International Journal of Serious Games

ISSN: 2384-8766
https://journal.seriousgamessociety.org/

Article

Serious Games and Cognitive Assessment. A psychometric approach to serious games analytics

César Mejía[1], Jorge Quimbaya Gómez[2] and Alejandra Herrera-Marmolejo[1]

[1] Laboratorio de Psicología, Universidad de San Buenaventura, Cali, Colombia; [2] Vortex Psychometrics, Cali, Colombia.
camzulua@usbcali.edu.co; ceo@vortexpsychometrics.net; ahmarmol@usbcali.edu.co

Keywords:

Serious Games
Psychometrics
Serious Games Analytics
Executive Functions
Game Based Assessment

Received: August 2023
Accepted: June 2024
Published: August 2024
DOI: 10.17083/ijsg.v11i3.667

Abstract

This study introduces a psychometric approach to serious games analytics, specifically focusing on a video game designed for the assessment of executive functions. The study presents methodologies for data capture and processing, employing reliability analysis and exploratory factor analysis (EFA) within the validation process. Serious games have emerged as valuable tools for cognitive assessment, particularly in domains such as executive functions. However, there is a need for robust methodologies to validate the effectiveness of these games as assessment tools. In this study we address the question: How can psychometrics be applied to serious game analysis in a video game designed to assess executive function? Utilizing psychometrics and task analysis, the study conceptualizes, and validates a measurement instrument embedded within a serious game. The evaluation process includes assessing the reliability and factorial structure of the test, testing its construct validity. The study demonstrates that serious games integrate seamlessly with psychometric techniques. This innovative approach contributes to the advancement of serious game analytics and has implications for enhancing cognitive assessment methodologies in various domains.

1. Introduction

Serious video games are digital games created with the primary purpose of improving players' skills and performance through training and instruction [1], [2]. They are not designed only to be played for fun but are designed for educational and skills development purposes. In this context, some form of evaluation is needed to ensure that serious games are effective in improving players' skills and performance [1]. In this sense, serious video games have a clearly defined pedagogical or training objective. In these contexts, it is usually necessary to provid some kind of evidence of how close trainees have come to achieving the outcomes expected by the educational agent. In the case of serious video games, this type of information can be used as evidence of the effectiveness of the application. According to this line of reasoning, if the purpose of serious games is learning, then those serious games that incorporate a mechanism for performance evaluation would be better equipped to fulfil their purpose.

In addition, the same information obtained from the performance evaluation can be used to provide feedback to the user at the precise moment when certain actions are performed during the game, thus enriching the learning [3] and gaming experience. This kind of feedback is fundamental in the teaching-learning process to guide the learner's actions. The game-based assessment then opens the possibility for flexible and in-real-time feedback to users in serious games. Furthermore, these same metrics can be used to optimize game design, by providing information on the performance of multiple learners facing the same tasks.

If assessment tools are used in serious games, then some questions can be raised about these assessment tools. A key question is: how do can we know that these metrics are actually capturing the information for which they were designed? This is the problem of validity that psychometrics has been studied for decades. For example, suppose we have a serious game for teaching some mathematical operations and concepts, and we record the number of successes. Suppose further that the task involves the actual manipulation of symbols (letters, numbers, etc.). In such cases, it could assumed that the score obtained by the users confuses arithmetic performance with reading performance. Ultimately, the validity of measurement instruments consists of gathering empirical evidence that the instrument measures the trait or attribute it is hypothesized to measure, and that it does so in a reliable manner.

There are different techniques and conceptual frameworks for validity research. Although we will not discuss these different perspectives here, it is necessary to mention factor analysis, as this is the approach used in this study. Factor analysis, in general terms, is a technique aimed at testing the hypothesis of a certain factor structure of the latent variables, which should be reflected in the way the scores are grouped [57]. If the results obtained do not form a structure consistent with the theory, this is considered strong evidence that the test is not valid (null hypothesis). Specifically, Exploratory Factor Analysis is recommended with novel instruments [4], [5], [6], [7], [8] and this will be the technique employed in this study. For this purpose, we designed a psychometric test embedded in a serious video game for the assessment of executive functions.

Currently many researchers [9], [10], [10], [11], [12], [13], [14] privilege the executive function (EF) model proposed by Miyake et al. [15] inhibition, shifting (flexibility), and updating (working memory). From this perspective, when people resolve a problem, information can be obtained based on the ability to reject distractions and resist prepotent response (inhibition), the ability to switch smoothly between tasks, routines, or contexts (flexibility) and the ability to maintain and mentally manipulate multiple ideas to achieve a goal (working memory). Additionally, in Luria's model [16], planning is considered a main component of EF. Planning refers to the people's ability to analyse what kind of procedures should be performed to achieve a goal, while monitoring their actions, keeping the plan in working memory and inhibiting irrelevant actions outside the goal [17], [18]. In this sense, serious video games could be very suitable for assessing EF.

Karr et al. [19] note that the popularization of the three-factor EF model has led researchers to focus on testing it to the exclusion of other types of models. In their research they found six possible factorial models of EF. In terms of dimensionality, researchers such as Arán [20], Brocki and Bohlin [6] and Wiebe et al. [8] suggest that children's neuropsychological development should be considered as a control variable, since factorial models seem to be influenced by the timing of executive functions development. The model proposed by Miyake et al. [15] was based on a sample of adolescents and adults, whereas studies with preschool children, such as those by Huizinga et al. [7] and Best et al. [10], suggest unifactorial and two-factorial models. Studies with adolescents have found more evidence has been found for the three-factor model, including inhibition, flexibility, and monitoring [21], [22], [23].

The central point of this study is that we are attempting to carry out such a validation of a psychometric test, in the context of an adventure video game. This brings us to the field of game analytics, or more specifically, serious games analytics. This is a novel approach that does not seem to be fully resolved in the specialized literature [2], [24]. Thus, the question guiding this study is: How can psychometrics be applied to serious games analytics in a video game designed to assess executive functions?

1.1. Related works

Serious video games may incorporate multiple measurement systems with different purposes, based on different information-gathering techniques. Smith et al. [25] found in a meta-analysis that the highest proportion of information collected in serious games occurs after or before the game (79%), while the information collected during the game accounts for the lowest proportion of publications in the field (21%). In turn, they found that techniques such as interviews, focus groups, questionnaires, and various types of observation are commonly used. Interviews and focus groups tend to be particularly useful in post-game evaluations, to ask about the user's experience and satisfaction with the game.

