

A Serious Game for Soft Skills Assessment in Human Ressources: A Convergent Validity Study

Maxime Boutrouille, Léo Fichet, Jérôme Dinet

Submitted to: JMIR Serious Games
on: October 17, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 47

Figures 48

Figure 1..... 49

Figure 2..... 50

Figure 3..... 51

Figure 4..... 52

Figure 5..... 53

Figure 6..... 54

Figure 7..... 55

A Serious Game for Soft Skills Assessment in Human Ressources: A Convergent Validity Study

Maxime Boutrouille^{1*} PhD, MD; Léo Fichet^{1*} MD; Jérôme Dinet^{2*}

¹ Yuzu Metz FR

²LPN Campus Lettres et Sciences Humaines Université de Lorraine Nancy FR

*these authors contributed equally

Corresponding Author:

Maxime Boutrouille PhD, MD

Yuzu

2 RUE MAURICE BARRES

Metz

FR

Abstract

Background: Soft skills are increasingly valued in the labor market, yet their assessment remains challenging. Traditional approaches (eg, interviews, questionnaires, cognitive tests) often face limitations related to subjectivity, ecological validity, and bias. Serious games have emerged as promising alternatives by offering immersive environments and behavioral data collection.

Objective: This study aimed to evaluate the convergent validity of a serious game (Yuzu) designed for soft skills assessment in human ressources, by comparing three modules assessing active-empathic listening, decision-making under uncertainty, and teamwork dialogue with their respective reference psychometric tools.

Methods: A sample of 39 participants completed both the Yuzu modules and their psychometric reference versions: the Active-Empathic Listening Scale (AELS), the Iowa Gambling Task (IGT), and a simplified SYMLOG questionnaire. Statistical analyses included correlation, concordance, equivalence testing, and descriptive comparisons.

Results: Strong correlations were found between the Yuzu listening module and the AELS, with high agreement demonstrated through Bland-Altman and concordance analyses. The decision-making module showed statistical equivalence with the IGT, replicating exploration-exploitation dynamics. In contrast, the dialogue-based teamwork module failed to demonstrate convergent validity with the SYMLOG questionnaire due to ceiling effects and social desirability bias.

Conclusions: These findings support the potential of gamified assessments to provide ecologically valid and engaging measures of soft skills. Adaptations of standardized tools (AELS, IGT) preserved validity while enhancing participant experience. However, dialogue-based assessments require further refinement through richer branching scenarios and adaptive responses. This study contributes to bridging psychology, ergonomics, and game design and supports continued validation of serious games in professional selection contexts.

(JMIR Preprints 17/10/2025:86054)

DOI: <https://doi.org/10.2196/preprints.86054>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/86054>
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/86054>



Original Manuscript

A Serious Game for Soft Skills Assessment in Human Ressources: A Convergent Validity Study

Abstract

Background. Soft skills are increasingly valued in the labor market, yet their assessment remains challenging. Traditional approaches (e.g., interviews, questionnaires, cognitive tests) often face limitations related to subjectivity, ecological validity, and bias. Serious games and gamification have emerged as promising alternatives by offering immersive environments and behavioral data collection.

Methods. This study examined the convergent validity of *Yuzu*, a serious game designed to assess soft skills through three gamified modules: (1) active-empathic listening, (2) decision-making under uncertainty, and (3) teamwork via dialogue choices. A sample of 39 participants completed both the gamified modules and their psychometric reference versions: the Active-Empathic Listening Scale (AELS), the Iowa Gambling Task (IGT), and a simplified SYMLOG questionnaire. Statistical analyses included correlation, concordance, equivalence testing, and descriptive comparisons.

Results. Strong correlations were found between *Yuzu*'s listening module and the AELS, with high agreement demonstrated through Bland-Altman and concordance analyses. The decision-making module showed statistical equivalence with the IGT, replicating exploration–exploitation dynamics. In contrast, the dialogue-based teamwork module failed to demonstrate convergent validity with the SYMLOG questionnaire, due to ceiling effects and social desirability bias.

Conclusion. These findings support the potential of gamified assessments to provide ecologically valid and engaging measures of soft skills. Gamified adaptations of standardized tools (AELS, IGT) preserved validity while enhancing participant experience. However, narrative-based dialogue

assessments require further refinement, particularly through richer scenarios and adaptive responses. This study contributes to bridging psychology, ergonomics, and game design, and calls for continued validation of serious game in professional contexts.

KEYWORDS : soft skills; assessment; serious game; gamification; video game; validity



1. Context

In the era of digitalization and the transition toward Industry 5.0, work environments are undergoing profound transformations, with routine and technical tasks increasingly automated. While technical skills remain essential for carrying out activities across all sectors and professional domains, employers and recruiters are placing growing emphasis on the evaluation of non-technical skills, or soft skills [1].

As a result, soft skills are becoming central not only as a factor of individual adaptability in uncertain and volatile contexts, but also as a lever of organizational competitiveness [2]. An analysis of a corpus of 19,000 companies shows that transversal skills such as critical thinking, problem-solving, creativity, cognitive flexibility, and communication now account for more than 40% of the competencies required in job postings [3]. These inherently behavioral and contextual skills help distinguish professionals who are able to integrate, innovate, and collaborate within complex work dynamics.

Scientific interest in soft skills has also grown exponentially over the past two decades. A bibliometric analysis of 753 publications indexed in Scopus (1999-2021) shows an average annual growth rate of 20.5%, with the number of articles rising from just two in 1999 to more than 120 in 2021 [4]. This trend reflects the increasing recognition of soft skills as a central topic, both in research and practice.

However, this growing importance raises a major challenge at the core of contemporary research: how can soft skills be assessed in a valid, objective, and reliable manner in selection contexts? Traditional methods (i.e., interviews, self-report questionnaires, cognitive tests, etc.) face limitations in terms of subjectivity and ecological transferability. In this context, gamification, and more

specifically the use of serious games, emerges as a promising avenue. By mobilizing immersive and interactive environments, these tools make it possible to simulate dynamic situations and observe, in situ, behaviors that reveal transversal competencies [5].

The objective of this study is therefore to examine the extent to which gamification, and more specifically a serious game, can constitute a valid and objective approach to assessing soft skills, in comparison with traditional methods.

This cross-disciplinary approach, at the intersection of psychology, ergonomics, and management sciences, aims to address a dual need: to provide HR professionals with strategic decision-making tools, and to offer researchers a robust methodological framework for measuring non-technical skills.

Our contribution is structured as follows: Section 1 contextualizes soft skills, including a review of the main existing assessment approaches, and presents and justifies the design of Yuzu, an innovative serious game for soft skills assessment. Section 2 details the study methodology. Section 3 presents the quantitative results. Section 4 discusses the findings and study limitations. Section 5 concludes the paper.

1.1. Soft skills and their importance in the workplace

Today, there is no consensus on the definition of soft skills [3,6], which are often contrasted with hard skills [7]. First mentioned in U.S. Army documents [8], the term gradually gained popularity, to the point that its definition has become blurred [9]. Some scholars describe soft skills as a combination of interpersonal and intrapersonal competencies [10,11]; others as interpersonal skills, thinking skills, and personal skills [12]; and still others as people management skills [13].

Despite their ubiquity in job postings, soft skills lack a standardized nomenclature: thousands of terms are used to describe them, making their definition and categorization challenging [3,14]. The

notion thus remains vague and heterogeneous, encompassing qualities, traits, values, and attitudes that are sometimes very different from one another [9]. Moreover, soft skills are sometimes referred to as non-technical skills [15], transversal skills [16], adaptive skills [17], or socio-cognitive skills [17].

Given this plurality, we adopt the definition proposed by Haselberger et al. [18], which encompasses multiple dimensions and has been widely cited [19-21]: *“Soft Skills represent a dynamic combination of cognitive and meta-cognitive skills, interpersonal, intellectual and practical skills.”* According to the authors, *“Soft skills help people to adapt and behave positively so that they can deal effectively with the challenges of their professional and everyday life”*.

Technological advances have further reinforced their importance: while machines excel at routine tasks, flexibility, creativity, and judgment remain uniquely human capacities [1,22]. Soft skills enable practitioners to make sense of the dynamic, social, and complex contexts of projects [23]. In technology-driven fields, competencies such as critical thinking, problem-solving, communication, and creativity remain essential, and achieving a balance between human and digital skills is crucial to adapting to technological change [3,24].

Nevertheless, despite their widely recognized importance, soft skills raise a central question: how can they be measured? The following section addresses different approaches to measurement, ranging from what individuals say about themselves (explicit approaches) to what they actually do in practice (implicit approaches).

1.2. Existing methods for assessing soft skills

The assessment of soft skills has become a critical issue in recruitment, management, and training. Unlike hard skills, however, soft skills are less tangible and more subject to interpretation, as well as to various biases, particularly on the part of recruiters. Nevertheless, several methods have been

developed to evaluate them as reliably as possible.

1.2.1. Qualitative Methods

The job interview remains the most popular tool for assessing candidates' soft skills [25].

Unstructured interviews resemble free-flowing conversations: questions vary depending on the candidate and are not based on a fixed framework. This lack of structure generates biases (e.g., halo effect, illegal questions) and leads to low predictive validity [26]. Some authors have even highlighted that unstructured interviews can be worse than conducting no interview at all [27-29].

In contrast, structured interviews rely on standardized questions (behavioral or situational) asked to all candidates and scored using a rubric. This standardization improves inter-rater agreement and reduces biases. It allows for the evaluation of soft skills through a behavioral approach (i.e., asking about past experiences) or a situational approach (i.e., asking how one would react to a hypothetical scenario). When meticulously conducted by a well-trained evaluator, this method improves the prediction of future performance [26]. Moreover, as Wingate and Bourdage [25] point out, interviews serve not only an evaluative function: they also contribute to recruitment by providing an attractive image of the organization, and to candidate socialization by offering information about the position and establishing an initial relationship.

