UNIVERSITY
INSTITUTE
OF LISBON

Does the gamification process suggest better psychometric qualities than the traditional situational judgement tests (SJT)?

Gonçalo Nuno Pacheco Silva

Masters in Human Resources Management and Organizational Consulting

Supervisor:
PhD Aristides Ferreira, Associate Professor,
ISCTE Business School

September, 2023

BUSINESS
SCHOOL

Department of Human Resources and Organizational Behavior

Does the gamification process suggest better psychometric qualities than the traditional situational judgement tests (SJT)?

Gonçalo Nuno Pacheco Silva

Masters in Human Resources Management and Organizational Consulting

Supervisor:
PhD Aristides Ferreira, Associate Professor,
ISCTE Business School

September, 2023

Dedication and acknowledgements

A few years ago, I didn't imagine myself taking this path. My interest in this area awoke me in the last year of my degree, when I was introduced to Human Resources and Team Management, which I identified myself with in my daily life and among my group of friends without realizing it. I wanted to learn more and more about the area, so much so that over the following month, during my academic break, I completed several courses in this field and my desire to learn more was not satisfied.

My degree in IT and Business Management came from, on the one hand, my liking for games since I can remember, unleashing my curiosity in the IT side, mixed with my background in economics coming from high school, which I enjoyed. The degree allowed me to acquire IT and management fundamentals on how to optimize systems and business processes using technological systems. It is with great pleasure that in developing my thesis I was able to combine my passions to create something that I truly believe will have an impact and could be the future in the field.

I want to show gratitude to ISCTE for giving me this opportunity not only in my master's but also on my bachelors' degree. A great institution with great facilities and with high-quality experienced professors that have taught me a lot, not only school related, and I can proudly say that I have changed completely since my first year at this University.

Particularly I want to thank my supervisor Aristides Ferreira that has given me valuable advice throughout the whole journey, support, and encouraging words, being an exceptional mentor that gave me the guidelines and the support that I needed to successfully be completed this master thesis. I want to show my appreciation towards all my close friends and family for supporting me through this journey.

Finally, I want to thank Maria (my girlfriend) for being very patient, kind, and supportive, with constant encouraging words and love, that even in my struggles believed in me and lift me up, to keep working hard towards my objectives, and so deserves this paragraph and much more.

Resumo

A tese explora o uso de gamificação no recrutamento, mais especificamente quando a testar a criatividade do individuo. A gamificação ou avaliações gamificadas, é a adição de elementos semelhantes a jogos às avaliações existentes e permite aos utilizadores uma melhor experiência e desenvolvimento de sentimentos positivos. Quando aplicada ao recrutamento, molda a maneira como o candidato perceciona a empresa. Este estudo investiga se a gamificação sugere melhores qualidades psicométricas do que os tradicionais testes de julgamento situacional (SJT). Através de um teste de julgamento de situação tradicional desenvolvido por Beatriz Trigo, o investigador desenvolveu uma versão gamificada do SJT e compara as duas amostras e os seus resultados. A versão gamificada é constituída de uma amostra de 48 portugueses, que concluíram a gamificação e o questionário após o jogo para a obtenção de resultados e recodificação dos mesmos. Após, foi extraído um único fator que passou por diversos passos de validação e unidimensionalidade para a criação de um modelo criativo. De forma a julgar se a gamificação é o ponto de viragem do recrutamento e seleção tradicional.

Palavras-chave: Gamificação, teste de julgamento situacional, criatividade, recursos humanos, recrutamento e seleção

Classificação JEL: J24 (Capital Humano; Competências); O15 (Recursos Humanos)

Abstract

The thesis explores the use of gamification in recruitment, more specifically when testing an individual's creativity. Gamification, or gamified assessments, is the addition of game-like elements to existing assessments and allows users to have a better experience and develop positive feelings. When applied to recruitment, it shapes the way the candidate perceives the company. This study investigates whether gamification suggests better psychometric qualities than traditional situational judgment tests (SJT). Using a traditional situation judgment test developed by Beatriz Trigo, the researcher developed a gamified version of the SJT and compares the two samples and their results. The gamified version consists of a sample of 48 Portuguese people, who completed the gamification and the questionnaire after the game to obtain results and recode them. Afterwards, a single factor was extracted and went through several validation and unidimensionality steps to generate a creative model. From there it was put under judgement whether these gamified SJTs are the turning point to a more efficient recruitment and selection process.

Keywords: Gamification, situational judgement tests, creativity, human resources, recruitment and selection

JEL Classification: J24 (Human Capital; Skills); O15 (Human Resources)

Index

Index of tables

Index of figures

Introduction

In human resources management, recruitment and selection has been perceived as an outdated field in the fast-paced world that we live in, with procedures that have not been changed for years (Lievens & Chapman, 2019). The implementation of e-recruitment, for example, has helped the field move forward and opened doors for new promising interventions like gamification (Lievens & Chapman, 2019). Throughout my thesis it is my intention to focus the main research problem on whether the gamification process suggests better psychometric qualities than the traditional situational judgement tests (SJT).

The addition of game elements (e.g., fantasy, narrative, avatars) to assessments, also known as gamified assessments, scholars have found that it can forecast performance in academic and work-related contexts, while bringing enjoyment and flow to the candidate (Georgiou & Lievens, 2022; Nikolaou et al., 2019). Moreover, gamification can present itself as an innovative and interesting alternative to traditional SJT. These game-a-like assessments can be more compelling, engaging, less invasive, potentially useful, and a more enjoyable way for the Human Resources to evaluate candidates' psychometric qualities for recruitment and selection (Brown et al., 2022; Georgiou et al., 2022; Gkorezis et al., 2020; Landers & Sanchez, 2022; Stoeffler et al., 2020). Furthermore, the researchers Brown et al. (2022), have suggested that game-based assessments might help mitigate subgroup differences (e.g., gender, ethnicity, race), since it is less language dependent and includes less cultural aspects than the traditional SJT.

Scholars suggest that gamification and other types of game objects in assessments might improve applicants' attractiveness towards the organization. The reduction of fake results, better reliability, better measurement, reduced bias, improve fairness, and other attractive qualities to the selection process (Georgiou et al., 2022; Gkorezis et al., 2020; Landers et al., 2022).

The gaming experience is normally associated with the feelings of enjoyment which provides a better flow to the candidate. This is a consequence of the individual feeling less aware, more focused, challenged, and losing self-consciousness due to the control given over the environment (Gkorezis et al., 2020).

The Social Identity Theory (SIT) explains the relation the player can feel towards the character being played which can enhance engagement while reducing fake results attempts (Gkorezis et al., 2020; Teng, 2017).

Creativity as a cognitive ability and a personality trait plays an important part in cognitive, social and emotional activity. Subsequently, this trait intervenes in the ability of problem-solving and decision-making, that are valuable qualities for a candidate in recruitment and selection. This field of

research is highly investigated, to learn and create instruments involving the better understanding of creativity (Plucker et al., 2004; Runco et al., 2001; Sen, 2016). Nevertheless, due to this trait being hard to evaluate and non-measurable, researchers did not manage to create many high-quality assessments to the latter (Plucker et al., 2004; Runco et al., 2001; Sen, 2016).

The SJT's are typically used to evaluate cognitive abilities and personality traits, within work context, with the aim to predict the on-job performance of the respondent relative to the respective abilities/traits (Gkorezis et al., 2020; Lievens et al., 2008). Similarly, to other personality inventories, and cognitive ability tests, SJT's has its strengths and weaknesses. For instance, its strong validity, its perceived in a positive way, due to being less impactful to minorities (Lievens et al., 2008; Motowildo et al., 1990). On the other hand, the SJT is susceptible to fake results with the respondent choosing the answer that is morally more correct instead of what they would do, and being static and linear (Lievens et al., 2008; Motowildo et al., 1990). Using gamification and its properties, some of the strengths related to the SJT can be enhanced, as well as some of the weaknesses can be mitigated as mentioned previously (Gkorezis et al., 2020; Lievens & Chapman, 2019).