Learning, on the other hand, is more accurately recorded using questionnaires or indirect observation during gameplay. According to Smith et al. [25], indirect observation does not requiere the presence of an observer and, since it can be programmed to operate during gameplay, "is particularly attractive for serious games as event logs can be tailored to collect any pertinent information; for example, task sequences, task completion times, and/or percentage of task completion" (pp. 35).

Accordingly, in-game techniques seem to be the most appropriate for measuring learning and cognitive processes in serious video games. One of the strongest lines of research in technology and cognitive assessment is gamified neuropsychological testing. In this case, classical cognitive assessment tasks are used in computerized environments to measure various cognitive abilities. There are several arguments in favour of these tools: The design can be more eye-catching with graphics, animations, sounds and other gamification features. They provide a more attractive and motivating experience for participants, and data can be obtained more accurately collected through automated capture on computers or tablets [26]. In addition, they are thought to have the potential to increase ecological validity by creating virtual scenarios that more closely resemble natural environments [27], [28]. Ecological validity is important, among other things, because cognitive tasks have been criticized in this regard because they are often simplified and tightly controlled, emphasizing on internal rather than external validity [1], [2], [3], [24], [25], [29].

Vladisauskas, et al. [28] developed a freely available computer-based task software for children aged 4 to 8 years called Mate Marote. It consists of a battery of five tests traditionally used in neuropsychology to train and assess executive functions. They found that the children's performance was in line with what was expected in terms of developmental milestones reported in the literature. They concluded that this computerized battery can effectively and reliably assess cognitive performance in a typical classroom setting and without the presence of researchers. TOWI is also an instrument based on standardized neuropsychological tests that uses gamification [26]. It was developed to measure working memory, planning, inhibition, sustained and selective attention. They found age as a predictor of performance and similarities in performance between TOWI and normative data.

These technological developments also include testing in virtual reality devices. For example, there is Aula-Nesplora [30], a gamified test in virtual reality that simulates a school classroom. It was designed to assess items of attention and executive functions such as selective and sustained attention, motor activity, impulsivity, and reaction times. Díaz-Orueta et al., [31] found significant correlations between Aula-Nesplora and the Continuous Performance Test

(CPT) in a group of children of average cognitive ability diagnosed with attention deficit/hyperactivity disorder (ADHD).

Transforming an existing test into a gamified version is an interesting alternative. However, the possibilities offered by video games when it comes to creating scenarios and tasks are unlimited. It is possible, for example, to create open-world scenarios with tasks that occur simultaneously, just like in everyday life. In these cases, scores needed to be obtained in a different way, since they are not discrete and sequential items, as in traditional psychometric tests. It is in these types of games where that serious game analytics could be useful.

In the dynamic field of serious video games, there is a growing interest in implementing various data analysis techniques. These include game analytics, serious game analytics, telemetry, learning analytics and game-based assessment stand out. In this context, Loh et al. [24] highlight a fundamental distinction between game analytics and serious game analytics. While game analytics focuses on the identification of game patterns in order to adjust difficulty, challenges and thus improve the game experience, serious game analytics prioritizes the measurement and assessment of performance in education and training contexts to improve the learning process.

According to Kang et al. [32] and Loh et al. [24], serious game analytics involves extracting information from gameplay data within a serious game in order to measure, evaluate, or improve performance. In these settings, students' actions and behaviours are tracked in real time through numerical variables known as in-situ data, as opposed to ex-situ data, such as self-reported surveys or any data collected outside the game system. Kang et al. [32] also state that traditional achievement tests and data mining techniques that lack theoretical guidance often fail to fully capture how students learn complex skills in a gaming context. Squire [33] highlights that one of the challenges of open-ended serious games is identifying the progression of students' learning, as they may take different paths to solve a problem. According to Kang et al. [32] it is important to understand how information is collected in serious games, as this determines the type of analysis that can be performed on the data. This approach combines statistical analysis with data mining to examine learning patterns among students with different levels of experience.

Kiili et al. [34] present a comprehensive array of in-game metrics designed to assess various aspects of player performance within serious games. These metrics include evaluations of overall game performance based on task accuracy, comparison of performance between users, and order and time of task completion. Each task category is examined, providing insights into how players approach and solve the challenges presented within the game environment. The study also examines the maximum level achieved by players, providing an indication of their progression within the game's hierarchical structure. Taken together, these metrics contribute to an understanding of player engagement, strategy, and achievement within serious games, facilitating evaluation and refinement of game design and educational effectiveness.

Validating learning and neuropsychological assessment instruments is a challenge for serious game analytics. Gibson and Freitas [35] point out that traditional psychometric models face difficulties in these contexts, leading to uncertainty about the validity of the measures. In this scenario, the use of exploratory and iterative psychometric approaches is a promising option. Currently, there is a group of studies identified by Raković et al. [36] that includes theoretical, methodological, and data analytic contributions aimed at improving the validity of assessment in learning analytics methods. The authors argue that the combination of game-based assessment and learning analytics methods is a promising approach to assess students' cognitive skills. This argument makes it possible to think about game-based assessment beyond learning itself, as this approach makes it possible to measure cognitive processes that underlie all learning.