However, while structured interviews are subject to fewer biases than unstructured ones, they are by no means bias-free. Structured interviews remain subjective, with interviewers introducing their own biases (e.g., confirmation bias, similarity bias) and judgments (e.g., personal prejudices), which undermines the reliability of the assessment [30]. Interviews may also facilitate direct or indirect discrimination, for instance on the basis of gender [31], age [32], personality (extraversion vs. introversion) [28], or physical appearance [33].

Blinded interviews, which anonymize candidate information prior to or during the interview, can eliminate certain biases [26], particularly in the early stages of soft skills evaluation. However, candidates are still exposed to the aforementioned biases once direct contact occurs in later recruitment phases. On another front, AI-driven interviews for assessing soft skills may offer more objective evaluations, less prone to human bias [34]. Yet, due to their lack of transparency (black box) and the risks of algorithmic bias and discrimination, concerns remain [35].

Peer assessment is another method used to evaluate soft skills, whereby individuals of similar status or expertise (peers) evaluate one another based on predefined criteria [36]. This approach offers several advantages: it fosters reflexivity and critical thinking among evaluators [36], increases motivation and engagement, and makes feedback appear more direct and less hierarchical [37].

Nevertheless, its reliability strongly depends on the number of evaluators [38]. It is also vulnerable to socio-cultural biases: evaluators may be lenient toward peers, avoid reporting poor performance, or temper their judgments due to cultural norms of politeness [39]. Other studies reveal demographic biases, with less favorable evaluations of students from racial minorities [37]. Finally, because peer assessment often relies on normative grids or prescribed models, it may leave little room for the emergence of creative or culturally diverse practices.

In sum, the assessment of soft skills through interviews or peer assessment remains subjective, depends heavily on the evaluator's skills, and is highly sensitive to bias. Furthermore, it may disadvantage certain populations.

1.2.2. Psychometric questionnaires

Self-report questionnaires are sometimes used to measure soft skills. However, due to the wide diversity of components associated with this concept [3], their measurement remains complex.

Several recent instruments can nevertheless be cited:

- Soft Skills Questionnaire for Nurses (SSQ-N) [40]: a 25-item questionnaire assessing communication, management skills, emotional intelligence, and confidentiality. Analyses revealed high content validity and satisfactory internal consistency ($\alpha = 0.824$).
- SKILLS-in-ONE [41]: a 74-item inventory assessing 13 soft skills (e.g., communication, leadership, creativity). Reliability was established through Cronbach's alpha and McDonald's omega.
- [42] Inventory: a 49-item tool covering six domains (self-determination, resilience, empathy, assertiveness, social support, teamwork), validated through factor analysis.

The self-report method is relatively easy to implement, particularly because of its easy interpretability, richness of information, motivation to report, causal force, and sheer practicality [43,44].

However, it suffers from the same limitations as other methods: “anchoring effects, primacy and recency effects, time pressure, and consistency motivation” [43]. Additional biases include socially desirable responding, acquiescent responding, and extreme responding [45]. Moreover, such questionnaires often measure self-reported traits rather than actual performance.

1.2.3. Cognitive tests applied to soft skills

Cognitive tests assess certain general abilities (e.g., reasoning, working memory, perception) or decision-making processes, which are sometimes associated with soft skills (e.g., problem-solving). For several decades, they have constituted a benchmark tool in recruitment, with Raven's Progressive Matrices being a classic example [46]. Their effectiveness lies in their high predictive validity: a meta-analysis of more than 20,000 studies [47] reported an average correlation of $r = 0.50$

between cognitive ability and job performance. Similarly, Schmidt and Hunter [48] found that general intelligence is one of the best predictors of performance ($r = 0.51$), comparable to structured interviews and slightly below work sample tests ($r = 0.54$). In a later article, Schmidt and Hunter [49] broadened this conclusion, showing that the benefits extend beyond the workplace: cognitive ability also predicts academic success, socioeconomic status, and career advancement. These findings explain their central role in selection processes.

However, meta-analyses also show that cognitive ability tests often generate negative reactions among candidates and may be perceived as poorly related to the job, which can affect organizational attractiveness [50]. Moreover, Leutner et al. [51] emphasize that traditional cognitive tests carry major disadvantages in terms of fairness. They produce the largest ethnic group differences among common selection methods [52,53]. In addition, they frequently induce a negative candidate experience: test anxiety, stereotype threat, and perceived lack of relevance (face validity) may undermine both performance and acceptability [54,50].

1.2.4. Serious game and gamification

Given the limitations of traditional tests, serious games appear to represent a promising avenue for assessing soft skills, as they allow behaviors to be observed in situ and go beyond the narrow scope of self-reports.

For example, Leutner et al. (2023) proposed a game-based assessment of cognitive abilities. Pedersen et al. [55] developed Skill Lab, a set of large-scale validated mini-games designed to predict various cognitive abilities through behavioral modeling. XBadges [56] relies on commercial video games (e.g., *Tetris*, *Pac-Man*, *Flappy Bird*) to measure competencies such as perseverance, spatial reasoning, and risk-taking, complemented by analyses of players' emotional expressions.

Other studies emphasize the explicit evaluation of soft skills through narrative environments. For

instance, Hernandez et al. [57] proposed a model in which dialogue choices in a visual novel are used to measure and train players' soft skills. This approach models narrative decisions to identify traits such as empathy or initiative-taking. Similarly, Altomari et al. (2023) introduced *Among the Office Criticality*, a serious game based on a branching decision tree, in which players' choices are translated into performance indicators for problem-solving, time management, and emotional regulation. Pouezevara et al. (2019) likewise demonstrated that digital games can assess a wide range of dimensions, including reliability (task completion, time management), learning skills (compliance with instructions), and interpersonal abilities such as self-regulation.

Finally, several recent studies have explored the use of serious games to assess emotional intelligence [59] or to develop algorithmic models for real-time behavioral analysis [60].

To summarize, numerous comparative studies between gamified and non-gamified tools, partly synthesized by Leutner et al. (2023), show that serious games provide a dual added value. On the one hand, compared with traditional paper-based or computer-based tests, they offer a more immersive and engaging experience [61,62], reduce evaluation-related anxiety, and enhance the perceived attractiveness of the assessment tools [63-65]. On the other hand, unlike self-report questionnaires, which are limited to subjective perceptions, serious game environments allow for the dynamic observation of behaviors (e.g., response sequences, reaction times), while maintaining psychometric rigor comparable to classical methods [5,58]. Finally, some studies suggest that they may also mitigate social biases, for example by reducing stereotype threat effects on performance [66].

However, while gamification appears to be an innovative and complementary approach to traditional recruitment tools, several limitations remain. First, the predictive validity of gamified assessments with respect to actual job performance is still insufficiently documented: although convergent

validity with traditional tests has been demonstrated, it is not yet clear whether these formats retain the same predictive power for real-world performance [51]. Second, some candidates express doubts about the perceived relevance (face validity) of the games, particularly when tasks appear too abstract or disconnected from the job-related skills expected. It should also be noted that, since it is difficult to isolate precisely which soft skill is activated or measured at each stage, serious games often adopt a broad approach (e.g., mixing leadership, communication, creativity), which can make evaluation ambiguous or imprecise. In addition, the “black box” effect of algorithms (e.g., machine learning models) may reduce the acceptability and trust of users or evaluators [51]. Finally, the development of serious games is often time-consuming and costly.

Table 1. Summary of the main advantages and limitations of soft skills assessment methods in professional contexts

Assessment Method	Advantages	Limitations
Unstructured interview	Flexibility, freedom of expression, natural exchanges [25]	High subjectivity, halo/affinity biases, low predictive validity [26]
Structured interview	Standardized questions, higher inter-rater reliability [26]	Candidate-related biases and interviewer-related biases [30]
		Risks of direct/indirect Behavioral and situational discrimination (gender, age, approaches to target soft personality, appearance) [28, skills [26] 31-33]

	Socialization and employer branding functions [25]	
Blinded interviews	Reduction of some identity-related biases [26]	Biases reappear at later stages involving direct contact [26]
AI interview	Potential for objectivity, reduction of human biases [34]	Algorithmic bias risks, lack of transparency [35]
Peer assessment	Encourages reflexivity and critical thinking among evaluators [36]	Reliability depends on number of evaluators [38] Socio-cultural biases [39]
	Increases motivation and engagement [37]	
	Perceived as more direct and less hierarchical than external evaluation [37]	
Questionnaires & self-assessments	Demonstrated psychometric validity and easy administration [44]	Anchoring, social desirability, and response biases [43,45]
		Assess reported traits rather than actual performance [43]
		Potentially stressful context
Cognitive tests	High predictive validity for	Adverse impact: ethnic group

	job performance ($r \approx .50$)	differences	[52,53]
	[47,48]		
		Ethical, financial, and legal	
	Predict academic success, concerns		
	socioeconomic status, and		
	career progression [49]	Negative candidate experience	
		(anxiety, low face validity)	
	Standardization	and [50,54]	
	objectivity [46]		
Serious games & gamification	Immersive and engaging	Insufficiently demonstrated	
	experience, reducing	predictive validity [51]	
	evaluation-related anxiety		
	and enhancing the perceived	Few tools validated	
	attractiveness of assessment	scientifically [51]	
	tools [61-65]		
		Doubts about perceived	
	Enable dynamic observation	relevance (face validity) [51]	
	of behaviors (e.g., response		
	sequences, reaction times)	Algorithmic “black box” effect	
	while maintaining [51]		
	psychometric rigor		
	comparable to classical	Time-consuming and costly	
	methods [5,58]	development [51]	
	May help mitigate social	Not always perceived as fun	

biases, for instance by
reducing stereotype threat
effects on performance [66]

Despite notable advances, the assessment of soft skills continues to face several limitations. Traditional methods largely rely on self-assessments, and thus on subjective perceptions. Cognitive tests, while robust predictors of general performance, fail to capture interpersonal or behavioral dimensions in real-world contexts. Serious games appear to be a promising avenue, as they allow the collection of innovative behavioral data while offering engaging and attractive experiences for candidates, thereby reducing biases and improving acceptance. However, empirical validations of such tools remain scarce.