Nowadays, it is increasingly hard for organizations to create and sustain a competitiveadvantage. There is empirical evidence that gamification in human resources has amplified organization attractiveness, employee productivity, and in recruitment and selection how the candidate perceives the company. If a company is more innovative, modern, technologically advanced, that will lead to a higher job pursuit intention consequently increases recommendation intention, especially on candidates with high experience with video games (Ellison et al., 2020; Georgiou et al., 2022; Gkorezis et al., 2020). These characteristics are more and more important nowadays in the world full of Volatility, Uncertainty, Complexity and Ambiguity (VUCA), embracing the constant change and uncertainty that we live in (Bennett & Lemoine, 2014). One of the only sustainable advantages that companies have is through human capital (App et al., 2012; Elidemir et al., 2020). Sustainable Human Resources Management is what sets a company apart with a high-quality workforce (App et al., 2012). To do so, organizations need to attract the best candidates and become an attractive employer, which can be enhanced through gamification in the recruitment and selection process (App et al., 2012; Georgiou et al., 2022).

Game-based assessments development and validity is still an under-researched topic, more specifically, the gamification of psychometric assessments is a recent emerging trend in organizations and empirical studies (Brown et al., 2022; Georgiou et al., 2022). However, with the technological advancements society faces nowadays, and the fast paced and volatile world we live in, the option of a more modern and suitable way should be considered (Woods et al., 2020).

The remaining of this paper will be structured into 5 main topics, being the Literature Review, where one will explore in depth the existing research on the topics explored in the investigation. How

these are related and the importance in this context, with scientific articles, books and other sources to support their validity. These subjects include gamification, situational judgement tests, creativity, creativity dimensions, analytical concepts, and the conceptual model.

Followed by the Methodology, to lay out how the research was conducted and display the processes that were used and how the gamified SJT was developed. The sample and the creation of the game and questionnaire. Results, to showcase the conclusions of the gamification as well as the questionnaire. The discussion will explore the compare with the previous study (Trigo, 2021) as well as correlate the findings with the Concepts themselves to come up with supported conclusions. Lastly, Limitations and Recommendations, where the setbacks encountered during the development of this research will be reviewed, as well as recommendations based on the limitations found and that could be valuable for future research in this topic. Followed by the conclusion with a brief closure onto the main research questions that this paper intends to answer.

Considering the previous, it is intended to find the answers to the following questions:

- To discover what are the main differences between creativity assessment through traditional SJT and gamification?
- To explore how gamification assessment can be different from a traditional SJT?
- To evaluate how to develop a gamified assessment to evaluate creativity?

CHAPTER 1

1. Literature review

Recruitment and selection are crucial strategic domains in human resources management (HRM), but it is often perceived as an outdated field of HRM. With procedures that have not altered for years which do not catch up with the internal and external organizational environment of constant change developed nowadays (Lievens & Chapman, 2019). Moreover, recruitment and selection has been overgeneralized, to a question of formal or not formal tests, whether the interviews were conducted in structured way or not, and efficient recruitment being associated to highly capable applicants (Lievens & Chapman, 2019). However, the latter still tackles crucial aspects of recruitment and selection but diminishes the advanced level that research on the field has accomplished over the years (Lievens & Chapman, 2019). E-recruiting, renewed status of the recruiter, and applicant responses to the selection process, have been recent and innovative trends in the field. This has helped to mitigate some of the stereotypes of traditional recruitment. The future direction for more promising interventions in the field of recruitment and personnel selection, are seen to be gamification (Lievens & Chapman, 2019).

Creativity has been shown to be a crucial psychometric quality in recruitment and selection. Today's world is in constant improvement, and like that, it is crucial for companies to hire minds that evolve and become assets to the company itself. Through the development of new and original ideas to answer difficult work problems from a new perspective, being valuable to a business by boosting its success and competitiveness (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020).

1.1. Gamification

There are three different game-related classes of assessments which are Game-based; Gameful design; and Gamification (Landers & Sanchez, 2022). Game-based is seen as when the applicants are the players in a core gameplay experience with traits implied. Gameful design, on the other hand, is when the mechanics or concepts of a certain game guides applicants to the decision making. Lastly, Gamification, also known as gamified assessments, is the addition of game elements (e.g., fantasy, narrative, avatars) to assessments (Chan et al., 2018; Gkorezis et al., 2020; Landers & Sanchez, 2022; Salikutluk et al., 2019).

Over the last few years, gamification has been trending in a variety of online context, including online learning, business and education for a positive reaction in user experience of enjoyment, fun (fascinating, entertaining and captivating) and engagement (Chan et al., 2018; Georgiou & Lievens, 2022; Gkorezis et al., 2020; Salikutluk et al., 2019). Companies are adopting gamification to enrich

costumer's reliability, employee productivity, and organizational attractiveness (Chan et al., 2018; Georgiou & Lievens, 2022; Gkorezis et al., 2020). The use of gamified assessments in the recruitment and selection process has positive effects on the applicant's perception of the company, giving feelings of technologically advanced, modern and innovative, as well as recommendation and job pursuit intentions (Georgiou & Lievens, 2022; Gkorezis et al., 2020). These feelings can also translate to symbolic attributes of trendy, creative, intelligence, effective and original, that enhances organizational attractiveness and a more pleasant experience for the applicant compared to traditional assessments (Georgiou & Lievens, 2022; Gkorezis et al., 2020).

Moreover, assessments with game elements, such as avatars, narrative, fantasy, are more enjoyable or satisfying for the applicants compared to their counterpart without these elements and enables the execution of psychometric scales (for example, personality traits, motivation, values), due to data collection through technology. (Georgiou & Lievens, 2022; Gkorezis et al., 2020; Ostrow & Heffernan, 2018).

1.2. Situational Judgement Tests

Situational judgement tests (SJT) are a method to measure cognitive ability and personality inventory traits, in job-related situations. The applicant answers with the action that it would most likely be the one performed by the former (Gkorezis et al., 2020; Lievens et al., 2008). The answers available on these assessments are not right or wrong, they are based on possible behaviors one will pursue when faced with the situation in question (Lievens et al., 2008). In other words, it can be perceived as a job simulation, and it may predict how well the applicant will perform on the job. Compared to the self-reports, where the participant evaluates their own abilities, SJT's due to the introspection that these games might need, applicants are less vulnerable to faking the results (Oostrom et al., 2019).

The SJT initially developed by Motowidlo, Dunnet, and Carter (1990) like other personality inventory and cognitive ability tests, has its strengths and weaknesses. When developing these kinds of assessments, allows the researcher to have a strong criterion-related validity, incremental validity, the applicants perceive the assessment in a positive way, due to being job related and less impactful to minorities and be accessible to a larger number of applicants through the internet (Lievens et al., 2008). On the downside, applicants can fake results by impersonating what they should do and not what they would do, creating a character that would please the recruiters instead. However, this action can be somehow related to coaching in order to get better results in the SJT (Lievens et al., 2008).

Moreover, in the last years SJT's had a significant increase in adherence on research and on practice, especially has a predictor of job performance (Olaru et al., 2019; Oostrom et al., 2019). Research has shown that SJT's can explain variance in job performance, when monitoring the results

of other traditional predictors (for example, cognitive ability, personality, and job experience) (Gkorezis et al., 2020; Oostrom et al., 2019). Typically, SJT's can be perceived as static and linear, hence realistic stimulus for example, 2D animated, video-based, 3D animated, have been developed and can break that barrier through gamification, or in other words, "Serious games" (Gkorezis et al., 2020; Lievens & Chapman, 2019).