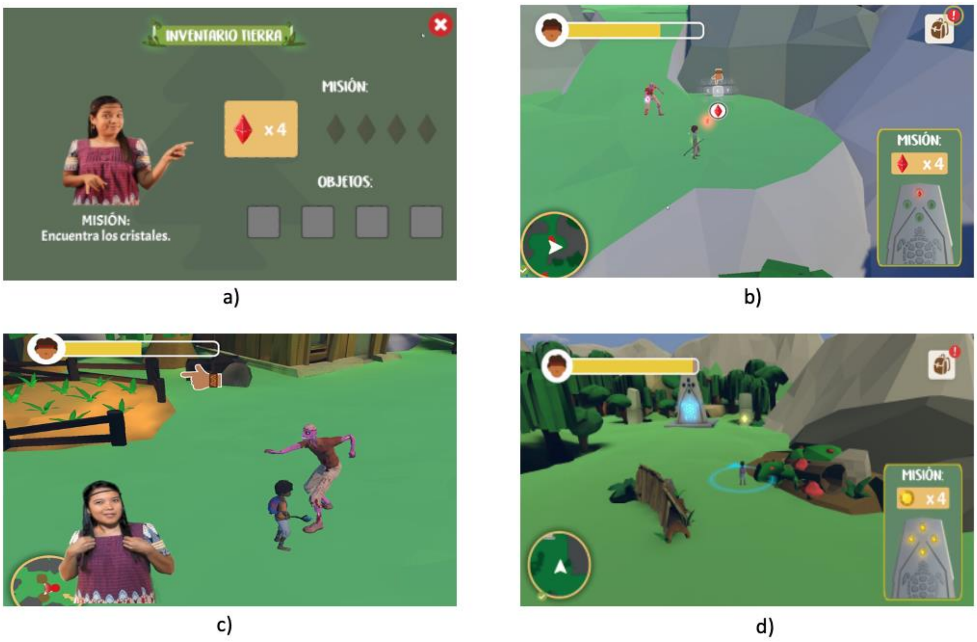
The purpose of this study is to present a psychometric approach to serious game analytics of a video game designed to assess executive functions. To our knowledge, this is the first time this has been done. We will present data collection and processing procedures, as well as a series of psychometric procedures used in an exploratory validation process, including, reliability analysis, and exploratory factor analysis.

2. Methods and Material

The methodological design of this study is based on an exploratory validation phase of a psychometric test embedded in a video game. Visor 2.0 is not intended as commercial software, but rather it should be considered as a laboratory prototype. This is important because we are not presenting a finished version of a psychometric tool. Instead, we are interested in describing the basic concept of the design and presenting an exploratory phase of the validation. On an empirical level, we seek to test the hypothesis of a factorial structure that groups the game scores into different domains of executive functioning. Conceptually, we seek to approach serious games analytics using psychometric procedures to assess the reliability and validity of the measures.

2.1. Instruments

Visor 2.0 is a serious video game that features an embedded psychometric test [37]. The video game scenarios are in 3D and the software was developed using the Unity engine. It is a strategy and adventure video game with six levels of a general scenario (Figure 1). The instructions are presented in audio (Spanish) and in Colombian Sign Language. The story of the game takes place in a village, where people are suffering from a curse that turns them into zombies. The main mission is to collect some objects (crystals), then to open a portal, and solve a challenge to transform the zombies of that scenario back into human beings.



MISIÓN:
Encuentra los cristales.

a)

b)

c)

d)

Figure 1. Screenshots illustrating different moments of the game: a) Inventory of collected crystals and mission instructions; b) The player collects a red crystal corresponding to the mission; c) Encounter with an adversary (zombie); d) The player approaches the portal to complete the mission.

There are six stages in all, with the same structure, but variations in the map, objects (shape, colour, and location) and adversaries (number and location). Due to of the limited time to play, not all players will make it to the final stage. In each of these six stages, the same items were recorded (e.g., number of zombies destroyed, or number of wrong crystals collected). To obtain a single score for each item for each participant, the corresponding scores were added together.

Table 1 shows the 24 items included in the factor analysis, after removing items with low discrimination levels (see Results). Classification into domains, prior to Exploratory Factor Analysis (EFA), was carried out by the research team based on the task analysis methodology for video games [38]. All items included aim to assess executive functions, which are divided into four sub-domains: planning, inhibition, flexibility, and monitoring.

Table 1. Items of Executive Functions in Visor 2.0

Item	Operational Definition	Monitoring	Planning	Inhibition	Flexibility
Target crystal	Target crystals collected			x	
Opening the backpack	Frequency of access to backpack menu	x			x
Non-target crystals	Non-target crystals collected			x	
Shoots	Number of times power (o the staff) is cast			x	
Evasion	Number of times the player exits the chase	x			
Fountain	Visits to the fountain to restore health	x			
Successful evasion	It is recorded when 10 seconds elapse after evasion, without activating the pursuit again	x			
Zombie mode	Total depletion of the first life bar (human mode). Equals ¨zombification¨1 + 2				x
Completed levels	Number of missions successfully completed		x		x
Type 1 impact	Impact a zombie, leaving him frozen (non-lethal impact)	x	x		
Type 2 impact	Lethal impact to a zombie			x	
Impacts received	Impacts received (in human or zombie mode)	x			x
Chasing	Total number of initiated chases by (any number) of zombies	x	x		
Portal	Entrance to portals		x		x
Successful hidden stealth	Enter the zombie's range of vision, covered by a barrier, and then exit the zombie's range of interaction undetected		x		
Failed hidden stealth	Enter the zombie's range of vision, covered by a barrier, and then move into the zombie's range of vision (activates pursuit)		x		
Successful stealth	Leave the zombie's interaction radius undetected	x	x		
Failed stealth	Move from zombie interaction range to viewing range		x		
Suicidal stealth	Move from zombie interaction range to proximal range (activates unavoidable attack)		x		
Target crystal visits	Direct contact with target crystals		x		
Visits to non-target crystals	Direct contact with non-target crystals			x	
Portal visits	Approaches to the portals			x	
Health regenerators	Collection of crosses for health restoration	x			
Static zombie activated	Activation of zombies at rest	x	x		

2.2. Procedure

The dataset used in this study comes from a larger dataset, as the research project includes other instruments and procedures that are not discussed here. However, as the data collection and filtering are relevant, we describe the earlier stages of the research project here to provide context for the data set extracted for this study. In the first phase of the project, gameplay tests were conducted using qualitative observation techniques. Although these qualitative results will not be discussed in this study as they are not directly related to the validation process, it should be mentioned that they played an important role in the correction of the software. This information provided the development team with an indication of the extent to which the game would be comprehensible to users. Tests were also carried out with university students to test the stability of the software and to detect possible bugs in the gameplay, in the online server where the databases were hosted and in the connection between the game and the databases.

The debugging of the software through this process resulted in the current version of the software, which was used with a large sample of school children. The application of the video game was carried out in the educational institutions, in group sessions. The groups consisted of between 7 and 10 users per session. Each user had three game sessions of about 30 minutes each, on consecutive days. The group application was chosen in order to obtain a larger sample. To control the game environment, all participants were equipped with headphones, while the professionals of the research team controlled the possible interactions between the children. All this was done by a neuropsychologist, accompanied by a multimedia engineer.