1.3. Presentation of Yuzu and research question

In light of these findings, the guiding question of our study is as follows: Can a serious game be developed that, while maintaining psychometric rigor, provides a multimodal and immersive assessment of soft skills tailored to the needs of the professional world?

Designed to address this challenge, a tool such as Yuzu appears particularly relevant. Yuzu is a serious game for soft skills assessment that integrates multiple complementary modalities within a single narrative environment. The player embodies a crew member in a futuristic space station, interacts with non-player characters, solves cognitive tasks, and makes decisions under constraints. Every action is automatically recorded through embedded trackers, allowing observed behaviors to be linked to precise psychometric indicators (e.g., decision times, choice sequences, interaction profiles).

Yuzu distinguishes itself from existing tools in three ways. First, it integrates multiple assessment modalities within a single 3D environment (i.e., gamified questionnaire, cognitive task, and dialogue choices inspired by psychosocial models), thus enabling a multimodal measurement of soft skills. Second, Yuzu is built around immersion and engagement, with realistic environments developed in Unreal Engine 5 (Figure 1), where players can freely navigate and explore. Finally, Yuzu immerses users in a science fiction narrative, in which they embody the captain of a space station and where all their actions are automatically analyzed using indicators (trackers) designed in collaboration with work psychologists, ergonomists, and video game specialists. The objective is to ensure both scientific rigor and a credible gameplay experience for candidates.



Figure 1. Actual images from the Yuzu serious game

Accordingly, this study aims to investigate the convergent validity of three gamified modules (questionnaire, cognitive task, dialogue choices) integrated into Yuzu, in order to determine which provide a reliable and transferable measurement of soft skills. It follows the empirical study of Dinét et al. [67], which demonstrated that a serious game adaptation of classical psychometric tests (e.g., WCST) is feasible within Yuzu.

2. Methods

2.1. Research objective

The objective of this study is to evaluate the convergent validity of three gamified approaches to soft skills assessment integrated into a serious game: (1) an active listening questionnaire (inspired by the *Active Empathic Listening Scale* - AELS) [68], (2) a decision-making task (inspired by the *Iowa Gambling Task* - IGT) [69], and (3) a dialogue-choice measure (SYMLOG) [70].

More specifically, the aim is to determine the extent to which these modalities yield results comparable to their reference versions, and whether they can constitute reliable and acceptable alternatives for soft skills assessment.

2.2. Research hypotheses

To address this objective, one general hypothesis and three operational hypotheses were formulated:

- **H1.** The gamified modules of Yuzu demonstrate convergent validity with their respective psychometric reference measures.
- **H1a.** Scores from Yuzu's gamified active-empathic listening module positively correlate with the AELS questionnaire.
- **H1b.** Performance in Yuzu's gamified decision-making module is equivalent to performance observed in the reference version.
- **H1c.** Dimensions from Yuzu's gamified teamwork module positively correlate with those measured by the SYMLOG questionnaire.

2.3. Sample

The sample included 23 women (59.0%) and 16 men (41.0%), with a mean age of 27.79 years ($SD = 7.96$, range = 20-54). Regarding education and occupational status, 41.0% were students, 35.9% executives or managers, 15.4% employees, and a small proportion were retired (2.6%), entrepreneurs (2.6%), or job seekers (2.6%). Participants reported a relatively high computer literacy ($M = 7.91$,

SD = 2.11) and moderate to high familiarity with video games ($M = 6.67$, $SD = 2.69$). Table 2 presents the sociodemographic characteristics of the sample ($N = 39$).

Table 2. Sample Characteristics ($N = 39$)

Variable	N (%) / M (SD)
Genre	Female = 23 (59.0 %) Male = 16 (41.0 %)
Âge	$M = 27.79$ ($SD = 7.96$), range = 20-54
Niveau d'études	Students = 16 (41.0%) Executives/Managers = 14 (35.9%) Employees = 6 (15.4%) Retired = 1 (2.6%) Entrepreneur = 1 (2.6%) Job seeker = 1 (2.6%).
Aisance avec les ordinateurs	$M = 7.91$ ($SD = 2.11$)
Aisance avec le JV	$M = 6.67$ ($SD = 2.69$)

Each participant signed an informed consent form that outlined their rights (e.g., the ability to withdraw at any time, access to their data) as well as the researchers' responsibilities (e.g., ensuring anonymity, managing access to sensitive data, secure data storage).

2.4. Instruments and validation process

2.4.1. Active-Empathic Listening Module (Questionnaire)

The assessment of active listening in Yuzu was based on an approach inspired by the theoretical dimensions of the Active-Empathic Listening Scale (AELS) [68,71]. The original AELS distinguishes three fundamental components: sensing (the ability to detect and identify cues from the

interlocutor), processing (the ability to interpret and organize received information), and responding (the ability to provide appropriate verbal or nonverbal feedback).

In Yuzu, these dimensions served as the conceptual framework for the development of an original module to measure active listening. Items were specifically formulated for the game and integrated into a gamified format. Responses were collected on a five-point Likert-type scale. The data included both an overall active listening score and three sub-scores corresponding to the theoretical dimensions of active-empathic listening, which were compared with the original AELS questionnaire [68].

In the game, this module appears immediately after the player awakens from prolonged sleep: the station's physician administers a short questionnaire to verify the participant's condition and adaptation to the situation (Figure 2).

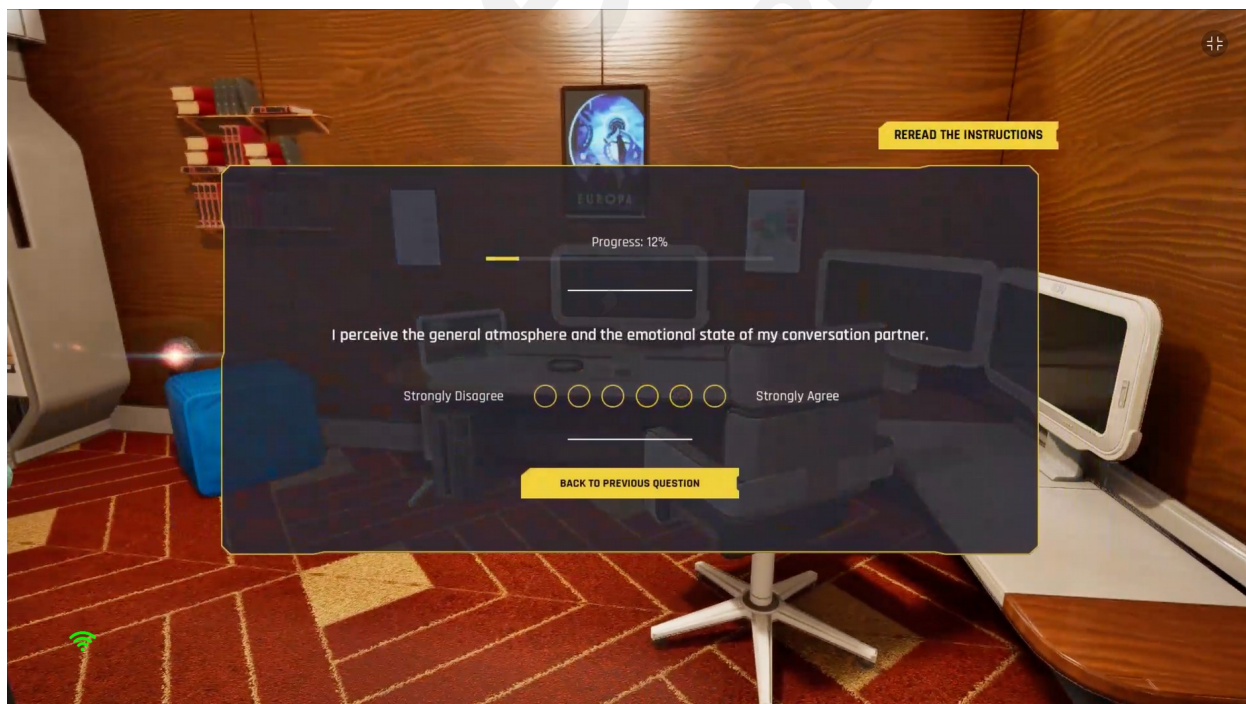


Figure 2. Screenshot of the Yuzu active listening module, inspired by the Active-Empathic Listening

Scale (AELS)

2.4.2. Decision-Making Module (Task)

Decision-making under uncertainty was assessed using an adapted version of the *Iowa Gambling Task* (IGT) [69]. In its classical form, the task consists of 100 trials in which the participant chooses among four card decks (A, B, C, D). Decks A and B are disadvantageous (high immediate gains but long-term losses), while decks C and D are advantageous (lower immediate gains but long-term benefits). Expected performance is reflected in an increasing preference for advantageous decks over the course of the task.

In Yuzu, the core principles of the IGT were retained but transposed into a serious game format, where participants had to collect meteorite samples via drones across four planetary excavation sites (Figure 3). The 100 trials were divided into 10 blocks of 10 choices. The first 40 trials constituted the exploration phase (blocks 1-4), while the remaining 60 trials formed the exploitation phase (blocks 5-10), consistent with the literature [72]. The primary performance indicator was the Net Score $((C + D) - (A + B))$ calculated during the exploitation phase. Secondary measures included the proportion of risky choices $(A + B)$ per block, which allowed for the analysis of learning trajectories.