The traditional SJT are known to be when the applicant is presented with different written scenarios and has to explain the proper response alternative, these can be video-based or multimedia SJT's. More specifically, the scenarios are presented in a video-format, these are always job-related situations, and the video will end at a critical moment it is then that the applicant must select which action or multiple actions would be the most correct course of action (Lievens et al., 2008). Despite its content, SJT's are always presented in a multiple-choice format, however, depending on the response type, its reliability degree changes. If the responses are written, it is considered to be low reliability, video-based medium reliability and when reacted to the behavior without being asked to, high reliability.

Even though, SJT's have lower reliability compared to other personality inventories, by being answer restricted instead of an ongoing behavior evaluation, it captures a variety of personality traits that other experiments would not detect (Lievens et al., 2008).

Personality traits in the workplace environments, has the literature supports, is of major importance. Therefore, HR professionals rely frequently on performing personality assessments (Oostrom et al., 2019). It was in a 2015 survey conducted to HR specialists (Ryan et al., 2015) the most common aspect recruiters look for when evaluating the candidate are the inner characteristics towards the job itself but most important the overall personality. Personality assessments are frequently done through self-reports, but over the last few years, the debate over self-reports and faking their results has increased in the literature. Due to the latter, researchers of this field are in the search for new techniques for alternatives to the self-reports, with a great degree of reliability and validity as well (Oostrom et al., 2019). Over the years researchers have found different techniques to increase the criterion-related validity of personality measures. It is through the contextualization of the questions presented to the applicants and by adapting them to workplace environments, the validity of the results will increase (Holtrop et al., 2014; Oostrom et al., 2019). SJTs have never been so used as they are right now. Due to its present accuracy, consequence of its specificality, businesses trust more and more in these processes to evaluate intangible skills from its recruits (Oostrom et al., 2019).

1.3. Creativity

Creativity has a critical role in cognitive, social and emotional activity. However, high-quality creativity assessments are still lacking (Runco et al., 2001). Researchers believe that creativity as a cognitive ability and personality trait, is hard to evaluate. The researchers Plucker et al. (2004), suggested that the oversimplification of creativity had repercussions on the stereotype of creativity and in research potential, due to being non measurable, prone of reduction. For a long time, the latter has been a major interest for researchers in the field with decades of theoretical and empirical research that has generated many instruments allowing the conduct of various aspects of creativity (Sen, 2016).

Because creativity plays a crucial role in problem-solving and decision-making, it is a crucial psychometric quality for recruitment and selection. It enables people to develop fresh, original ideas that can be helpful in a variety of circumstances, such as when dealing with a difficult work or issue or when looking for a solution to a complicated problem (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020). Additionally, creativity can encourage individuals to think creatively and develop novel solutions to issues, which can be very useful in today's fast-paced and constantly evolving work world (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020). Creativity is crucial for personal development. In addition to its many practical advantages, it enables people to express themselves and investigate their own thoughts and ideas, which may be satisfying and rewarding (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020). It is crucial for businesses to look for applicants that exhibit creative thinking abilities during the recruitment and selection process, since these people are more likely to approach their work from fresh angles. Overall, because it can enhance the success and competitiveness of the firm, creativity should be considered when hiring and selecting workers (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020).

There has been an increase in studies that explore the variables that can positively impact increased creative and/or innovative behavior. These can be perceived as predictors (Runco & Chand, 1995; Runco et al., 2001; Tan et al., 2019; Trigo, 2021). In the traditional SJT developed by Trigo (2021), the variables that were used to inspire its development were divergent thinking, risk taking, intrinsic motivation, emotional management, and teamwork. However, in the current research it was measured the variables of divergent thinking (Plucker et al., et al., 2006; Runco & Chand, 1995; Runco et al., 2001; Sen, 2016), intrinsic motivation (Chan et al., 2018; Georgiou & Lievens, 2022; Runco & Chand, 1995; Salikutluk et al., 2019), and openness to experience (McCrae, 1987; Tan et al., 2019; Woo et al., 2013) that were hypothesized has having a strong relationship with creativity, that will be further explained in the next section.

1.4. Creativity Predictors

The researchers Runco & Chand (1995) hypothesized that divergent thinking (DT), and intrinsic motivation are the main predictors of creativity or creative behavior. However, there are also other important components such as, problem finding, strategy, judgement, extrinsic motivation, procedural and declarative knowledge (Plucker et al., 2006).

Moreover, Hocevar (1981) concluded that self-reported creative activities and accomplishments were superior to other methods of assessment because of its high face validity and predictive utility. On the other hand, the latter has a critical limitation of evaluating creative performance, instead of creative potential (Sen, 2016). Despite creativity having a serious gap for assessment, due to socioeconomic status and cultural environmental factors, not everyone has access to an engaging level of creative activity. Which results in rewarding factors such as chance, culture, education, status, and affluence, rather than the target construct of creativity itself (Sen, 2016).

The researchers Runco et al. (2001), developed a new scale to measure creative ideation. This was based on the belief that ideas can be treated as the product of original, divergent, and creative thinking. The same researcher (2001) developed the Runco Ideational Behavior Scale (RIBS), this scale assesses the ideational capacity that could be the target for creativity assessment. Where most items are described as real behaviors, that reflect an individual's usage of skillful ideas.

The RIBS appears to be a reliable instrument, both in individual and group usage. In contrast to other scales developed at the time which did not provide unambiguous evidence about the construct validity of the RIBS. Furthermore, in Plucker et al. (2006), it is hypothesized that the RIBS is a useful measure for cross-cultural research. The RIBS scores evaluate creative potential that is not caused by cultural traditions and social constraints. Opposed to, the Torrance Tests of Creative Thinking (Torrance, 1974) which are culturally bounded to the Hong Kong subjects (Plucker et al., 2006).

Divergent thinking (DT) tasks are one of the most used techniques, allowing to quantify produced ideas through their multitude. The categories they belong to and their infrequency, can predict creative potential (Plucker et al., 2006; Sen, 2016). The researchers Runco and Plucker believe that everyday creativity can be quantified by people's ideas (Runco et al., 2001).

Nowadays, students are more technologically connected, motivation is a critical factor in their engagement while learning. To retain knowledge, educators globally have adopted gamification to motivate the new generation of students through more diverse and challenging rewards (for example, points, badges, level-up's). As hypothesized by the researcher Salikutluk et al. (2019), feedback and reward systems through gamification generally enhance extrinsic motivation. On the other hand, these processes balances with the long-term performance and engagement, that are important for intrinsic motivation. Which increases user's overall satisfaction and well-being (Salikutluk et al., 2019).

As the researcher Chan et al. (2018) explored, intrinsic motivation can be referred to as "people's interests and values that are in line with their basic psychological needs". Furthermore, these can be evaluated through the Motivation Scale (SIMS), developed by the researchers Guay et al., (2000), that references the hierarchical model of intrinsic and extrinsic motivation (Vallerland, 1997). This is composed of 16 items that focus on the reasons why an individual would be involved in the execution of an activity, through a Likert scale going from 1 to 7. This scale has been used in a variety of studies (Gamboa et al., 2013; Gillet et al., 2012; Ratelle et al., 2005), where it presents good internal consistency values with the 0.95 in intrinsic motivation (Guay et al., 2000). Afterwards, it was developed a Portuguese adaptation of the Motivation Scale (SIMS) (Gamboa et al., 2013), where the internal consistency levels were not has high as the original model, nevertheless with the removal of the items 10 and 11, it spiked the consistency levels from 0.81 to 0.87 which is satisfactory, and obtained results supports the use this scale (Gamboa et al., 2013).