For this study, a selection of cases was made from the database. Initially, only cases of neurotypical children were selected. As the database contains records of deaf children, these cases were excluded because the factorial structure of the items could be influenced by the particularities of their development. Finally, the age of the participants was restricted to between 5 and 12 years, and sessions lasting more than 45 minutes were excluded.

Finally, with regard to the selection of the items tested in this study, the scores corresponding to direct user actions were selected, without taking into account the duration (time stamps) of the events. This resulted in 32 items.

2.3. Data set and Participants

The sample obtained for this study consisted of 171 participants (88 boys and 83 girls) between the ages of 5 and 12 years (M=8.27, SD=1.589) who were between in the first to fifth year of primary school (M= 3.1, SD= 1.456). Participants were randomly selected from six educational institutions. The participants' teachers did not report any diagnosis of developmental or neurological disorders. An informed consent form was signed by both the participants' families and the educational institutions.

2.4. Statistical analyses

The database was debugged by excluding trials that lasted longer than 45 minutes of play. This criterion was assumed because some sessions could be left open, or some participants could spend an excessive amount of time in a single attempt. The participants trials (1128) were scored by summing all trials. In the end, the data set contained a total of 171 cases. The statistical treatment included, first, an item discrimination analysis based on the item-total correlation. Secondly, the Kaiser-Meyer-Olkin sample adequacy test and Bartlett's sphericity test were carried out as a preliminary step to the exploratory factor analysis. Thirdly, the reliability analysis was carried out using Cronbach's alpha and McDonald's omega coefficients of the factors.

Finally, the fit indices for factor analysis were examined: root mean squared error of approximation (RMSEA), comparative fit index (CFI) and H-latent, using the bootstrap sampling technique with 5000 samples simulated from the original sample, which allows a robust calculation of factor loadings and correlations. The statistical software JASP (2023, version 0.17.2) [38] was used for item discrimination analysis and reliability calculations, and the software FACTOR (version 10.10.3) [40] was used for the elaboration of the Exploratory Factor Analysis (EFA).

3. Results

Before the EFA, a selection of scores was made by domain. They were divided into planning, inhibition, monitoring, and flexibility items. Items with a low level of discriminability were excluded, i.e., whose item-total correlation coefficient did not exceed 0.35, following the recommendations of Willoughby et al. [41].

The Inhibition domain originally had 9 items. Based on the item-total correlation, 3 items were eliminated. The standardized Cronbach's alpha for this set of items was found to be 0.898 and McDonald's omega was found to be 0.897. Of the 11 monitoring items, one item was dropped. The standardized Cronbach's alpha for this set of items was set at 0.931 and McDonald's Omega at 0.934. Of the 15 planning items, four were eliminated. The standardized Cronbach's alpha for this set of items was set at 0.961 and McDonald's omega at 0.962. Of the six flexibility items, one item was dropped. The standardized Cronbach's alpha for this set of items was set at 0.881 and the McDonald's omega at 0.888.

Table 2. Items count in the various validation phases.

Phase	# of items
Initial items	32
Items eliminated due to low discrimination	8
Selected items for the EFA	24
Items eliminated due to low factor loadings	6
Items in the final factorial model	18

After item discrimination analysis, 24 items remained. Then the first factorial explorations show that eight items significantly affected the EFA model, due to their low communalities and loadings on the factors (<0.2). Finally, the EFA was performed with the remaining 18 items. Following the recommendations of Field and Miles [42] recommendations, only items with factor loadings greater than 0.40 were included. The robust unweighted least squares (RULS) extraction method with an oblique rotation (Oblimin direct) was used. The Kaiser-Meyer-Olkin measure verified the adequacy of the sampling for the analysis (KMO = 0.9147), indicating a value considered excellent according to Field and Miles [42]. Bartlett's test of sphericity [χ^2 (153) = 1878.5, p <.001] indicated that the correlations between items were sufficiently large for the EFA. Two components had eigenvalues above the Kaiser criterion of 1 and together explained 79.89% of the variance (see Table 3). The reliability analysis of the set of items included in the EFA yielded excellent results (McDonald's Omega=0.973; Cronbach's Alpha=0.933 CI=0.918 - 0.947).

Table 3. Factor loadings after rotation

	Items	Inhibition/ Planning	Monitoring/ flexibility
v1	Type 1 impact	0.950	
v12	Target crystal	0.919	
v15	Portal	0.893	
v11	Target crystal visits	0.891	
v13	Visits to non-target crystals	0.856	
v18	Completed levels	0.841	
v17	Static zombie activated	0.777	
v2	Received impacts	0.718	
v6	Suicidal stealth	0.713	
v14	Non-target crystals	0.709	
v16	Portal visits	0.672	
v5	Failed stealth	0.542	0,498
v3	Chasing		0.972
v4	Successful stealth		0.756
v7	Successful hidden stealth		0.637
v8	Failed hidden stealth		0.407
v9	Evasion		0.988
v10	Successful evasion		0.940
	Eigenvalues	12.113	1.599 10.167 0.988 0.882
	% Variance	69.732	
	H-Latent	0.986	
	α	0.887	

Note: Omitted loadings below 0.40. Extraction method: Robust unweighted squares (RULS). Oblimin Direct Rotation. According to Ferrando and Lorenzo-seva, (2017) values >.80 in H-Latent suggest a well-defined latent variable.

In terms of goodness of fit, the root mean square residual approximation (RMSEA) was less than 0.01, while the comparative goodness of fit (CFI), goodies of fit index (GFI) and overall goodness of fit index (AGFI) were greater than 0.99. The values obtained in this study indicate an excellent fit, according to the parameters proposed by Ferrando and Anguiano (RMSEA <0.05) [42], Fleming and Merino (AGFI >0.95) [43] and Lai and Green (GFI - CFI >0.95) [44]. As can be seen in Figure 2, the grouped items suggest that factor 1 (RC1): represents Inhibition and Planning and factor 2 (RC2): Monitoring and Flexibility.

