐ Eens
- ☐ Helemaal mee eens

6. De omgeving voelde als een spel.

- ☐ Helemaal mee oneens
- ☐ Oneens
- ☐ Neutraal
- ☐ Eens
- ☐ Helemaal mee eens

7. Ik heb het gevoel dat ik in deze taak dezelfde vaardigheden nodig had als bij het plannen van een surveillanceroute.

- ☐ Helemaal mee oneens
- ☐ Oneens

☐ Neutraal

☐ Eens

☐ Helemaal mee eens

Geef een korte toelichting: ____

8. Ik denk dat mijn prestatie op deze taak een goede inschatting kan geven van mijn vaardigheid om een surveillanceroute te plannen.

☐ Helemaal mee oneens

☐ Oneens

☐ Neutraal

☐ Eens

☐ Helemaal mee eens

Geef een korte toelichting: ____

Einde vragenlijst.

Appendix C
Crusade

De digitale omgeving *Crusade*

- Ga naar: https://Crossroads.web.app/
- Kies voor Crusade
- Log in met de code: **SNDZV**

Volgende week donderdag ga je samen met je klas op excursie naar Amersfoort. Aan jou is gevraagd om te bepalen wat jullie die dag allemaal gaan doen.

Een aantal praktische zaken om rekening mee te houden:

- De bus zet jullie om 9:00 in Amersfoort en komt jullie rond 20:00 weer ophalen. Zorg ervoor dat jullie tussen **19:45 en 20:15** weer terug zijn.
- Jullie zijn in totaal met 20 leerlingen.
- De excursie eindigt met een diner in Amersfoort; houdt bij de keuze van het restaurant rekening met de dieetwensen van je klasgenoten (Bijlage 1).
- Jullie lopen van activiteit naar activiteit.
- Jullie hebben allemaal een lunchpakketje meegekregen, dus met lunchen hoef je **geen** rekening te houden in de planning.

Bijlage 1. Dieetwensen klasgenoten

Dieetwens	**Glutenvrij**	**Vega**	**Vegan**
Aantal klasgenoten	1	3	2

De school betaalt voor alle kosten die gemaakt worden tijdens de excursie, maar om het geld ook echt te krijgen moet de excursie niet alleen maar bestaan uit leuke activiteiten. **De school wilt dat de excursie tenminste drie culturele activiteiten bevat.** Voorbeelden zijn het theater, een museum, een kunsthal, en/of een rondleiding.

Je hebt in totaal **30 minuten** de tijd om de excursie te plannen. Voor het plannen van de excursie krijg je een aantal bronnen toegereikt:

 Een blog over Amersfoort;

Een lijst met informatie en recensies van activiteiten en restaurants in Amersfoort;

Mails van je klasgenoten en docenten.

Taakomschrijving

De taak bestaat uit 2 delen:

1. Maak een dagplanning op basis van de informatie uit de bronnen.
    - Kies voor elke activiteit of je het wel of niet gaan **PLANNEN**.
    - Geef per activiteit in de planning aan wat de verwachte **DUUR** is.
    - Zorg ervoor dat jullie tussen **19:45 - 20:15** weer terug zijn bij de bus.
    - Als je route klaar is druk je op **ROUTE KLAAR**.
2. De excursie is begonnen.
    - Je krijgt wel nog een paar verzoeken om op het laatste moment mee te nemen. **Let op, jullie moeten nog steeds tussen 19:45 en 20:15 bij de bus zijn**.
    - Als je dat gedaan hebt klik je op **INLEVEREN.** Je bent dan klaar met de opdracht.

Appendix D
Computer Questionnaire

1. Naam (voor- en achternaam):

   ____

2. Heb je binnen je opleiding HTV al stage gelopen?

   Ja / Nee

   Als het antwoord ja is:

   a. Heb je in deze stage ervaring opgedaan met het plannen van een surveillance?

      Ja/Nee

   b. Waar heb je stage gelopen?

      ____

Digitale apparaten

3. Heb je thuis een of meer van deze digitale apparaten?

	Ja, en ik gebruik het	Ja, maar ik gebruik het niet	Nee
Desktop (vaste) computer	☐	☐	☐
Laptop	☐	☐	☐
Tablet	☐	☐	☐
Internet	☐	☐	☐
Spelcomputer	☐	☐	☐
Mobiele telefoon (zonder internet)	☐	☐	☐

Mobiele telefoon (met internet)	☐	☐	☐

4. Heb je op school/werk toegang tot een of meer van deze digitale apparaten?

	Ja, en ik gebruik het	Ja, maar ik gebruik het niet	Nee
Desktop (vaste) computer	☐	☐	☐
Laptop	☐	☐	☐
Tablet	☐	☐	☐
Internet	☐	☐	☐

5. Op welke leeftijd gebruikte je voor het eerst een digitaal apparaat?

☐ jonger dan 3 jaar

☐ 4 - 6 jaar

☐ 7 - 9 jaar

☐ 10 - 12 jaar

☐ 13 jaar of ouder

☐ Ik heb voor vandaag nooit een digitaal apparaat gebruikt.

6. Hoe lang gebruik je digitale apparaten op een gewone schooldag/werkdag?

☐ niet

☐ 1 - 30 minuten per dag

☐ 31 - 60 minuten per dag

☐ tussen 1 en 2 uur per dag

☐ tussen 2 en 4 uur per dag

☐ tussen 4 en 6 uur per dag

☐ langer dan 6 uur per dag

7. Hoe lang gebruik je digitale apparaten op een gewone dag in het weekend (buiten school/werk)?

☐ niet

☐ 1 - 30 minuten per dag

☐ 31 - 60 minuten per dag

☐ tussen 1 en 2 uur per dag

☐ tussen 2 en 4 uur per dag

☐ tussen 4 en 6 uur per dag

☐ langer dan 6 uur per dag

8. Hoe vaak gebruik je digitale apparaten voor de volgende activiteiten?

	(Bijna) nooit	1 tot 2 keer per maand	1 tot 2 keer per week	bijna elke dag	elke dag
alleen een spel spelen (single player)	☐	☐	☐	☐	☐
met anderen een (online) spel spelen (multiplayer)	☐	☐	☐	☐	☐
mailen (bijv. Outlook of Gmail)	☐	☐	☐	☐	☐
chatten (bijv. Whatsapp, Telegram)	☐	☐	☐	☐	☐
sociale media (bijv. Facebook, Instagram)	☐	☐	☐	☐	☐
het nieuws lezen	☐	☐	☐	☐	☐

browsen op het internet voor vermaak (bv. filmpjes kijken op YouTube)	☐	☐	☐	☐	☐
praktische informatie opzoeken (bijv. locaties, lestijden, evenementen)	☐	☐	☐	☐	☐
downloaden van muziek, films, spellen of software	☐	☐	☐	☐	☐
tekstverwerking (bijv. voor huiswerk, rapporten, brieven)	☐	☐	☐	☐	☐

Einde vragenlijst.