Gamification can bring intrinsic factors experienced while playing games, such as flow and enjoyment that can be perceived in a positive way to the system being used (Georgiou & Lievens, 2022). When applicants are within the flow, these immerse themselves in the experience where awareness and self-consciousness become limited to the experience itself. This is the result of the perceived feeling that they are in control of the environment, which consequently affects their actions and behaviors. This could mitigate the problems of traditional SJT's and consequently traditional recruitment and selection, such as faking, impersonating, coaching, and being answer restricted. (Georgiou & Lievens, 2022).

Openness to experience is a major personality that is poorly understood, and poorly explored in terms of the relationship with creativity (Tan et al., 2019; Woo et al., 2013). This trait can be perceived as being imaginative, having high sensibility in arts and having an intellectual curiosity (Vasconcellos, 2008). Among the Five-Factor Model, openness to experience had consistently a positive relationship with creativity (McCrae, 1987; Tan et al., 2019), that has been barely investigated.

The researcher Tan et al., (2019), explored this relationship and even further analyzed with the use of intrinsic motivation as a mediator between them, which presented great results with the replication being successful, but also when the respondents feel more open to the experience, they would be more motivated and enhance their creativity.

Moreover, it was also found another linkage of openness to experience with divergent thinking in the article of the researcher McCrae (1987), where it was explored the connections of creativity and divergent thinking with openness to experience, which it was concluded that creativity from the Five-Factor Model, had the most strong and consistent relationship with openness to experience, but also divergent thinking was consistently correlated to the latter.

The openness to experience can be measured through the revised NEO personality Inventory (NEO-PI-R) (Costa & McCrae, 1992), as well as the other personality traits from the Five-Factor Model, with solid fundamentals and good reliability. However, the openness to experience can be considered an independent personality dimension (Garcia et al., 2005). The researcher Vasconcellos (2008), developed a scale of openness to experience adapted in Portuguese, based on the NEO-PI-R, which the dimensions obtained goes in concordance with the structure of the inventories, and the facets presented. The adaptation showed good psychometric properties, as well as an adequate internal consistency for each of the scales extracted (Vasconcellos, 2008).

1.5. Analytical Concepts

The basic Rasch model was initially developed with the purpose of comprising tests with dichotomous items measuring one latent attribute (Rasch, 1960/1980; Wright & Stone, 1979). The dichotomous Rasch model or also known as simply the Rasch Model, uses a dichotomously scored items that can be described as binary results, wrong or right, and even “Fail” or “Pass”, meaning that they are mutually exclusive response alternatives (Aryadoust et al., 2021; Wright & Masters, 1982). This format is mostly used in scoring performances on educational tests data, as well as in construction and maintenance of item banks (Wright & Masters, 1982; Wright & Stone, 1979). In other words, answers are given either number 0 if incorrect or 1 if correct, and it is the sum of these values, which will determine applicants' degree of what is being investigated in the experiment.

The idea behind the Rasch measurement is that the variation in performance from the participants is due to a latent attribute that the test sets out to evaluate (Linacre, 2005). Similarly, this is the idea behind item response theory (IRT) model. The relation of both is extensively debated between research (Aryadoust et al., 2021) on how IRT models are descriptive and use a model-to-data fit, while the Rasch is prescriptive and uses a data-to-model fit (Aryadoust et al., 2021; Bond & Fox, 2015). The IRT can also be used as an alternative to the Classical Test Theory (CTT), that due to not being sample dependent, uses a parameter invariance to mitigate these problems (Ferreira et al., 2011). Additionally, it has also been demonstrated that the use of IRT can be reliable in cases where the CTT fails. IRT due to his algorithm, the sample size does not play a role in determining the results. Uncommonly this approach balances results by calibrating the scale to enable an easier analysis and interpretation between them by bringing people versus items into similar terms (DeMars, 2010; Ferreira et al., 2011). The researchers Ferreira et al., (2011/2012), have provided evidence in the use of IRT to obtain unidimensionality and measure cognitive aptitudes. More specifically working memory, where in both cases the IRT demonstrated several advantages compared to the CTT using calibrated measures (DeMars, 2010; Zickar & Bradfoot, 2009). On the second case it was considered

the IRT characteristics such the difficulty of the questions (Magno, 2009) to improve the model and differentiate participants ability levels. This led to less inconsistencies and less measurement errors (Magno, 2009). The use of IRT can also be complemented with the partial credit model for polytomous scoring and dichotomous Rasch model for dichotomously scored items, depending on the most proper use for the data in question (Ferreira et al., 2011/2012).

Over the years, this model kept evolving and extended to a new concept of parameterizing polytomous scales. These scales reward the respondents for achieving intermediate levels, in contrast to the use of dichotomously scored responses with the only answers being right or wrong (Masters, 1982; Wright & Masters, 1982). More specifically, instead of scaling between the absolutes of 0 or 1, this scale tests between higher intervals, considering all answers, expanding the possible results. The use of partial credit scoring is usually to lead to more precise data of a people's capability compared to the dichotomous scores, and the polytomous scales includes Likert scale, which is very common in surveys and self-appraisals (Aryadoust et al., 2021; Masters, 1982).

The creation of more models derived from the Rasch model, for instance the rating scale model (Andrich 1978; Wright & Stone, 1979), partial credit model (Masters, 1982; Wright & Masters, 1982), and later on, the development of many more probabilistic models that are part of the Rasch measurement family which spiked interest within the psychometric experts in major centers, and being adopter in a larger scale (McNamara & Knoch, 2012; Rasch, 1960/1980; Wright & Stone, 1979).

1.6. Conceptual Model

Later in the study, it will be verified the validity of the model and its essential to understand the relationship between the variables and why they were chosen in the first place. Even though to this point of validating the model, the internal consistency and content validity should be secure, it is valuable to explore the relationship with variables that is expected to correlate, in other words check the criterion-related validity (Hinkin, 1988).

Criterion-related validity, it is used to analyze the relationship between variables that have a hypothesized relation and to create a nomological network (Hinkin, 1998; Cronbach & Meehl, 1995).

These relations should be theory based and be analyzed using correlation or regression analysis, and the criterion-related validity is proven when the hypothesized relationships have statistical significance (Hinkin 1988).

As mentioned previously, creativity plays a crucial role in terms of problem-solving, decision-making, as well as in the cognitive, social and emotional activity, however as a cognitive ability and personality trait, it is hard to evaluate (Runco et al., 2001; Anderson et al., 2014; Klepić et al., 2020).

To verify the criterion-related validity, since the model is designed to try to evaluate creativity, it is necessary to find the hypothesized relationships with creativity which could be its predictors.

Firstly, there is not a “best predictor” due to not having available many high-quality tests of creativity, rather the predictors of Divergent thinking (DT) and Intrinsic Motivation (IM) are considered major predictors of creativity. According to the researchers Runco & Chand (1995). The last predictor is Openness to experience, its hypothesized to relate closely with creativity (McCrae et al., 2014; Woo et al., 2013).

Divergent thinking can be perceived as the inclination to explore original ideas and being able to think outside the box (Runco et al., 2001). For intrinsic motivation, it is a crucial factor in evaluating creativity, where a person’s willingness to do an activity can foster higher levels of creativity and inventive thinking (Chan et al., 2018; Salikutluk et al., 2019). Openness to experience is considered to be a personality attribute that allows for the subject to be open to new ideas or experiences and be curious or willing to explore new concepts (McRae et al., 2014; Woo et al., 2013).

In sum, to obtain criterion-related validity it will be correlated the model with divergent thinking, intrinsic motivation and openness to experience, and analyze if the correlation between them is significant to check the validity as a creativity model.

CHAPTER 2

2. Methodology

The aim of this study is to test a gamified version of the traditional SJT (Trigo, 2021), gathered results through the gamification, develop a factor and check its reliability and validity. Followed by the discussion of the results, the finding and the respective conclusions. To develop the judgement on whether the gamification process suggests better psychometric qualities than the traditional SJT and compare some of the results.