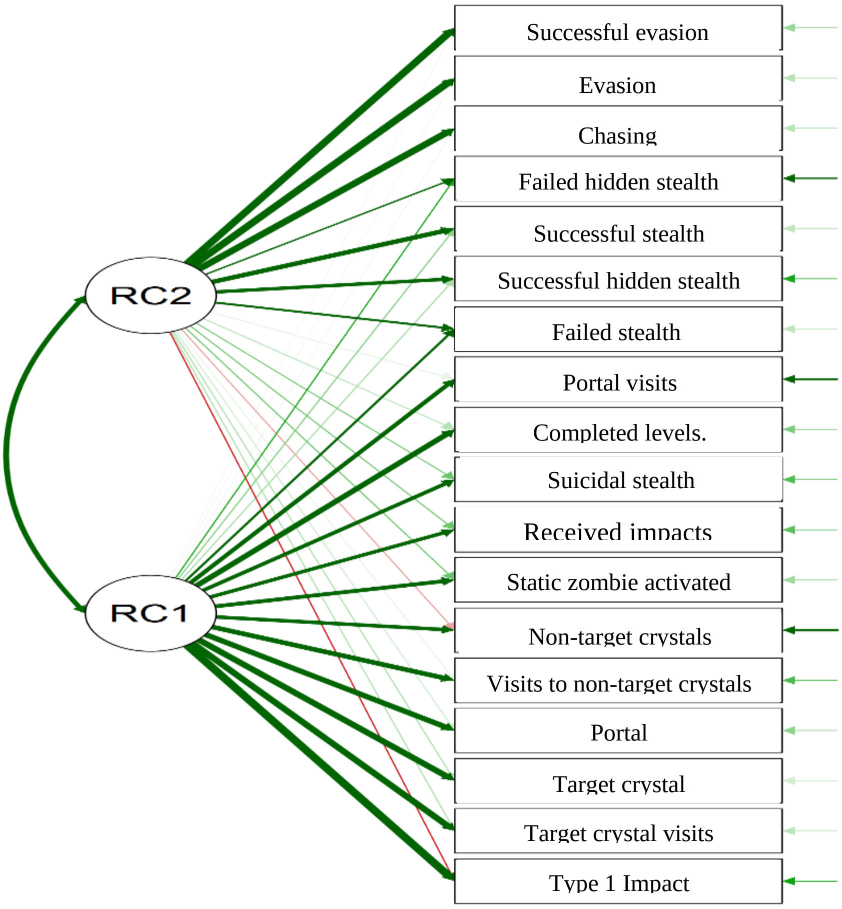


Figure 2. Visor 2.0 Exploratory Factor Analysis Diagram.

4. Discussion

In this study, we have approached serious games analytics from a psychometric perspective. The discussion is divided into three topics: Data collection during gameplay, identification of the factorial structure, and ecological validity.

4.1 Data capture during gameplay

The first challenge focused on information gathering, exploring how and what data should be collected during the player's interactions with the video game. Methods such as task analysis and gameplay analysis were used. These methods are in line with the perspective of Smith et al. [25], who suggest that learning is most accurately captured through indirect observations during gameplay. Indirect observation, according to these authors, do not require an observer and can be programmed to operate during gameplay, thus avoiding interfering with the user's gameplay experience. In addition, Guardiola and Natkin [45] highlight that collecting data specifically during game sessions ensures the authenticity of players' responses and behaviours.

Bakhtiari and Habibzadeh [2] emphasize the need to evaluate the quality of the serious games. Guardiola and Natkin [45] stress the importance of scientifically validating the psychological model, to ensure that the measures obtained during the gameplay accurately reflect the constructs being assessed. During the development of Visor 2.0, we applied task analysis and then operationalized variables (items) to collect scores from distributed tasks in a virtual scenario.

The design of Visor 2.0 has sought to capture the essence of strategy video games, and some of the activities that are typical of the genre. In this way, the aim is to achieve better gameplay and a greater player engagement. From the point of view of the design of the test items, this entails a possibility that may seem counter-intuitive at first glance, as the operationalization of the items is drawn from preconceived activities. For example, stealth is a recurring aspect in this type of game and consists of performing an activity without being detected by adversaries. In Visor 2.0 we used task analysis to identify the cognitive processes involved in this activity, which allowed us to extract executive function items.

Once the data had been collected and with regard to the procedure prior to the EFA, it was found that, with information from a number of trials per scenario, the first decision to be made was how to choose the way to summarize the scores. Summation was chosen as the optimal method for grouping the data, as demonstrated by the first reliability analyses by domain, since it provided a higher level of reliability than the use of the average. However, we are aware that reducing multiple scores to a single measure is not the most interesting option in terms of dynamics, variable trajectories showing ups and downs, advances and setbacks, and intrasubject variability, a condition of the change and learning process [46].

4.2 Factorial structure of executive functions in Visor 2.0

Looking at user data in Visor 2.0 a factor structure emerges that is consistent with a hypothesized structure of executive functions. This is considered important evidence in the process of validating an instrument given that "The degree to which the subtest cluster in patterns that align with a working hypothesis or theory provides one form of evidence that the subtest actually reflects the constructs" [4, pp. 290].

In terms of construct validity, executive functions, as a broad concept of different cognitive functions associated with the frontal lobes, have been the subject of discussion regarding the dimensionality of the construct [23], [47], [48], [49]. Our results show a multidimensional model, consisting of the inhibition and planning factors correlated with the monitoring factor fused with flexibility. With regard to factor analysis of executive functions, it has been reported that attentional inhibition and action inhibition tasks are correlated and form a single factor [43]. Factor 1 in our data consists of a group of items related to inhibition and planning. This finding is consistent with the approach of Brocki and Bohlin [6], who claim that inhibition is the most important element for the development of other executive functions, especially during school age. Factor 1 items are related to inhibition because they are precision tasks that require a higher level of inhibition in actions. According to Diamond [50], Brocki and Tillman [51], planning tasks require greater inhibitory control.

With respect to the items that the analyses suggested be removed from the factorial model, it was observed that these were actions that were not fundamental to winning the game. This analysis highlights the importance of the functionality of the game for the effectiveness of the neuropsychological test items. In this sense, if the gameplay does not activate certain actions during the game, the items associated with these specific actions may not play the intended role or may have a negative impact on the test variance. It is important to recognize that the integral functioning of the gameplay is critical to the success and usefulness of the test items; otherwise, these items may lose their effectiveness and relevance in measuring the target constructs.