Figure 3. Screenshot of the Yuzu decision-making module (gamified adaptation of the Iowa Gambling Task)

Scores from this module were compared with those obtained from an original Iowa Gambling Task administered online.

2.4.3. Teamwork module (dialogue choices)

Interpersonal skills were assessed through a narrative module inspired by the SYMLOG model [70]. The scenario places the player in an emergency situation: a spacecraft is approaching the space station dangerously, and rapid decisions must be made (Figure 4). Participants were confronted with four sets of six dialogue choices during interactions with non-player characters. Response options varied not only in content but also in tone (e.g., more or less dominant or friendly). In addition, some proposals originated from Earth, representing an external authority whose directives the player could either follow or reject.

Each option was pre-categorized according to the three axes of the model (i.e., dominance vs. submission, friendliness vs. hostility, and acceptance vs. rejection of norms and tasks). For each

player, average scores were then calculated on these dimensions, enabling the establishment of an in-game interpersonal behavioral profile. This profile was subsequently compared with that obtained through the simplified SYMLOG questionnaire by Blumberg (2006), administered separately.

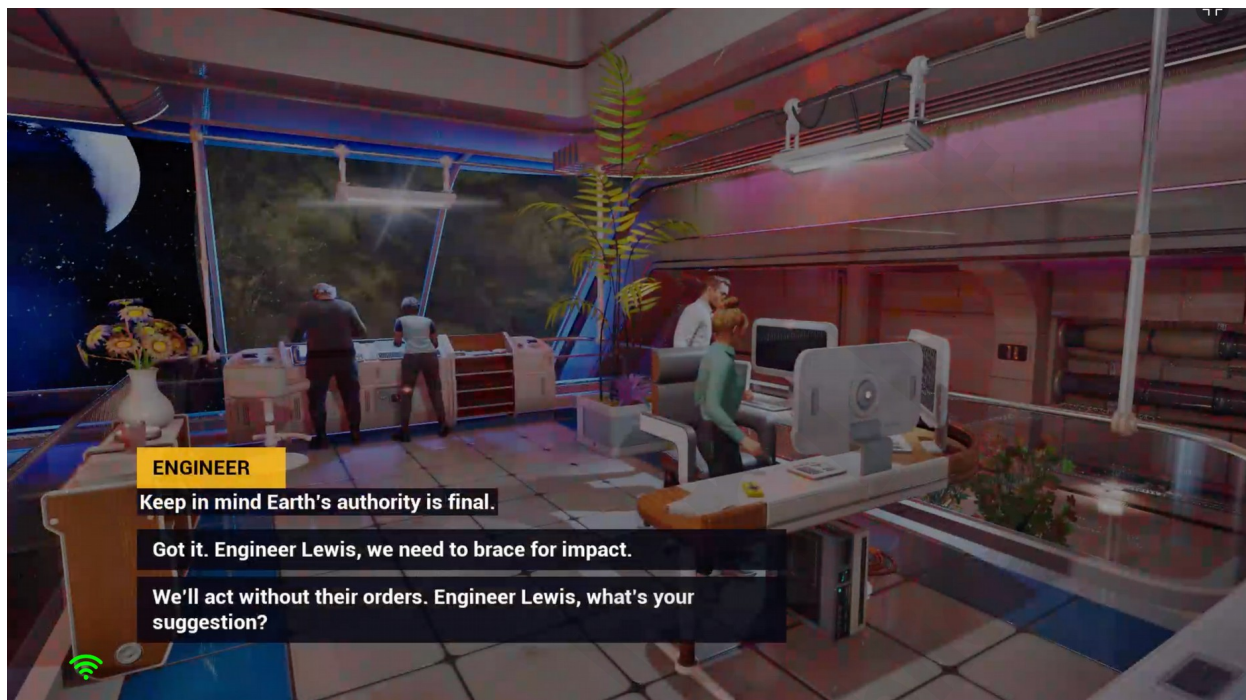


Figure 4. Screenshot of the introduction to the Yuzu teamwork module (gamified adaptation of the SYMLOG model)

2.5. Experimental procedure

Participants were welcomed in controlled settings (universities, laboratories), specifically in quiet rooms without distractions to ensure standardized testing conditions. Each session lasted approximately 2 hours, including a 10-minute break. This duration allowed participants to complete all tasks without inducing excessive fatigue.

To control for potential order effects, the administration was counterbalanced: half of the participants began with the reference version and then completed the gamified version, while the other half followed the reverse order (game first, then the paper-based or online reference tests). In addition,

measures of computer literacy and technophilia were collected to verify that performance was not biased by participants' relationship with digital technology.

The procedure was conducted as follows:

1. Welcome and briefing: Explanation of the study objectives and session outline (10 minutes).
2. Yuzu tasks and reference tests: Half of the participants began with Yuzu, followed by the reference tests; the other half started with the reference tests, followed by Yuzu (60-90 minutes).
3. Debriefing and final questionnaires: Debriefing interview, game experience questionnaire, and demographic questionnaire (20 minutes).

2.6. Statistical analyses

2.6.1. Active-empathic listening module

Agreement between the reference version of the AELS and the gamified version inspired by the AELS was examined using three approaches:

- Pearson correlation matrices were computed between the tests. Significant correlations ($p \leq .05$) were interpreted as evidence of good internal validity.
- A Bland-Altman plot [74] was used to assess absolute agreement between the two versions. This method estimates the mean bias (average difference between the two measures) and calculates the 95% limits of agreement ($\text{bias} \pm 1.96 \times \text{SD of the differences}$), indicating the interval in which most individual differences fall. The absence of systematic bias and narrow limits of agreement were interpreted as evidence of good agreement between methods. Analyses were performed using the *blandr* module in Jamovi (version 2.6.26.0).

- Lin's Concordance Correlation Coefficient (CCC) was calculated to quantify relative agreement, that is, the extent to which the two methods provide consistent values in terms of both location and dispersion. The CCC combines information on precision (Pearson correlation) and accuracy (proximity to the line of perfect equality). Interpretation followed standard thresholds: < 0.50 = poor, $0.50-0.75$ = moderate, $0.75-0.90$ = good, ≥ 0.90 = excellent. Analyses were performed using the *SimplyAgree* module [75] in Jamovi (version 2.6.26.0).

2.6.2. Decision-making module

Data from the Iowa Gambling Task (IGT) were divided into 10 blocks of 10 trials. Following the literature [72], the exploration phase was defined as the first 40 trials (blocks 1-4), and the exploitation phase as the subsequent 60 trials (blocks 5-10).

The primary endpoint was the mean Net Score calculated during the exploitation phase, defined as $(C + D) - (A + B)$. Descriptive statistics were also computed for the proportion of risky choices $((A + B)/10)$ per block.

Equivalence between the reference version and the Yuzu version was tested using the Two One-Sided Tests (TOST) procedure [76], with an equivalence margin set a priori at ± 5 points for the Net Score. This margin was chosen for its practical relevance and because it corresponds to a small effect size. Analyses were conducted in Jamovi (version 2.6.26.0) with $\alpha = .05$. Reported results included the two TOST p-values, the 90% CI of the difference, the 95% CI for transparency, and effect size (Hedges' g).

In addition, block-wise descriptive analyses (10 trials each) were performed to visualize the evolution of the Net Score throughout the task. A learning curve (mean $\pm$ 95% CI) was plotted separately for each condition to examine trajectories and identify potential early or late divergences.

2.6.3. Teamwork module

For the dialogue-choice assessment, each response selected in Yuzu was categorized according to the three SYMLOG dimensions [70]: dominance vs. submission, positivity vs. negativity, and task orientation vs. task rejection. For each participant, a mean score was calculated on each dimension.

These scores were then correlated, using Spearman's rho, with those obtained from the simplified SYMLOG questionnaire [73], administered separately. The presence of significant positive correlations would have indicated coherence between the two assessment modalities.

3. Results

3.1. Active-empathic listening

Results showed strong positive correlations between Yuzu's active listening module and the AELS questionnaire ($r = .743$, $p < .001$; Table 3).

Table 3. Spearman correlation matrix between Yuzu's active listening module and the reference AELS

	Active Listening Yuzu	AELS test
Active Listening Yuzu	Spearman's rho	—
	p -value	—
AELS test	Spearman's rho	0.743***
	p -value	< .001

The Bland-Altman analysis (Table 4, Figure 5) indicated a non-significant mean bias between the reference and gamified versions (Mean Bias = 0.0238; 95% CI [-0.0306, 0.0783]), suggesting the absence of systematic differences. The 95% limits of agreement ranged from -0.306 to 0.353 points, representing a maximum variation below 7% of the total scale.

Table 4. Bland-Altman analysis between Yuzu's active listening module and the reference AELS

	Estimate	95% Confidence Interval	
		Lower	Upper
Bias (n = 39)	0.0238	-0.0306	0.0783
Lower limit of agreement	-0.3056	-0.3995	-0.2117
Upper limit of agreement	0.3533	0.2594	0.4472

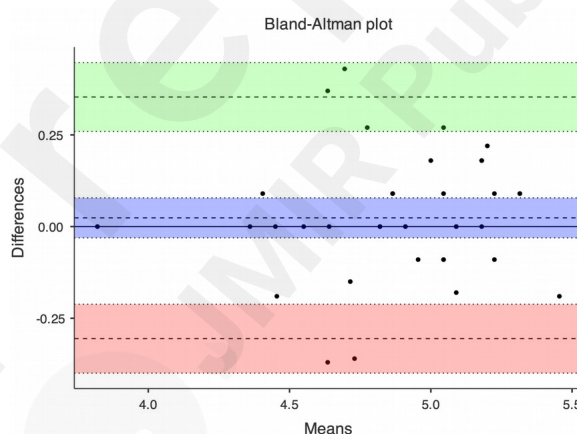


Figure 5. Bland-Altman plot between Yuzu's active listening module and the reference AELS

Lin's concordance correlation coefficient (CCC) between the two versions was 0.899 (95% CI [0.817, 0.946]), indicating good to excellent agreement according to standard thresholds (Table 5). These results suggest that the two versions produce highly similar total scores, both in absolute values and in relative participant rankings.