In doing so, the results will be recoded into a partial credit model (PCM) and a dichotomous Rasch model, converting the 4 possible answers in a scale of 0 to 3 from worst answer to best according to Trigo (2021), and a scale of 1 for the best answer and the remaining 0 respectively, to verify which model is more adequate for the data. The factor will be developed through the reliability analysis. Due to the sample size being low it is not possible to generate an exploratory factor analysis, and afterwards the item-response theory (IRT) will be applied to check the factors unidimensionality. To examine the validity, the factor will be correlated to the creativity predictors, divergent thinking, intrinsic motivation, and openness to experience, which had a section for each one in the questionnaire with questions from researched papers to test the latter.

2.1. Timeframe

To evaluate whether gamification produces better psychometric qualities than the traditional situational judgement tests is mandatory to compare the two models in a proper way. These could be implemented in the future of recruitment and selection, for better job performance, organizational attractiveness and other advantages.

Gamification is the addition of game elements (e.g., fantasy, narrative, avatars) to assessments, in other words gamified assessments (Chan et al., 2018; Gkorezis et al., 2020; Landers & Sanchez, 2022; Salikutluk et al., 2019).

Considering the VUCA world and the constant change one lives in, organizations need to keep up with the market, and be constantly innovating. Employees are made accountable to come up with creative solutions, in order to stay competitive. Therefore, the human resources department is under a lot of pressure to ensure the hired workforce is a fit to the culture as well as creative in the market. This puts high stakes for the Human Resources specialists, to keep up with the expectations (Govendo, 2005).

The traditional recruitment and selection techniques are getting outdated, to the high technological advancements. Despite recruiters' usage of remote work and social media

advertisements to attract candidates to new vacancies is increasing, the usage of other methods such as, the traditional (self-reports) personality assessments are vulnerable to fake results which brings flaws to the hiring process of organizations (Lievens & Chapman, 2019).

Situational judgement tests (SJT) can measure cognitive ability and personality inventory traits, in job-related scenarios by answering critical questions on what would be the most likely course of action one would take at a given moment. In other words, job simulation (Gkorezis et al., 2020; Lievens et al., 2008). These assessments suggest a higher validity and engagements, which consequently makes applicants less prone to faking inputs resulting in a more accurate job performance prediction. Since these are mostly applies in a job-related context, candidates are made aware of the situations they might be put in and therefore, enhances engagement flow which reduces the possibility of fake results, and enables more attractiveness towards the organization with positive feelings of fun, modern, among others (Chan et al., 2018; Georgiou & Lievens., 2022; Gkorezis et al., 2020; Oostrom et al., 2019; Salikutluk et al., 2019).

Despite its importance, creativity is hard to evaluate due to being non-measurable and prone to reduction. Several techniques, such as behavioral interviews, case studies, and problem-solving activities, have been developed to evaluate this trait. There is not a single "best" predictor of creativity because there are not many high-quality tests available, and different approaches may be helpful depending on the context and assessment goals (Plucker et al., 2006). According to the literature, divergent thinking, intrinsic motivation, and openness to experience, are the main predictors to gauge creativity (Plucker et al., 2004; Plucker et al., 2006; Runco et al., 2001).

Divergent thinking is the capacity to come up with several alternatives or solutions to a given issue. It is sometimes regarded as a crucial component of creativity. Instead of focusing only on one predefined solution, divergent thinking entails thinking outside the box and taking into account a wide range of different ideas or approaches (Plucker et al., 2004; Plucker et al., 2006; Runco et al., 2001). In addition, divergent thinking can be evaluated and found to be useful in a variety of situations, including recruitment and selection, assessing creativity training programs, and determining an individual's overall creative potential (Plucker et al., 2004; Plucker et al., 2006; Runco et al., 2001). Divergent thinking assessments like the Runco Ideational Behavior Scale (RIBS) have proven to be reasonably accurate indicators of creative ability. The scale has 20 items that measure an individual's flexibility, originality, and ideational fluency (the quantity of ideas generated and the ability to shift between different perspectives) (Runco et al., 2001).

Intrinsic motivation is a crucial element to consider when evaluating creativity because it can affect a person's degree of involvement, willingness to explore new ideas, and ability to think creatively. People that are intrinsically motivated are more likely to be involved in what they do and

to follow their passions, which can foster better levels of creativity and inventive thinking (Auger et al., 2016; Chan et al., 2018; Salikutluk et al., 2019).

The Motivation Scale (SIMS) is a valid and reliable tool for measuring intrinsic motivation, which is the desire to carry out actions for their own sake as opposed to receiving benefits from others. The scale is comprised of 16 items in a Likert scale from 1 to 7, and this scale has been used in a variety of studies (Gamboa et al., 2013; Gillet et al., 2012; Ratelle et al., 2005), which presents good internal consistency values (Guay et al., 2000). The researchers Gamboa et al., (2013) have adapted the scale to a Portuguese population with good internal consistency and results that supports the latter.

Lastly, openness to experience, this predictor is a personality attribute that indicates the readiness one has to consider new concepts, experiences, and the curiosity and exploratory mindset.

According to research, creativity and openness to experience go hand in hand the more one is open to new ideas and experiences the more likely is to be responsive to new ideas and to be more flexible and adaptable in their thinking (McCrae et al., 2014; Woo et al., 2013).

Assessing openness to experience can be important in a variety of contexts during the recruitment process it is through the assessment of the later that one can start to have an idea of an individual's overall creative potential (McCrae et al., 2014; Woo et al., 2013). The Openness to experience can be measured through the revised NEO personality Inventory (NEO-PI-R) (Costa & McCrae, 1992), since it's part of the Five-Factor Model. However, a study developed by the researcher Garcia et al., (2005) suggests that this personality trait could be perceived as an independent personality dimension. The investigator Vasconcellos (2008), developed a Portuguese study to openness to experience scale. Based on the NEO-PI-R, with 42 items, similar dimensions, and facets, showed good psychometric properties, as well as an adequate internal consistency for each of the scales extracted.

The execution of the research was done over a time span of 1 year (Figure 9). In order to compare the two SJTs the researcher created a gamified version of an existing traditional SJTs so the results could be compared. Firstly, it was brainstormed what is the vision for the gamification and it was brainstormed with my supervisor and my colleague that developed the previous SJT (Trigo, 2021), how could one adapt her vision to a gamification without excluding vocal points of the original SJT and keep the same objective. From there a third person, office theme was chosen with the respondents having as their goal the completion of the daily tasks of a "regular day" at work.

Following this stage, it was explored the best game development tool to use. Unity was the chosen program to create a real-time 3D game and reflect the scenarios previously explored. The language used for programing was C#. The theme of the scenes was a translation of the environment the previous population was inserted in and from there the purchase of assets, characters and its customization started. The cast of the game wanted to be inclusive, where one could relate to the personalities encountered during its gaming journey. The creation of the scenarios as well as its

characters was completed in a period of 2 months. Since there was a lack of 3d modeling skills to create new assets, the Unity store was used for assets that fitted the idea and description. Through these assets, it was created an office scenario through the mix/addition/removal/modification of all the assets and the code behind some of them to match my idealization.

Following the creation of the 3D SJT, the investigator moved to adapt the original items of the traditional SJT to a dialogue form and split or merge some of the items into different dialogues. In order to give flow to the interactions, and reduce the play time for the user, more than 1 item was combined to mitigate the same feedback given previously of the situational assessment being too long for the respondents.

As the researcher developed the game there was a lot of testing, bug fixing, redefine game strategy and assess the viability of the game. This stage was dedicated to refining the gamified assessment, taking into account the feedback given and the results from the quantitative data, so it would work as intended and assembled a minimum viable product (MVP) to be used on the following phases.