4.3 Ecological validity

In addition to the construct validity discussed above, ecological validity is an important aspect of this discussion, as it is one of the most salient aspects of gamified tests and serious games. Part of the design of Visor involved introducing cognitive challenges into a technological device popular with children: computers and video games. Lumsden et al. [52] claim that gamified versions of traditional tests that incorporate elements characteristic of games (such as narrative, graphics, goals, and rules), can be more engaging and motivating for the participants [53]. This is important from a neuropsychological assessment point of view, as it favours eliciting maximum possible performance. Furthermore, gamified tasks may be more attractive to the users because of the balance between challenge and reward [54].

Video game scenarios usually recreate activities like those that users perform in their daily lives. This aspect is one of the most prominent in the use of serious video games. For Parsons [27], neuroscience research often employs simple, static stimuli that lack several potentially important aspects of real-world activities and interactions. He notes that there is a growing interest in human neuroscience in multimodal scenarios and highlights the potential of neuropsychological assessment in digital environments to improve ecological validity in neuroscience. In the case of Visor 2.0, the narrative, which focuses on a character who must save the world from the zombie curse, allows the different challenges presented throughout the game to be connected and promotes integration, task meaning, and children's motivation. Although some of the missions presented in Visor 2.0 are not everyday activities (e.g., facing a zombie), problem solving in video games is similar to real life activities in some important aspects: humans face simultaneous tasks, choose among a number of possibilities, make decisions that involve emotions, and keep useful information in memory to act on at another time. In strategy video games, the player's actions each time lead to a new configuration of the game states and update the dilemmas before which the player has to make a new choice (recharge life, freeze or kill the zombie, hide, etc.).

5. Conclusions

In this study we approach serious games analytics using some techniques from contemporary psychometrics. This approach can be a contribution to the field of serious games, and to serious games analytics, both in the process of conceptualizing and designing of measurement instruments, and their validation process.

Since the psychometric test is embedded in the video game, it is crucial in this model to test the gameplay. This is because the items are actions that are recorded directly in the flow of the game. If the gameplay fails, the test items will not work properly.

Reliability analyses show a good internal consistency of the items in Visor 2.0 and, together with the item discrimination analysis, allow the identification of scores that are uncorrelated with the rest of the scores in each conceptual category of the test. These uncorrelated scores measure something else (if anything). In such cases, the items could be corrected or removed. This practice allows the adjustment of the instrument, thereby improving its reliability.

In addition, the EFA technique provided valuable information about the factor structure of the test. This allows the working hypothesis regarding the grouping of the items, thus contributing to the assessment of construct validity. For serious games developers, this could be a useful technique for testing the validity of their assessment tools.

However, there are some limitations to this study. External measures are not used in this study. The lack of information on the convergence of Visor 2.0 with standard tests of EF may be considered as a limitation. Although the EFA can stand alone for validation purposes, convergent validation could provide additional and useful evidence.

Acknowledgments

This study was partially financed by the Departamento Nacional de Ciencia y Tecnología – Minciencias (CT-036-2018).

Conflicts of interest

There are no conflicts of interest.

2. References

[1] M. M. Hellström, D. Jaccard, and K. E. Bonnier, “A systematic review on the use of serious games in project management education,” *International Journal of Serious Games*, vol. 10, no. 2, pp. 3–24, 2023, doi: https://doi.org/10.17083/ijsg.v10i2.630.

[2] R. Bakhtiari and Z. Habibzadeh, “Designing a framework and validating a tool for evaluating the educational quality of serious games: a meta-synthesis,” *IJSG*, vol. 10, no. 2, pp. 61–83, Jun. 2023, doi: 10.17083/ijsg.v10i2.576.

[3] E. Pacheco-Velazquez, V. Rodes-Paragarino, L. Rabago-Mayer, and A. Bester, “How to Create Serious Games? Proposal for a Participatory Methodology,” *IJSG*, vol. 10, no. 4, pp. 55–73, Nov. 2023, doi: 10.17083/ijsg.v10i4.642.

[4] V. Arán, “Funciones ejecutivas en niños escolarizados: efectos de la edad y del estrato socioeconómico,” *Avances En Psicología Latinoamericana*, vol. 29, no. 1, pp. 98–113, 2011.

[5] J. R. Best, P. H. Miller, and L. L. Jones, “Executive functions after age 5: Changes and correlates,” *Developmental Review*, vol. 29, no. 3, pp. 180–200, Sep. 2009, doi: 10.1016/j.dr.2009.05.002.

[6] K. C. Brocki and G. Bohlin, “Executive Functions in Children Aged 6 to 13: A Dimensional and Developmental Study,” *Developmental Neuropsychology*, vol. 26, no. 2, pp. 571–593, Oct. 2004, doi: 10.1207/s15326942dn2602_3.

[7] M. Huizinga, C. V. Dolan, and M. W. Van Der Molen, “Age-related change in executive function: Developmental trends and a latent variable analysis,” *Neuropsychologia*, vol. 44, no. 11, pp. 2017–2036, Jan. 2006, doi: 10.1016/j.neuropsychologia.2006.01.010.

[8] S. A. Wiebe, K. A. Espy, and D. Charak, “Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure.,” *Developmental Psychology*, vol. 44, no. 2, pp. 575–587, Mar. 2008, doi: 10.1037/0012-1649.44.2.575.

[9] N. Bardikoff and M. Sabbagh, “The differentiation of executive functioning across development: Insights from developmental cognitive neuroscience,” in *New perspectives on human development*, New York, NY, US: Cambridge University Press, 2017, pp. 47–66. doi: 10.1017/CBO9781316282755.005.

[10] J. R. Best and P. H. Miller, “A Developmental Perspective on Executive Function,” *Child Dev*, vol. 81, no. 6, pp. 1641–1660, 2010, doi: 10.1111/j.1467-8624.2010.01499.x.

[11] F. Collette, M. Hogge, E. Salmon, and M. Van der Linden, “Exploration of the neural substrates of executive functioning by functional neuroimaging,” *Neuroscience*, vol. 139, no. 1, pp. 209–221, Apr. 2006, doi: 10.1016/j.neuroscience.2005.05.035.