Table 5. CCC between Yuzu's active listening module and the reference AELS

	Estimate	Lower C.I.	Upper C.I
CCC	0.899	0.817	0.946

Overall, these findings validate the relationship between the two versions and empirically support the use of a gamified tool to assess active listening.

3.2. Decision-making module

For the decision-making module, analyses supported equivalence between the two modalities. During the exploitation phase, the mean Net Score was 3.57 (SD = 4.85) for the reference version and 4.93 (SD = 4.61) for the Yuzu version.

The mean difference (Yuzu vs. reference test) was 1.35, with a 90% CI of [-1.22, 3.93], entirely within the equivalence interval [-5, +5]. The TOST results (Table 6) showed that both one-sided tests were significant (lower test: $p < .001$; upper test: $p = .012$), confirming statistical equivalence. The effect size was small (Hedges' $g = 0.21$, 90% CI [-0.31, 0.75]). A traditional paired t -test revealed no significant difference ($p = .374$), with a 95% CI of [-1.77, 4.48].

Table 6. TOST and t -test results comparing Net Scores on the Iowa Gambling Task between the reference version and Yuzu

Statistic	Estimate	df	p	90% Lower	CI Upper	90% CI	Equivalence Bounds (Low, High)
t-test (Net Score)	0.913	17	0.374	–	–	–	–
TOST	3.615	17	0.001	–	–	–	–

Lower						
TOST	-1.789	17	0.046	—	—	—
Upper						
Effect size	1.352	—	—	-1.224	3.927	-4.000, 4.000
(Raw)						
Effect size	0.210	—	—	-0.311	0.752	-0.637, 0.637
(Hedges's						
g(z))						

Figure 6 illustrates the distribution of the difference and the position of the confidence intervals relative to the equivalence bounds.

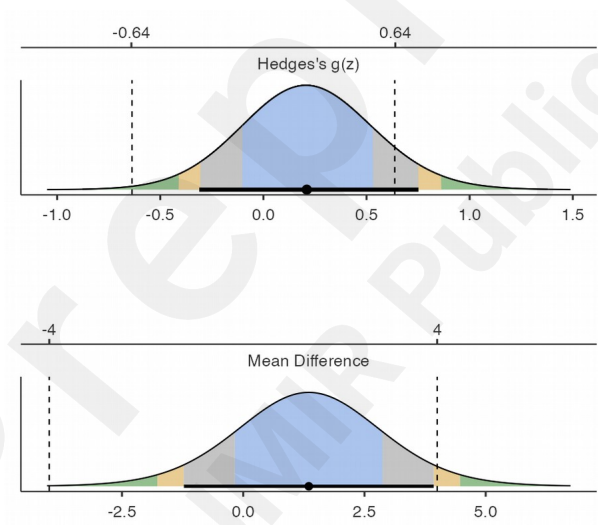


Figure 6. TOST graph (Hedges' *g* with bounds and Mean Difference)

Figure 7 illustrates the evolution of the mean Net Score across blocks of 10 trials for both modalities. In both conditions, performance progressively increased, marking the shift from exploration to exploitation starting at block 5. The trajectories were visually parallel, suggesting comparable learning dynamics across modalities.

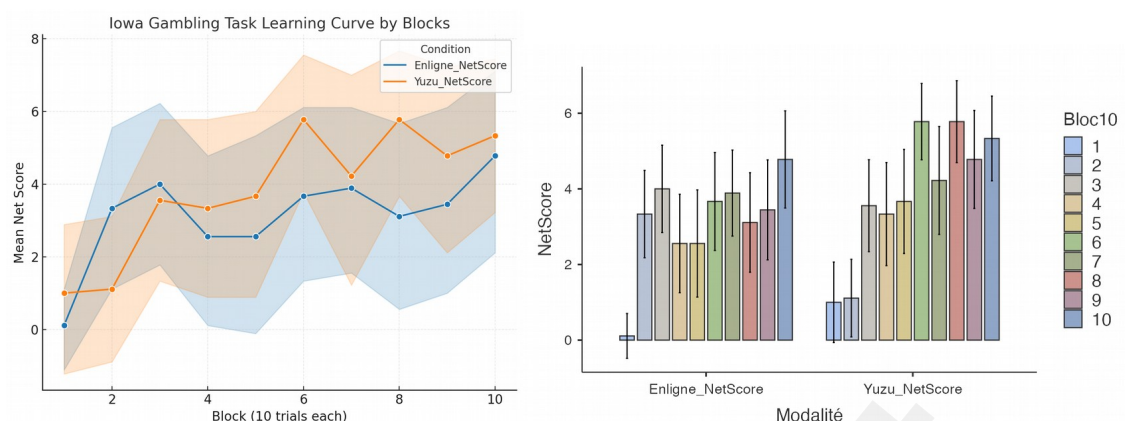


Figure 7. Evolution of mean Net Score in the Iowa Gambling Task by blocks of 10 trials for each condition (reference test vs. Yuzu)

3.3. Teamwork module

In contrast to the two previous modalities, the dialogue-choice assessment did not show convergent validity with the SYMLOG questionnaire.

Descriptive statistics indicated that nearly all participants adopted a collaborative teamwork style (high affection, high task orientation, high dominance), suggesting a ceiling effect. Spearman correlations between Yuzu's dimensions (dominance, positivity, task orientation) and those from the questionnaire were all non-significant (Table 7).

Table 7. Spearman correlation matrix between scores obtained from Yuzu's teamwork module and the SYMLOG questionnaire

Dimension	Variable	Spearman's rho	p-value
Dominance	Reference vs. Yuzu	-0.070	0.656
Positivity	Reference vs. Yuzu	0.082	0.319
Task	Reference vs. Yuzu	0.134	0.222

Descriptive statistics (Table 8) showed that SYMLOG questionnaire scores were generally balanced around the mean, with a wide distribution across the three axes, suggesting that the questionnaire differentiated individual profiles effectively. By contrast, Yuzu-derived scores displayed systematic biases: participants appeared more oriented toward friendliness (median = +0.56) and task orientation (median = +0.42), while dominance was slightly negative (median = -0.23). This asymmetry likely reflects a generalized preference for collaborative options in the game, confirming the ceiling effect observed in the profile distribution.

Table 8. Descriptive statistics of SYMLOG scores (questionnaire and Yuzu version, z-scores)

	Median	Minimum	Maximum
Paper_dominance – Z Score	-0.0778	-2.01	2.128
Yuzu_dominance – Z Score	-0.2395	-1.70	1.948
Paper_positivity – Z Score	-0.1340	-2.22	1.608
Yuzu_positivity – Z Score	0.5657	-2.45	0.996
Paper_task – Z Score	-0.1764	-2.71	2.359
Yuzu_task – Z Score	0.4219	-2.29	1.025

4. Discussion

Building on the pilot study conducted by Dinét et al. [67], our research extends serious game to additional modalities (gamified questionnaire, decision-making task, interactive dialogues).

For the active listening module, analyses showed a high level of agreement between the reference and gamified versions inspired by the AELS, thereby supporting hypothesis H1. This suggests that the gamified adaptation preserves the construct validity of the original tool and that active listening can be assessed through a serious game without major alteration [19]. This finding is particularly

noteworthy given the recurring criticisms regarding the subjectivity and social desirability biases of self-report questionnaires [43].

Regarding the decision-making module, measured with the IGT, results supported hypothesis H2. The gamified adaptation reproduced the decision-making dynamics under uncertainty previously demonstrated by the original version [69], supporting the relevance of this format for applied assessment contexts.

In contrast to questionnaires and the IGT, the assessment of soft skills through dialogue choices did not show significant correlations with the SYMLOG. Several interpretations may be advanced. First, the complexity of human interactions is difficult to reduce to a limited set of response options: the choices offered in Yuzu may have lacked sufficient nuance to faithfully reflect the dimensions of the circumplex model (dominance, positivity, task orientation). Second, a ceiling effect was observed: most participants selected collaborative responses, likely due to social desirability bias. Descriptive analyses confirmed this bias, showing that Yuzu scores converged toward a cooperative and task-oriented profile, whereas the reference questionnaire produced a wider distribution. In a testing context, it is indeed valued to appear cooperative. Finally, the SYMLOG format (self-report questionnaire) and the serious game format (contextualized interactive choices) may not be measuring exactly the same constructs, which could explain the divergence in results. This observation aligns with limitations already noted in the literature on scenario-based serious games (Altomari et al., 2023): while these approaches can operationalize certain behavioral dimensions, they also risk oversimplifying the complexity of human interactions [5,12].

Furthermore, Hernandez et al. [57] demonstrated that it is possible to analyze behavioral tendencies through narrative choices in a serious game, while also highlighting the limits of such approaches. Similarly, our findings suggest that the validity of dialogue-choice assessments could be strengthened by offering a greater variety of response options or by incorporating AI-based adaptive

dialogues that adjust character responses in real time.

More broadly, these results extend the conclusions of Leutner et al. [51]: gamification provides promising perspectives in terms of user experience and acceptability, but its psychometric foundations need to be further consolidated.

5. Limitation and suggestions for future research

Several limitations should be acknowledged. First, the relatively small sample size limits the statistical power of the analyses. It is therefore possible that effects of small magnitude were not detected. A promising avenue for future research would be to increase sample sizes in subsequent empirical studies.