Lastly, the investigator started gathering responses through social media posts, word of mouth, e-mails, and directly contacting institutions to gather respondents. Like that the population was created through a snowball effect.

2.2. Sample/Population

The sample used in this study is composed by a total of 48 Portuguese participants, from different backgrounds. Where 29.17% (N = 14) identified itself as the female gender, and the remaining identified as male, constituting 70.83% (N = 34) (Table 10; Figure 10). In the same sample, the average respondent had an age of 27.31 years old (SD ≈ 8.72; Table 11) with a right skewed distribution (Figure 11), going from the minimum of 19 to a maximum of 58 years old, with a vast majority being 24/25 years old with 56.25% (Table 12) of all candidates and a mode of 24 years old (Table 11) (Fisher & Marshall, 2009; Haden, 2019; Rendón-Macías et al., 2016). The population consented to participate in this study after a brief description of the study, its contribution and the objectives and goals to be achieved. It is important to note that the identity of the the respondent was kept anonymous and confidential at all times.

About 47.92% (N = 23) of the respondents are at the moment of the study, working, 50% (N = 24) are students, and only one respondent has chosen as the occupation “other” instead of the 2 previous categories (Table 13). The working sample has a wide variety of industries and therefore these cannot be organized. However, one can say that the majority of the sample (6.3% of the hospitality sector) was categorized as being part of the service sectors (Table 14).

2.3. Instruments

The development of the gamification was made through the program Unity and took the researcher about 6 months to fully develop, discarding implementations that were later made based on feedback. This application can develop games for a large range of devices. The programming language used in Unity it is C#, and has an online asset store with 3D models, to facilitate the investigator game development. Microsoft Visual Studio is a complementary platform that enables the writing of the code since one cannot code directly on Unity. This platform was used later on to connect the game with the online server and make it available to the participants.

All the code developed in Unity and the Azure Functions to connect the game and the online database was written through Visual Studio since it is a Microsoft tool that has a great adaptability to Azure.

For the connection between the game and an online database it was used Microsoft Azure, where the item responses and other relevant data like gender, age and identification (last 4 id digits for confidentiality and anonymity), where it was then converted into a JSON string and posted directly to the data storage container. Once the respondent finished the game through an Azure function it was enabled to get the data, increment it with the new data in the game and post it back to the database with the old and new data. It was chosen to use a JSON string type, due to being light in storage, connection size, and also being understandable to the naked eye and effortless code wise. Afterwards, once the data is posted in the Azure Container it can be easily downloaded and imported to Microsoft Excel as a structured database for further analysis.

In order to collect data of the predicts to later check the validity of the items, it was used Qualtrics to host the questionnaire.

To allow the gamification to be posted online and played, it is necessary for a server to host the game. To make the game available to external participants, it was necessary to use Itch.io. This platform is dedicated to host, sell and download indie video games of Indy developers (solo or small groups of game developers without vast resources), game assets, comics, zines and music, with a variety of personalization, freedom and stable servers that incorporates HTML 5, mandatory to host WebGL Unity games (browser played games).

Later, the collected data was analyzed through SPSS, a tool that allows to gain insights from small to large-scale data with sophisticated statistical analysis with powerful survey research and precise identification of trends.

Lastly, after developing a factor and examining the reliability analysis in SPSS, it was applied the Winsteps program to conduct a Rasch analysis to assess the factor reliability and separation, item and persons fit statistics to examine the factor's unidimensionality.

2.4. Procedure

The development of the gamification took 6 months to fully develop. Initially it was brainstormed what was the vision for the gamification and along with the developer of the previous SJT (Trigo 2021), how could the researcher adapt the vision to the gamification. The reproduction of the SJT could not lack on vocal points of the original SJT and had to keep the same objectives. The game happens in an office and the goal is to resolve the daily tasks given. To speed the process of development, all items used for the development of the game were purchased in a 3d Unity asset shop and reshaped to fit the idea and description. Through these assets, one created an office scenario with mix, addition, removal, modification of all the assets and a new coding on some of them to match the idealization.

Furthermore, it was proceeded to adapt the original items of the traditional SJT to a dialogue form and split or merge some of the items into different dialogues. The combination of items allowed a better flow to the game and conversations developed and enabled a reduction of play time for the user, since the original SJT was commented to be too long for the respondents.

Some of the scenarios for these dialogues are a job interview, a remote call, a presentation and so, to immerse the respondent in the narrative (Figures 1 to 5).