[12] N. Garon, S. E. Bryson, and I. M. Smith, “Executive function in preschoolers: a review using an integrative framework,” *Psychol Bull*, vol. 134, no. 1, pp. 31–60, Jan. 2008, doi: 10.1037/0033-2909.134.1.31.

[13] T. A. Niendam, A. R. Laird, K. L. Ray, Y. M. Dean, D. C. Glahn, and C. S. Carter, “Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions,” *Cogn Affect Behav Neurosci*, vol. 12, no. 2, pp. 241–268, Jun. 2012, doi: 10.3758/s13415-011-0083-5.

[14] U. Müller and K. Kerns, “The development of executive function,” in *Handbook of child psychology and developmental science: Cognitive processes, Vol. 2, 7th ed*, Hoboken, NJ, US: John Wiley & Sons, Inc., 2015, pp. 571–623. doi: 10.1002/9781118963418.childpsy214.

[15] A. Miyake, N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager, “The Unity and Diversity of Executive Functions and Their Contributions to Complex ‘Frontal Lobe’ Tasks: A Latent Variable Analysis,” vol. 41, no. 1, pp. 49–100, 2010, doi: 10.1006/cogp.1999.0734.

[16] A. Luria, *El cerebro en acción*. 1988.

[17] A. Diamond, “Executive Functions,” *Annu. Rev. Psychol.*, vol. 64, no. 1, pp. 135–168, Jan. 2013, doi: 10.1146/annurev-psych-113011-143750.

[18] J. M. Fuster, “Frontal lobe and cognitive development,” *J Neurocytol*, vol. 31, no. 3–5, pp. 373–385, 2002, doi: 10.1023/a:1024190429920.

[19] J. E. Karr, C. N. Areshenkoff, P. Rast, S. M. Hofer, G. L. Iverson, and M. A. Garcia-Barrera, “The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies.,” *Psychological Bulletin*, vol. 144, no. 11, pp. 1147–1185, Nov. 2018, doi: 10.1037/bul0000160.

[20] V. Arán and M. B. López, “Estructura Latente de las Funciones Ejecutivas en Adolescentes: Invarianza Factorial a través del Sexo,” *Av. Psicol. Latinoam.*, vol. 35, no. 3, p. 615, Sep. 2017, doi: 10.12804/revistas.urosario.edu.co/apl/a.4724.

[21] O. E. Arango, I. C. Puerta, and D. A. Pineda, “Estructura factorial de la Función ejecutiva desde el dominio conductual,” *Divers.: Perspect. Psicol.*, vol. 4, no. 1, Jun. 2008, doi: 10.15332/s1794-9998.2008.0001.05.

[22] J. C. Flores, R. E. Castillo-Preciado, and N. A. Jiménez-Miramonte, “Desarrollo de funciones ejecutivas, de la niñez a la juventud,” *analesps*, vol. 30, no. 2, pp. 463–473, May 2014, doi: 10.6018/analesps.30.2.155471.

[23] F. Xu, Y. Han, M. A. Sabbagh, T. Wang, X. Ren, and C. Li, “Developmental Differences in the Structure of Executive Function in Middle Childhood and Adolescence,” *PLoS ONE*, vol. 8, no. 10, p. e77770, Oct. 2013, doi: 10.1371/journal.pone.0077770.

[24] C. Loh, S. Yanyan, and D. Ifenthaler, “Serious Games Analytics: Theoretical Framework,” 2015, pp. 3–30. doi: 10.1007/978-3-319-05834-4_1.

[25] S. P. Smith, K. Blackmore, and K. Nesbitt, “A Meta-Analysis of Data Collection in Serious Games Research,” in *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, C. S. Loh, Y. Sheng, and D. Ifenthaler, Eds., in Advances in Game-Based Learning. , Cham: Springer International Publishing, 2015, pp. 31–55. doi: 10.1007/978-3-319-05834-4_2.

[26] M. F. Gómez-Tello, M. F. Rosetti, M. Galicia-Alvarado, C. Maya, and R. Apiquian, “Neuropsychological screening with TOWI: Performance in 6- to 12-year-old children,” *Applied Neuropsychology: Child*, vol. 11, no. 2, pp. 115–124, Apr. 2022, doi: 10.1080/21622965.2020.1764357.

[27] T. D. Parsons, “Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences,” *Front Hum Neurosci*, vol. 9, p. 660, 2015, doi: 10.3389/fnhum.2015.00660.

[28] M. Vladisauskas *et al.*, “The Long and Winding Road to Real-Life Experiments: Remote Assessment of Executive Functions with Computerized Games—Results from 8 Years of Naturalistic Interventions,” *Brain Sciences*, vol. 14, no. 3, Art. no. 3, Mar. 2024, doi: 10.3390/brainsci14030262.

[29] J. Jylkkä *et al.*, “Assessment of goal-directed behavior with the 3D videogame EPELI: Psychometric features in a web-based adult sample,” *PLoS One*, vol. 18, no. 3, p. e0280717, 2023, doi: 10.1371/journal.pone.0280717.

[30] G. Climent and F. Banterla, *Nesplora Aula Manual.*, Segunda Edición. 2016. [Online]. Available: https://www.researchgate.net/publication/329371095_AULA_NESPLORA_evaluacion_de_los_procesos_atencionales

[31] U. Díaz-Orueta, C. Garcia-López, N. Crespo-Eguílaz, R. Sánchez-Carpintero, G. Climent, and J. Narbona, “AULA virtual reality test as an attention measure: convergent validity with Conners’ Continuous Performance Test,” *Child Neuropsychol*, vol. 20, no. 3, pp. 328–342, 2014, doi: 10.1080/09297049.2013.792332.

[32] J. Kang, M. Liu, and W. Qu, “Using gameplay data to examine learning behavior patterns in a serious game,” *Computers in Human Behavior*, vol. 72, pp. 757–770, Jul. 2017, doi: 10.1016/j.chb.2016.09.062.

[33] K. Squire, *Open-Ended Video Games: A Model for Developing Learning for the Interactive Age*. The MIT Press, 2008. [Online]. Available: https://cdn-educators.brainpop.com/wp-content/uploads/2013/10/squire-open-ended-games-macarthur-salen.pdf

[34] K. Kiili, K. Moeller, and M. Ninaus, “Evaluating the effectiveness of a game-based rational number training - In-game metrics as learning items,” *Computers & Education*, vol. 120, pp. 13–28, May 2018, doi: 10.1016/j.compedu.2018.01.012.