Another point concerns the duration of the experimental sessions, which was relatively long in this study. Excessive session length may generate fatigue, reduce participant concentration, and affect engagement. This can particularly influence the quality of responses to questionnaires or the consistency of performance in game-based tasks. Future studies could explore shorter and more modular formats, for example by splitting the session into multiple parts or by offering abbreviated versions of the tests.

Finally, the limitations observed in the dialogue-choice assessment call for methodological improvements. The low level of correlation with the SYMLOG questionnaire may be explained by a lack of nuance in the response options, a social desirability bias (participants' tendency to choose the option perceived as the most socially acceptable), or the fact that in-game dialogues and questionnaires may capture only partially overlapping facets of interpersonal communication. Future studies could diversify scenarios, introduce more contrasting dilemmas, analyze not only the final choice but also response times and hesitation, and use AI-driven non-player characters to increase the

credibility and sensitivity of the assessment.

6. Conclusion

The results of this study provide several insights for the future development of serious game for soft skills assessment. First, they show that the gamification of standardized psychometric tests (AELS, IGT) is a promising approach: the versions integrated into Yuzu demonstrated validity comparable to that of the reference versions, while enhancing participant engagement. Second, the partial failure of the dialogue-based module highlights the specific challenges inherent to this type of assessment. Dialogue-choice tasks are intuitive and engaging, but their validity strongly depends on the richness of scenarios, the diversity of options, and the management of biases (e.g., social desirability, confirmation).

These findings underscore the importance of subjecting gamified tools to the same validation requirements as traditional psychometric instruments. To be credible in both academic and organizational contexts, they must be empirically validated. Gamification should therefore not be viewed as an end in itself, but as a means of collecting ecologically valid behavioral data, provided that methodological robustness is established, and as a complement to other assessment methods.

Yuzu thus positions itself as an innovative contribution to the literature on non-technical skills assessment, directly addressing calls in the literature for contextualized and scientifically validated tools. In the longer term, the integration of real-time feedback or AI-generated adaptive dialogues could further enhance measurement sensitivity. These avenues will be explored in future studies.

Funding statement

This research received no specific grant.

Conflict of interest

None declared.

References

- [1] Deming DJ. The value of soft skills in the labor market. NBER Reporter. 2017;(4):7-11.
- [2] Schislyaeva ER, Saychenko OA. Labor market soft skills in the context of digitalization of the economy. Social Sciences. 2022;11(3):91. doi:10.3390/socsci11030091
- [3] Poláková M, Suleimanová JH, Madzík P, et al. Soft skills and their importance in the labour market under the conditions of Industry 5.0. Heliyon. 2023;9(8):e18670. doi:10.1016/j.heliyon.2023.e18670
- [4] Espina-Romero LC, Aguirre Franco SL, Dworaczek Conde HO, et al. Soft skills in personnel training: Report of publications in Scopus, topics explored and future research agenda. Heliyon. 2023;9(4):e15468. doi:10.1016/j.heliyon.2023.e15468
- [5] Altomari L, Altomari N, Iazzolino G. Gamification and soft skills assessment in the development of a serious game: Design and feasibility pilot study. JMIR Serious Games. 2023;11:e45436. doi:10.2196/45436
- [6] Matteson ML, Anderson L, Boyden C. "Soft skills": A phrase in search of meaning. Portal: Libraries and the Academy. 2016;16(1):71-88. doi:10.1353/pla.2016.0009
- [7] Carvalho MMD, Rabechini Junior R. Impact of risk management on project performance: The importance of soft skills. International Journal of Production Research.

- 2015;53(2):321-340. doi:10.1080/00207543.2014.919423
- [8] Caudron S. The hard case for soft skills. *Workforce*. 1999;789:60-67.
- [9] Touloumakos AK. Expanded yet restricted: A mini review of the soft skills literature. *Frontiers in Psychology*. 2020;11:2207. doi:10.3389/fpsyg.2020.02207
- [10] Hurrell SA, Scholarios D, Thompson P. More than a “humpty dumpty” term: Strengthening the conceptualization of soft skills. *Economic and Industrial Democracy*. 2013;34(1):161-182. doi:10.1177/0143831X12444934
- [11] Marin-Zapata SI, Román-Calderón JP, Robledo-Ardila C, Jaramillo-Serna MA. Soft skills, do we know what we are talking about? *Review of Managerial Science*. 2022;16(4):969-1000. doi:10.1007/s11846-021-00474-9
- [12] Mitchell GW, Pritchett CC, Skinner LB. The importance of the integration of soft skills into the curriculum as identified by MBA students. *Academy of Business Research Journal*. 2013;I:8.
- [13] Parente DH, Stephan JD, Brown RC. Facilitating the acquisition of strategic skills: The role of traditional and soft managerial skills. *Management Research Review*. 2012;35(11):1004-1028. doi:10.1108/01409171211276918
- [14] Laurisz N, Gáspár T, Gałat W, Juhász T. The other side of the coin: Expectations of Polish and Hungarian students on soft skills in the labour market—a futures perspective. *European Journal of Futures Research*. 2024;12(1):13. doi:10.1186/s40309-024-00235-3
- [15] Nasir ANM, Ali DF, Noordin MKB, Nordin MSB. Technical skills and non-technical skills: Predefinition concept. In: *Proceedings of the IETEC’11 Conference*; 2011:01-p17; Kuala Lumpur, Malaysia.
- [16] Lucas B, Venckutė M. Creativity—a transversal skill for lifelong learning: An overview of existing concepts and practices. Literature review report. 2020. EUR 30405.
- [17] Lamri J, Barabel M, Lubart T, Meier O. Chapitre 1. Généralités sur les soft skills. In: *Le*

défi des soft skills: Comment les développer au XXI^e siècle ? Paris: Dunod; 2022:23-42.

Available at: <https://shs.cairn.info/le-defi-des-soft-skills--9782100830923-page-23?lang=fr>

- [18] Haselberger D, Oberheumer P, Perez E, Cinque M, Capasso D. Mediating soft skills at higher education institutions. Handbook of ModEs Project. Education and Culture DG Lifelong Learning Programme. European Union; 2012. Brussels.
- [19] Succi C, Wieandt M. Walk the talk: Soft skills' assessment of graduates. European Journal of Management and Business Economics. 2019;28(2):114-125. doi:10.1108/EJMBE-01-2019-0011
- [20] Andrés JC, Aguado D, García-Izquierdo AL. Big Four LinkedIn dimensions: Signals of soft skills? Revista de Psicología del Trabajo y de las Organizaciones. 2023;39(2):75-88.
- [21] Werner K, Junek O, Wang C. Event management skills in the post-COVID-19 world: Insights from China, Germany, and Australia. Event Management. 2022;26(4):867-882.
- [22] Autor DH. Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives. 2015;29(3):3-30.
- [23] Azim S, Gale A, Lawlor-Wright T, Kirkham R, Khan A, Alam M. The importance of soft skills in complex projects. International Journal of Managing Projects in Business. 2010;3(3):387-401. doi:10.1108/17538371011056048
- [24] Burbekova S. Soft skills as the most in-demand skills of future IT specialists. In: 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST). 2021:1-5. doi:10.1109/SIST50301.2021.9465935
- [25] Wingate TG, Bourdage JS. What are interviews for? A qualitative study of employment interview goals and design. Human Resource Management. 2024;63(4):555-580. doi:10.1002/hrm.22215
- [26] Bergelson I, Tracy C, Takacs E. Best practices for reducing bias in the interview process. Current Urology Reports. 2022;23(11):319-325. doi:10.1007/s11934-022-01116-7

- [27] Dana J, Dawes R, Peterson N. Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making*. 2013;8(5):512-520.
- [28] Räsänen J, Lippert-Rasmussen K. Personality discrimination and the wrongness of hiring based on extraversion. *Journal of Business Ethics*. 2024;195(3):681-694. doi:10.1007/s10551-024-05643-w
- [29] Ziegler M. Applicant faking: A look into the black box. *The Industrial and Organizational Psychologist*. 2011;49(1):29-36.
- [30] Aamodt MG, Custer H. Who can best catch a liar? *Forensic Examiner*. 2006;15(1):6.
- [31] Galos DR, Coppock A. Gender composition predicts gender bias: A meta-reanalysis of hiring discrimination audit experiments. *Science Advances*. 2023;9(18):eade7979. doi:10.1126/sciadv.ade7979
- [32] Batinovic L, Howe M, Sinclair S, Carlsson R. Ageism in hiring: A systematic review and meta-analysis of age discrimination. *Collabra: Psychology*. 2023;9(1):82194. doi:10.1525/collabra.82194
- [33] Johnson SK, Podratz KE, Dipboye RL, Gibbons E. Physical attractiveness biases in ratings of employment suitability: Tracking down the “beauty is beastly” effect. *The Journal of Social Psychology*. 2010;150(3):301-318. doi:10.1080/00224540903365414
- [34] Lal N, Benkraouda O. Exploring the implementation of AI in early onset interviews to help mitigate bias. *arXiv preprint*. 2025:arXiv:2501.09890.
- [35] Hunkenschroer AL, Kriebitz A. Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring. *AI and Ethics*. 2023;3(1):199-213. doi:10.1007/s43681-022-00166-4
- [36] O'Donnell AM, Topping K. Peers assessing peers: Possibilities and problems. In: Topping K, Ehly S, eds. *Peer-Assisted Learning*. London: Routledge; 1998:255-278.
- [37] Cho K, Schunn CD. Scaffolded writing and rewriting in the discipline: A web-based