Figure 1
In-game screenshot 1

Figure 2
In-game screenshot 2

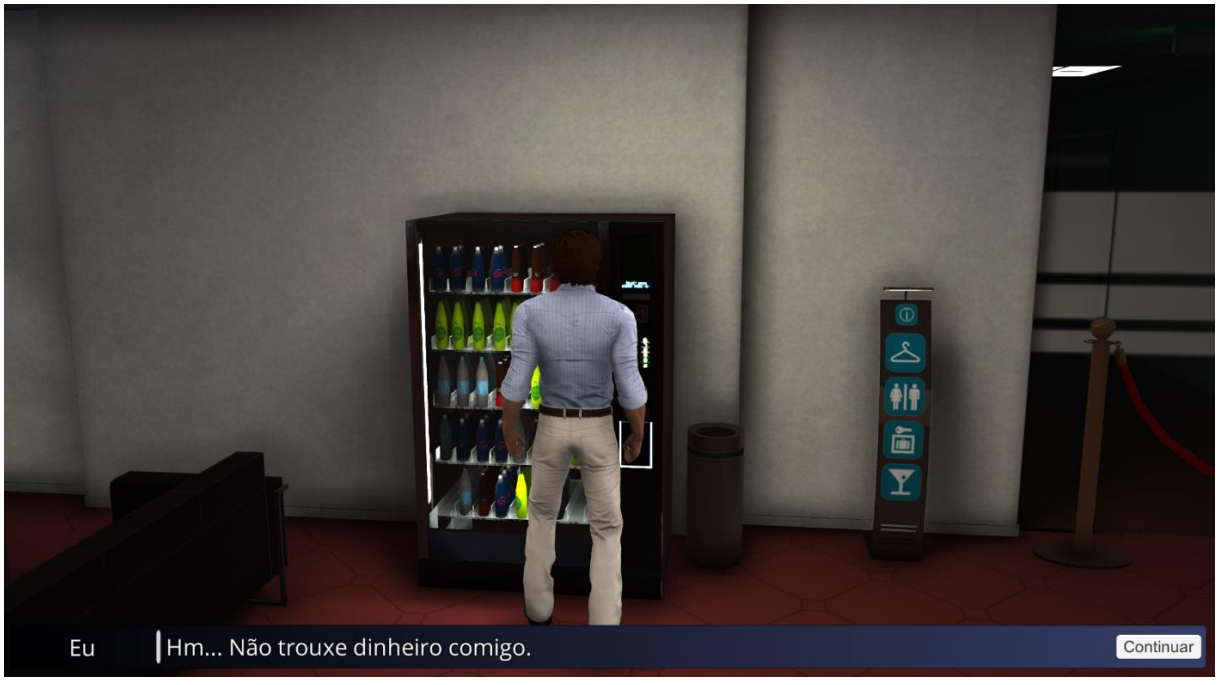


Figure 3
In-game screenshot 3

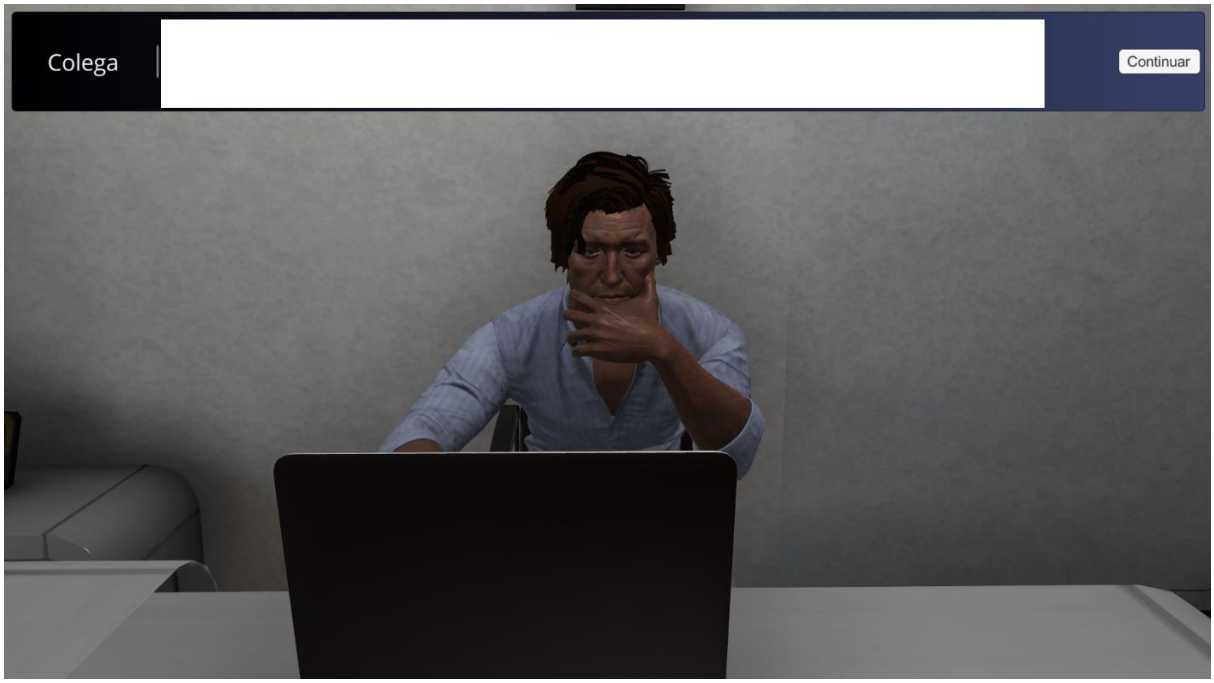


Figure 4
In-game screenshot 4

Figure 5
In-game screenshot 5

The asset used for the game cast was a cowboy themed asset and therefore all postures, gestures and clothing had to be modified to be more business attire. In addition, textures, colors, hairstyle/color, eye color, facial hair, skin tone had to be readjusted to create 15 unique characters based on only 4 models (Figure 1, 4 and 5).

Once the dialogue and the characters were created these had to be combined with its animation to give a sense of realism and interactivity between characters during dialogue within the context, they are in. The camera angles the player was in was a crucial part of the development to mitigate flaws in the animations. The overall experience of the game had to be tested consecutively with the addition of further writing or modification of the code for various parts to create a smooth experience and tackle the major problems presented. Moreover, there was an on-going developing gameplay quality, bug fixing and quality implementation of life features such as User Interface (UI). This includes an in-game pause menu, some of the tasks that are UI dependent, a minimap to help with user orientation around the building, top left corner icons (Figure 4, 6, 7) to remind the respondent on how to access the latter, dialogue boxes and buttons optimized and modified color scheme to adapt the environment.

Figure 6
In-game screenshot 6

Figure 7
In-game screenshot 7

It was created a pre/pos game scene, as well as a loading screen for user experience, this scene is compiled by a background representing what will be happening in the game, as well as a starting menu with empty fields for the user to fill with information to later be analyzed and match with the questionnaires. The post-game scene has the objective to let the user know that the game is finished with a friendly reminder to do the questionnaire and to thank the participation. This allows the right time for the answers to go to the database. In the case of the loading screen, its implementation has

as objective to have something interactable while the scene is being loaded so the user does not think the game crashed or froze.

Aside from the gamification development, with the gamification working as intended, it was started to investigate and develop an online database as stated in the instruments section. After the creation and configuration of the data base through the platform azure to contain a JSON file with the data, it was coded in C# a function that enables the flow of the data between the gamification and directly to the data base with a structured basis.

Moreover, it was developed some coding functions in the gamification to store the initial data and the items responses by the users. These responses were structured and converted to the JSON format before being sent. During the gameplay the data is stored every time an answer is selected and through the latter functions in conjunction with the azure functions, sends all the data to the database while loading the post-game scene.

For the gamification to be available online, a user-friendly platform was used to publish the game being simple to use for any external participant to access. However, the size exceeded the limit, and it was necessary to convert the meshes into smaller portions which affected negatively the visual quality of the game.

During the data collection phase, still some suggestions were being given and new game versions were created accordingly. These included an arrow to guide the player to the next nearest quest, to assist with the in-game orientation, some bug correction and the addition of easter eggs (non-game impactful game interactions) like talking to the vending machine, to give a more immerse and game a like sense for the respondents.

2.5. Questionnaire

The questionnaire, was developed on the platform Qualtrics. Firstly, the questions and its value to the research had to be investigated and validated with the data presented so it would be easy to correlate information at a later stage. In the first section, the participants had to fill in general questions like age, job, gender and so on for descriptive statistics.

For the validity, the predictive tools used are research proven tests about divergent thinking (Runco et al., 2001), intrinsic motivation (Gamboa et al., 2013), and openness to experience (Vasconcellos & Hutz, 2008), adapted in Portuguese to better be interpreted, since every respondent is Portuguese.

Later, the items with the greatest explanatory percentage were selected to reduce the questionnaire length time. After the latter, the data were extracted and transferred to SPSS database. Followed by the next objective, to organize the data collected by removing invalid or incomplete lines,

recode the scales to be all uniform, and creating dichotomous and partial credit model to be later on used on the analyses.

At this stage the ID given at the beginning of the gamified SJT and the questionnaire were matched to complement each other's information in the analyses creating a new database with the merged data. In the analysis phase, firstly a descriptive analysis was done on the population. The items and other relevant information, and the given feedback was put into excel for organization and interpretation purposes. Afterwards, it was tested to see which of the following models were better to develop a factor, with the dichotomous Rasch model and the partial credit model with decent result. However, due to the low sample, it was not possible to do an exploratory factor analysis, having a necessity to calculate a factor with a Cronbach's alpha superior to 0.70 (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011). A Rasch analysis was conducted through the program Winsteps with the model from the factor analysis to test the model reliability and separation, item and persons fit statistics and lastly its unidimensionality and local dependence to estimate the model's efficacy and internal consistency reliability. With the latter being verified, the factors validity was checked through the correlation between the factor and creativity predictors, to later conclude whether the model is viable.