[35] D. Gibson and S. de Freitas, “Exploratory Analysis in Learning Analytics,” *Tech Know Learn*, vol. 21, no. 1, pp. 5–19, Apr. 2016, doi: 10.1007/s10758-015-9249-5.

[36] M. Raković, D. Gašević, S. Hassan, J. A. Ruupérez, N. Aljohani, and S. Milligan, “Learning analytics and assessment: Emerging research trends, promises and future opportunities,” *British Journal of Educational Technology*, pp. 1–435, 2023, doi: 10.1111/bjet.13301.

[37] C. Mejía, A. Herrera-Marmolejo, M. Rosero-Pérez, J. Quimbaya, and J. F. Cardona, “Design of a video game for assessment of executive functions in deaf and hearing children,” *Applied Neuropsychology: Child*, vol. 0, no. 0, pp. 1–8, 2024, doi: 10.1080/21622965.2024.2311096.

[38] M. Rodríguez, J. Quimbaya, and D. Varón, “Bonus II: Análisis de tarea.Super Mario Sunshine,” in *I/O Videojuegos,Computadoras y seres humanos*, Editorial Bonaventuriana, 2009.

[39] “Introducing JASP 0.17.2.” Amsterdam University, 2023. Accessed: Jan. 22, 2024. [Online]. Available: https://jasp-stats.org/2023/05/30/jasp-0-17-2-blog/

[40] “Factor Analysis.” Accessed: Jan. 22, 2024. [Online]. Available: https://psico.fcep.urv.cat/utilitats/factor/

[41] M. Willoughby, S. J. Holochwost, Z. E. Blanton, and C. B. Blair, “Executive Functions: Formative Versus Reflective Measurement,” *Measurement: Interdisciplinary Research and Perspectives*, vol. 12, no. 3, pp. 69–95, Jul. 2014, doi: 10.1080/15366367.2014.929453.

[42] A. Field and J. Miles, *Discovering Statistics Using SAS: (and Sex and Drugs and Rock ‘n’ Roll)*. 2010.

[43] P. J. Ferrando and C. Anguiano-Carrasco, “El análisis factorial como técnica de investigación en Psicología,” *Papeles del Psicólogo*, vol. 31, no. 1, pp. 18–33, 2010.

[44] K. Lai and S. B. Green, “The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree,” *Multivariate Behavioral Research*, vol. 51, no. 2–3, pp. 220–239, May 2016, doi: 10.1080/00273171.2015.1134306.

[45] E. Guardiola and S. Natkin, “A Game Design Methodology for Generating a Psychological Profile of Players,” in *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, C. S. Loh, Y. Sheng, and D. Ifenthaler, Eds., in Advances in Game-Based Learning. , Cham: Springer International Publishing, 2015, pp. 363–380. doi: 10.1007/978-3-319-05834-4_16.

[46] H. Sánchez, E. Cerchiaro, and M. Guevara, “Cambio y variabilidad: un marco de referencia en los estudios sobre el primer año de vida,” *Acta Colombiana de Psicología*, vol. 16, no. 1, pp. 101–113, Jun. 2013.

[47] R. Chan, D. Shum, T. Toulopoulou, and E. Chen, “Assessment of executive functions: Review of instruments and identification of critical issues,” *Archives of Clinical Neuropsychology*, vol. 23, no. 2, pp. 201–216, Mar. 2008, doi: 10.1016/j.acn.2007.08.010.

[48] M. Huizinga, D. Baeyens, and J. A. Burack, “Editorial: Executive Function and Education,” *Front. Psychol.*, vol. 9, p. 1357, Aug. 2018, doi: 10.3389/fpsyg.2018.01357.

[49] P. D. Zelazo, L. Qu, and A. C. Kesek, “Hot executive function: Emotion and the development of cognitive control,” in *Child development at the intersection of emotion and*

cognition, American Psychological Association, 2010, pp. 97–111.

[50] A. Diamond, “The Early Development of Executive Functions,” in *Lifespan Cognition: Mechanisms of Change*, 2006, pp. 70–95. doi: 10.1093/acprof:oso/9780195169539.003.0006.

[51] K. C. Brocki and C. Tillman, “Mental Set Shifting in Childhood : The Role of Working Memory and Inhibitory Control,” *nfant and Child Development*, vol. 23, no. 6, pp. 588–604, 2014, doi: /10.1002/icd.1871.

[52] J. Lumsden, E. A. Edwards, N. S. Lawrence, D. Coyle, and M. R. Munafò, “Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy,” *JMIR Serious Games*, vol. 4, no. 2, p. e11, Jul. 2016, doi: 10.2196/games.5888.

[53] I. A. Chicchi, C. de Juan, E. Parra, and M. Alcañiz, “Are 3D virtual environments better than 2D interfaces in serious games performance? An explorative study for the assessment of executive functions,” *Applied Neuropsychology: Adult*, vol. 28, no. 2, pp. 148–157, Mar. 2021, doi: 10.1080/23279095.2019.1607735.

[54] H. Peters, A. Kyngdon, and D. Stillwell, “Construction and validation of a game-based intelligence assessment in minecraft,” *Computers in Human Behavior*, vol. 119, p. 106701, Jun. 2021, doi: 10.1016/j.chb.2021.106701.

[55] I. Bombín-González *et al.*, “Validez ecológica y entornos multitarea en la evaluación de las funciones ejecutivas,” *RevNeurol*, vol. 59, no. 02, p. 77, 2014, doi: 10.33588/rn.5902.2013578.

[56] G. Climent, P. Luna-Lario, I. Bombín-González, A. Cifuentes-Rodríguez, J. Tirapu-Ustárroz, and U. Díaz-Orueta, “Evaluación neuropsicológica de las funciones ejecutivas mediante realidad virtual,” *RevNeurol*, vol. 58, no. 10, p. 465, 2014, doi: 10.33588/rn.5810.2013487.

[57] J. Rust and S. Golombok, Modern Psychometrics: The Science of Psychological Assessment, 4th ed. London: Routledge, 2020. doi: 10.4324/9781315637686.