- reciprocal peer review system. *Computers & Education*. 2007;48(3):409-426. doi:10.1016/j.compedu.2005.02.004
- [38] Lertsakulbunlue S, Kantiwong A. Evaluating the dependability of peer assessment in project-based learning for pre-clinical students: A generalizability theory approach. *BMC Medical Education*. 2025;25(1):260. doi:10.1186/s12909-025-06772-0
- [39] Tandiono R, Limijaya A. Understanding rater bias in self and peer assessment among university students: A cultural psychology perspective. *The Asia-Pacific Education Researcher*. 2025;34(4):1561-1571. doi:10.1007/s40299-024-00967-7
- [40] Aridi M, Ahmad M, Ghach W, Charifeh S, Alwan N. Development and validation of the soft skills questionnaire for nurses. *SAGE Open Nursing*. 2023;9:23779608231159620. doi:10.1177/23779608231159620
- [41] Escolà-Gascón Á, Gallifa J. How to measure soft skills in the educational context: Psychometric properties of the SKILLS-in-ONE questionnaire. *Studies in Educational Evaluation*. 2022;74:101155. doi:10.1016/j.stueduc.2022.101155
- [42] Jardim J, Pereira A, Vagos P, Direito I, Galinha S. The Soft Skills Inventory: Developmental procedures and psychometric analysis. *Psychological Reports*. 2022;125(1):620-648. doi:10.1177/0033294120979933
- [43] Paulhus DL, Vazire S. The self-report method. In: *Handbook of Research Methods in Personality Psychology*. 2007:224-239.
- [44] Lucas RE, Baird BM. Global self-assessment. In: Eid M, Diener E, eds. *Handbook of Multimethod Measurement in Psychology*. Washington, DC: American Psychological Association; 2006:29-42. doi:10.1037/11383-003
- [45] Paulhus DL, Reid DB. Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*. 1991;60(2):307.
- [46] Raven J. The Raven's Progressive Matrices: Change and stability over culture and time.

- Cognitive Psychology. 2000;41(1):1-48. doi:10.1006/cogp.1999.0735
- [47] Kuncel NR, Ones DS, Sackett PR. Individual differences as predictors of work, educational, and broad life outcomes. *Personality and Individual Differences*. 2010;49(4):331-336. doi:10.1016/j.paid.2010.03.042
- [48] Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*. 1998;124(2):262.
- [49] Schmidt FL, Hunter J. General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*. 2004;86(1):162-173. doi:10.1037/0022-3514.86.1.162
- [50] Hausknecht JP, Day DV, Thomas SC. Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*. 2004;57(3):639-683.
- [51] Leutner F, Codreanu S-C, Brink S, Bitsakis T. Game based assessments of cognitive ability in recruitment: Validity, fairness and test-taking experience. *Frontiers in Psychology*. 2023;13:942662. doi:10.3389/fpsyg.2022.942662
- [52] Hunter JE, Hunter RF. Validity and utility of alternative predictors of job performance. *Psychological Bulletin*. 1984;96(1):72.
- [53] Ployhart RE, Holtz BC. The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*. 2008;61(1):153-172. doi:10.1111/j.1744-6570.2008.00109.x
- [54] Smither JW, Reilly RR, Millsap RE, Stoffey RW. Applicant reactions to selection procedures. *Personnel Psychology*. 1993;46(1):49-76.
- [55] Pedersen MK, Díaz CMC, Alba-Marrugo MA, et al. Cognitive abilities in the wild: Population-scale game-based cognitive assessment. 2020.
- [56] Alloza S, Escribano F, Delgado S, Corneanu C, Escalera S. XBadges: Identifying and

- training soft skills with commercial video games. arXiv preprint arXiv:1707.00863. 2017.
- [57] Hernandez J, Muratet M, Pierotti M, Carron T. How to model a visual novel game to train and identify players' soft skills. In: European Conference on Game Based Learning. 2021:322-330.
- [58] Pouezevara S, Powers S, Moore G, Strigel C, McKnight K. Assessing soft skills in youth through digital games. In: ICERI2019 Proceedings. 2019:3057-3066. doi:10.21125/iceri.2019.0774
- [59] Marengo A, Pagano A, Soomro KA. Serious games to assess university students' soft skills: Investigating the effectiveness of a gamified assessment prototype. Interactive Learning Environments. 2024;32(10):6142-6158. doi:10.1080/10494820.2023.2253849
- [60] Marengo A, Pagano A, Lund B, Santamato V. Research AI: Integrating AI and gamification in higher education for e-learning optimization and soft skills assessment through a cross-study synthesis. Frontiers in Computer Science. 2025;7:1587040. doi:10.3389/fcomp.2025.1587040
- [61] Jennett C, Cox AL, Cairns P, et al. Measuring and defining the experience of immersion in games. International Journal of Human-Computer Studies. 2008;66(9):641-661. doi:10.1016/j.ijhcs.2008.04.004
- [62] Weibel D, Wissmath B, Habegger S, Steiner Y, Groner R. Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment. Computers in Human Behavior. 2008;24(5):2274-2291. doi:10.1016/j.chb.2007.11.002
- [63] McPherson J, Burns NR. Assessing the validity of computer-game-like tests of processing speed and working memory. Behavior Research Methods. 2008;40(4):969-981. doi:10.3758/BRM.40.4.969
- [64] Georgiou K, Nikolaou I. Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. Computers in

- Human Behavior. 2020;109:106356. doi:10.1016/j.chb.2020.106356
- [65] Mavridis A, Tsiatsos T. Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*. 2017;33(2):137-150. doi:10.1111/jcal.12170
- [66] Albuquerque J, Bittencourt II, Coelho JAPM, Silva AP. Does gender stereotype threat in gamified educational environments cause anxiety? An experimental study. *Computers & Education*. 2017;115:161-170. doi:10.1016/j.compedu.2017.08.005
- [67] Dinet J, Kitajima M, Fichet L, Paquet C, Coursac V. A gamified sorting test to assess cognitive flexibility in personnel selection: A pilot study. In: *COGNITIVE 2023: The Fifteenth International Conference on Advanced Cognitive Technologies and Applications*. 2023:2-5.
- [68] Bodie GD. The Active-Empathic Listening Scale (AELS): Conceptualization and evidence of validity within the interpersonal domain. *Communication Quarterly*. 2011;59(3):277-295. doi:10.1080/01463373.2011.583495
- [69] Buelow MT, Suhr JA. Construct validity of the Iowa Gambling Task. *Neuropsychology Review*. 2009;19(1):102-114. doi:10.1007/s11065-009-9083-4
- [70] Bales RF, Couch AS. The value profile: A factor analytic study of value statements. *Sociological Inquiry*. 1969;39(1):3-17. doi:10.1111/j.1475-682X.1969.tb00934.x
- [71] Drollinger T, Comer LB, Warrington PT. Development and validation of the active empathetic listening scale. *Psychology & Marketing*. 2006;23(2):161-180.
- [72] Schmitz F, Kunina-Habenicht O, Hildebrandt A, Oberauer K, Wilhelm O. Psychometrics of the Iowa and Berlin Gambling Tasks: Unresolved issues with reliability and validity for risk taking. *Assessment*. 2020;27(2):232-245. doi:10.1177/1073191117750470
- [73] Blumberg HH. A simplified version of the SYMLOG trait rating form. *Psychological Reports*. 2006;99(1):46-50.

- [74] Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986;327(8476):307-310.
- [75] Caldwell AR. SimplyAgree: An R package and jamovi module for simplifying agreement and reliability analyses. *Journal of Open Source Software*. 2022;7(71):4148. doi:10.21105/joss.04148
- [76] Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. 1987;15(6):657-680. doi:10.1007/BF01068419

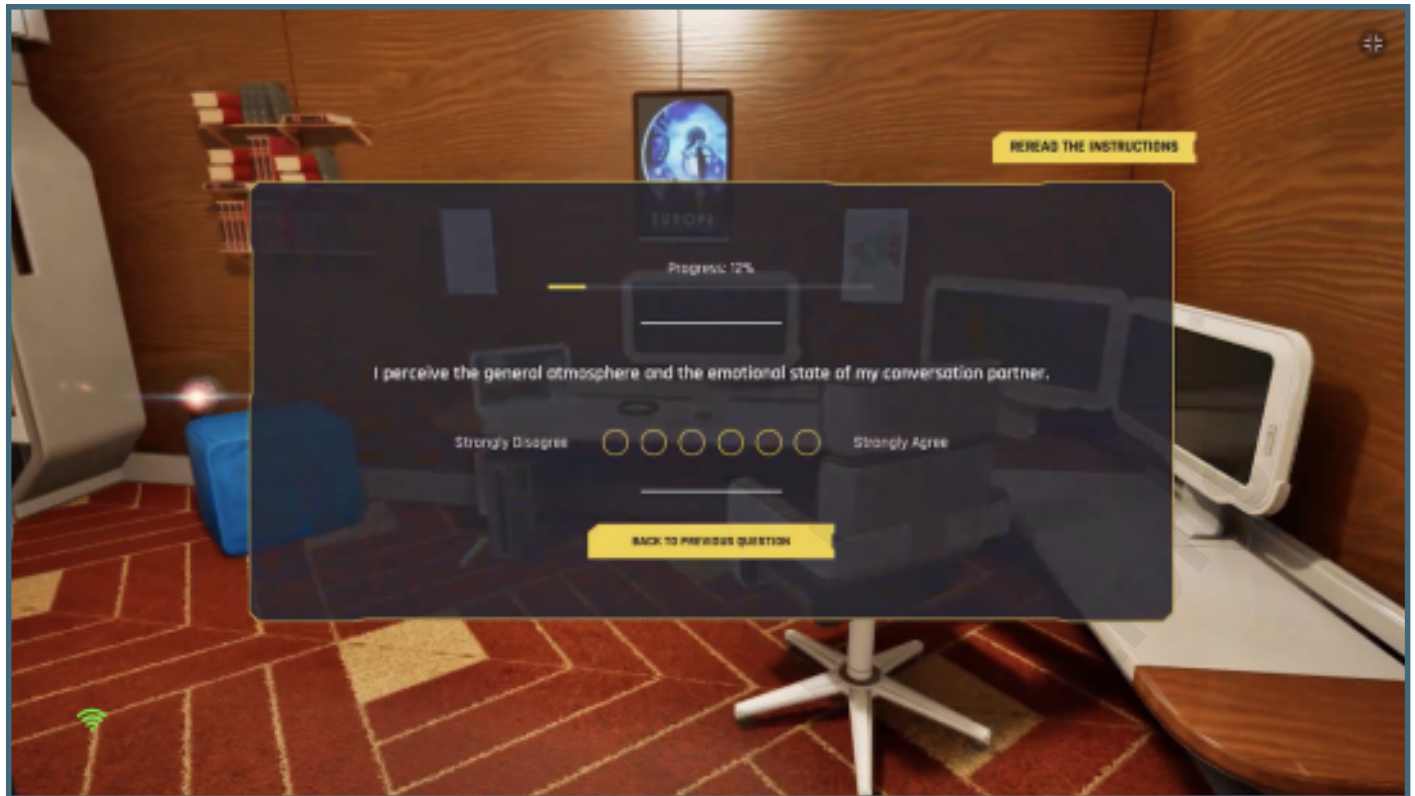
Supplementary Files

Figures

Actual images from the Yuzu assessment video game.



Screenshot of the Yuzu active listening module, inspired by the Active-Empathic Listening Scale (AELS).



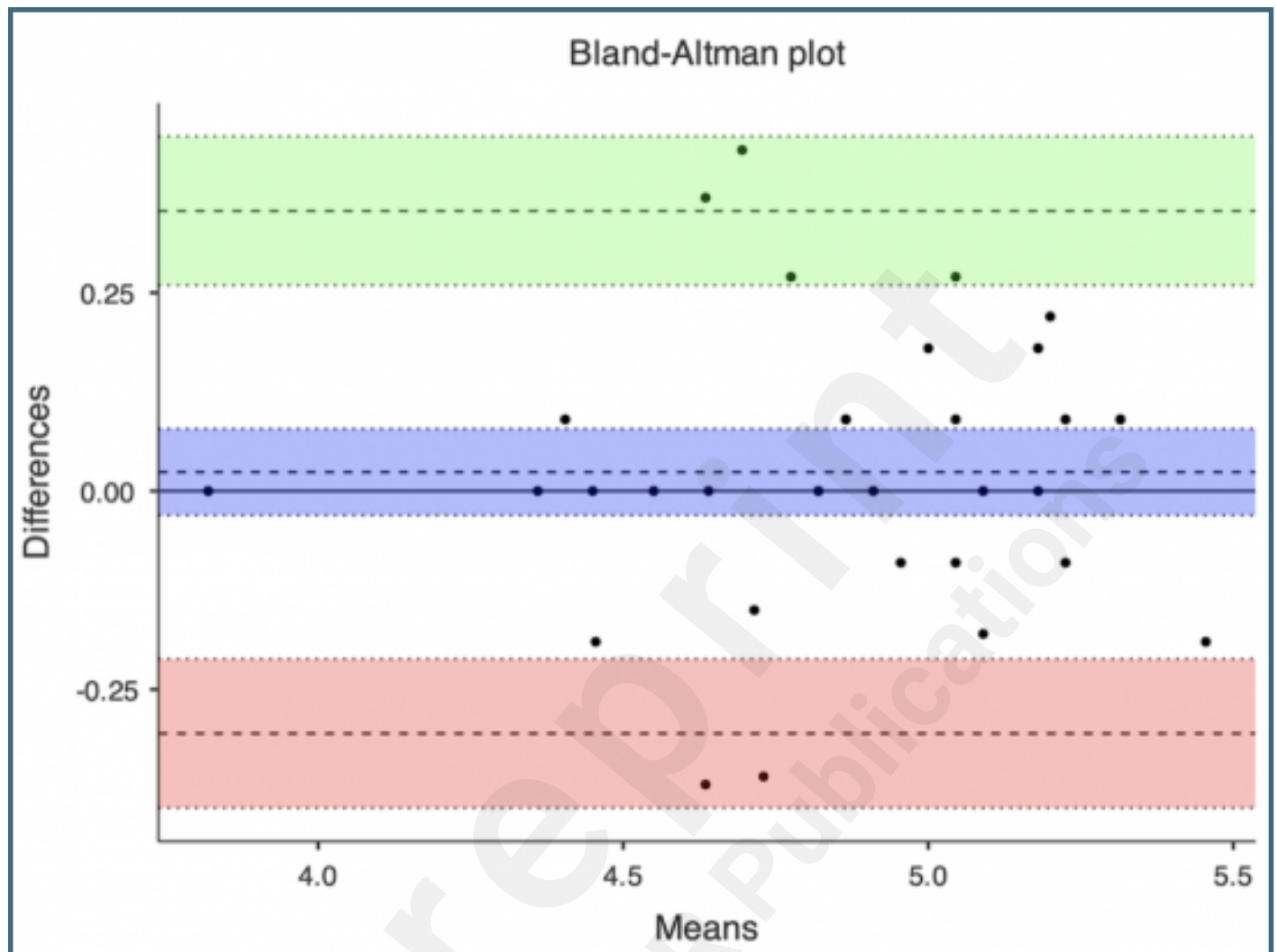
Screenshot of the Yuzu decision-making module (gamified adaptation of the Iowa Gambling Task).



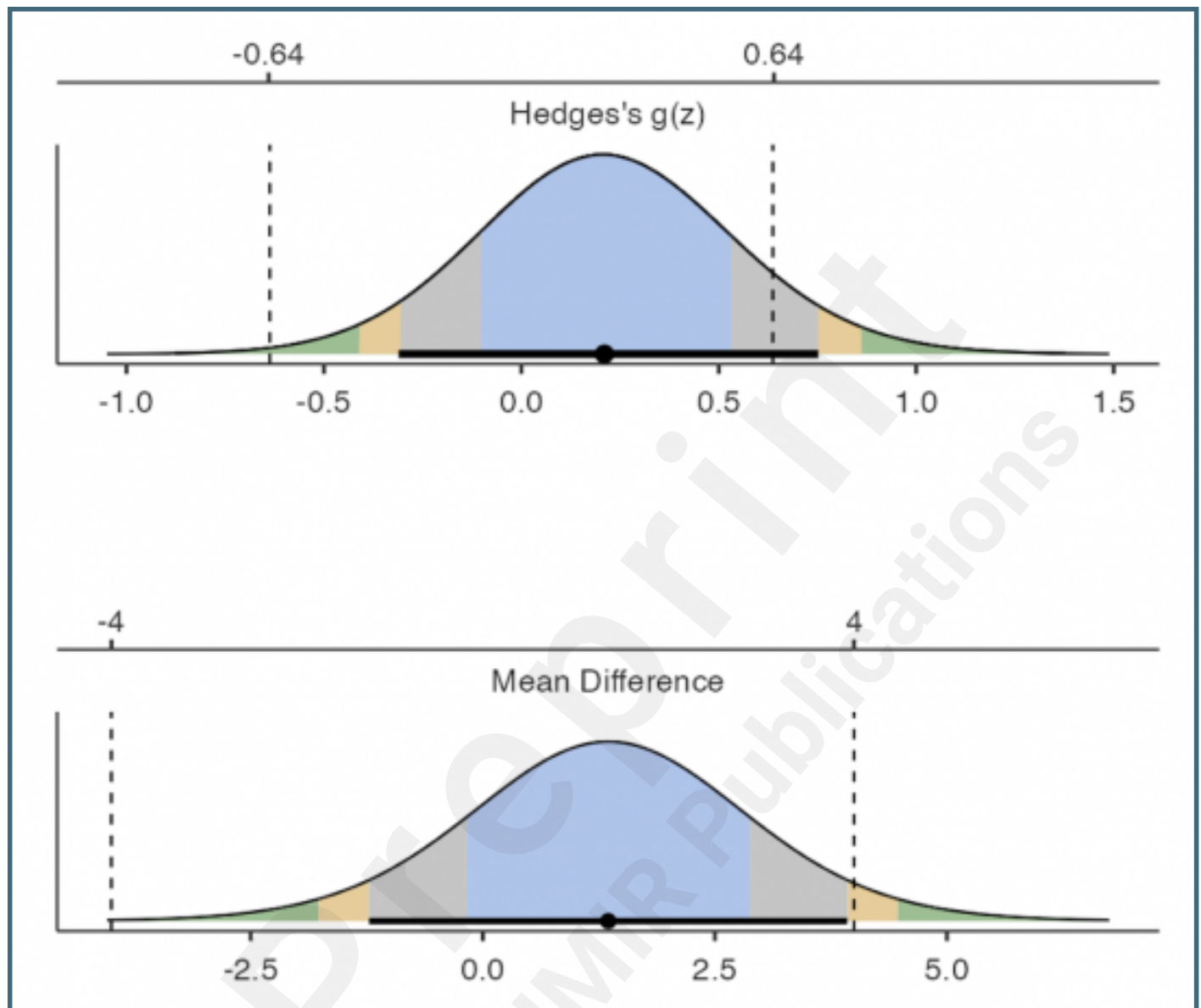
Screenshot of the introduction to the Yuzu teamwork module (gamified adaptation of the SYMLOG model).



Bland-Altman plot between Yuzu's active listening module and the reference AELS.



TOST graph (Hedges' g with bounds and Mean Difference).



Evolution of mean Net Score in the Iowa Gambling Task by blocks of 10 trials for each condition (reference test vs. Yuzu).