CHAPTER 3

3. Results

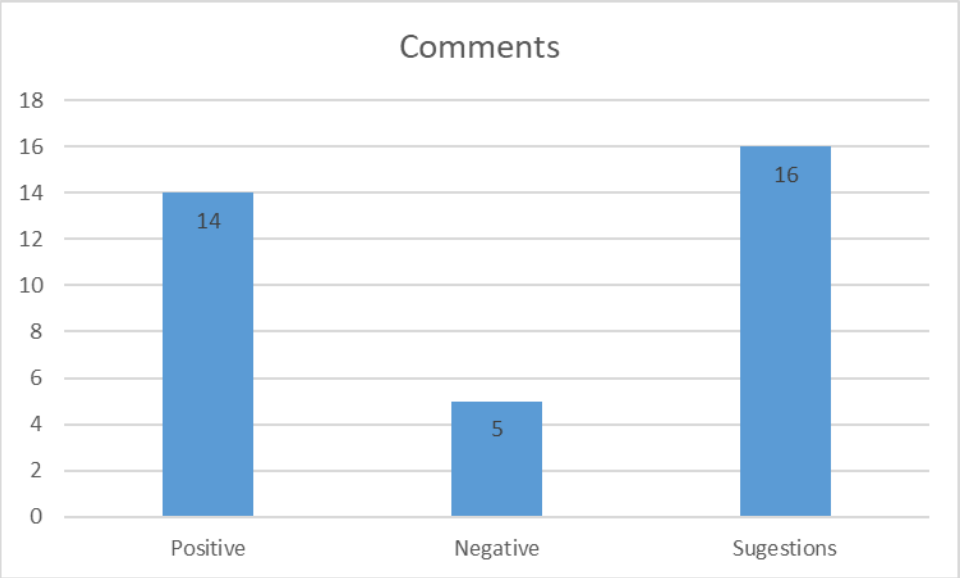
Overall, the objectives for the gamification were completed successfully, and as of right now the game is fully developed, and everything is working as intended without any kind of problems. One of the most challenging parts of the development was to fix in-game problems that do not happen on my end with little to no information to back up, in a series of trial and error mixed with code optimization in the hopes of fixing the problems.

After removing the outliers through a boxplot, the average respondent took about 239 seconds to answer the whole questionnaire (Table 11; Figure 12), and it's estimated that the average respondent took between 20 to 40 minutes to do the gamification.

From the after-game questionnaire it was given a not mandatory section for comments. The latter was divided into 5 sections, being the first information about the respondent, second section feedback about the game, and the 3 remaining sections focus on the 3 predictors of creativity, more specifically, divergent thinking (Runco et al., 2001), intrinsic motivation (Gamboa et al., 2013), and openness to experience (Vasconcellos & Hutz, 2008), based on a researched Portuguese adaptation of each the predictors scale, to later analyze the validity of the gamification items.

In total there was 35 comments (Figure 8), being them 40% (N = 14) positive, 14.29% (N = 5) negative, and the majority of 45.71% (N = 16) suggestions for improvement. The positive comments are 25.71% (N = 9) higher than the negative comments, and in brief are comments of enjoyment towards the gamification and the way it is structured (Table 15). In terms of negative comments (Table 15), were about the gamification being too long (a problem that the original SJT already had), navigability and knowing where to go and how to get there, and lastly bugs that allow the gamification to freeze indefinitely. Concerning suggestions, some of the main comments (Table 15) were about the gamification being too long, make the choices throughout the game feel more meaningful for the gameplay, for example with the implementation of different endings, different reactions from the NPCs, and probability of something happening when you choose X, lastly the last suggestion is to improve quality of gameplay through graphics, camera control, and settings to change the latter.

Figure 8
Bar chart Comments



For the polytomous items, it has a Kaiser-Meyer-Olkin (KMO) adequacy of the sample, it has a score of 0.400 (Table 16), and for the Bartlett's Test of Sphericity it has a sig of <0.001 (Table 16), which is lower than α (0.050), and so one can reject the null hypothesis and conclude that the variables are not orthogonal and the data is suitable for data reduction (Hooper, 2012; Shrestha, 2021).

In terms of Communalities (Table 17), which shows how much information it would be lost if there was a deleted item, the goal is to preserve those items with an average value of 0.500 score or higher which is considered acceptable since the sample as a value under 100 (Shrestha, 2021), which most of them are, being the exceptions the items 7_CI, 10_CI, 11_CI, and 21_CI with a close call on item 3_DTE with 0.497. Since it is compared the adequacy of the partial versus dichotomous scales, none of the items were removed.

For the Total Variance Explained (Table 18), if it was followed the Kaiser criterion method (Kaiser, 1960), that suggests that the total initial Eigenvalues should be above 1 to have as much information as a one variable, that indicates a total of 10 factors. If Hair's method was followed (Hair et al., 1998) the total variance explained (Table 18) should be greater than 60% that indicates 8 factors. Both of these number of factors are quite undesirable and will be reduced if the partial credit model options are used later on. Since there is no point of doing a deeper analysis without the reduction of factors, it will be compared the dichotomous scale to the partial.

For the dichotomous scale, the same factor analysis was conducted with a KMO and Bartlett's test, the communalities, and the total variance explained. In the KMO and Bartlett's test (Table 19), the results were not as promising as the partial credit model, the KMO adequacy of the sample had a value

of 0.389, in other words less 0.011 than the partial credit model, which is expected since the sample it's the same and the conclusions applied before can be concluded again here (Hooper, 2012; Shrestha, 2021). The Bartlett's Test of Sphericity (Table 19) in contradiction to the partial credit model was not significant with a value of 0.436, which is greater than α (0.050), and one cannot reject the null hypothesis and conclude that the variables are orthogonal, and the data is not suitable for data reduction (Hooper, 2012; Shrestha, 2021).

Even though the poor results from the Bartlett's Test of Sphericity would exclude the use of the dichotomous scale, for good measure it will be briefly analyzed the remaining tables from the factor analysis.

In terms of communalities (Table 20), most of the items have a lower value compared to the partial credit model option apart from items 5_DTE, 5_CI, 7_CI, 10_CI, 11_CI, 12_CI (with 0.001 more), and 21_CI. All of the items that were going to be removed in the partial credit model option, scored higher in this scale, and most of the items that had a higher value in the partial credit model, had a significant difference compared to the dichotomous scale, that the majority of the items on average had values around the 0.550 that are acceptable but not great (Shrestha, 2021).

Afterwards, the last table of the factor analysis is the total variance explained (Table 21) in the dichotomous scales where the values were once again slightly worsened. Interpreting through the Kaiser's method (Kaiser, 1960) mentioned before, the table indicates 11 factors, one more than the partial credit model. For the Hair's method, also stated previously, it suggests 9 factors, one more compared to the partial credit model (Hair et al., 1998).

Due to the small sample size, it was not possible to conduct an exploratory factor analysis, in this sense there was a need to calculate Cronbach's alpha, until obtaining a structure with an internal consistency greater than 0.700 (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011), obtaining a factor with 16 items (Item_1_DTE_Parcial, Item_5_DTE_Parcial, Item1_CI_Partial, Item4_CI_Partial, Item6_CI_Partial, Item7_CI_Partial, Item9_CI_Partial, Item12_CI_Partial, Item13_CI_Partial, Item14_CI_Partial, Item15_CI_Partial, Item17_CI_Partial, Item18_CI_Partial, Item20_CI_Partial, Item21_CI_Partial, Item22_CI_Partial).

Examining the factor Reliability Statistics (Table 22), this factor has an acceptable alpha of Cronbach with a value 0.726, meaning there is proper interval consistency within the items, and they are well explained (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011). Afterwards, it was evaluated if there is any room for improvement or optimization through the Item-total Statistics (Table 1), which one could check on the "Cronbach's alpha if item deleted", but there is not a single item that could have been removed in order to increase the alpha, being fully optimized.

Table 1

Item-Total Statistics of the Factor

	SCALE OF MEAN IF ITEM DELETED	*SCALE OF VARIANCE IF ITEM DELETED*	*CORRECTED ITEM-TOTAL CORRELATION*	*CRONBACH'S ALPHA IF ITEM DELETED*
ITEM1_DTE_PARTIAL	*30,771*	*40,691*	*,388*	*,705*
ITEM5_DTE_PARTIAL	*30,500*	*44,170*	*,317*	*,715*
ITEM1_CI_PARTIAL	*30,542*	*41,998*	*,395*	*,706*
ITEM4_CI_PARTIAL	*29,708*	*42,849*	*,385*	*,708*
ITEM6_CI_PARTIAL	*30,167*	*43,418*	*,214*	*,724*
ITEM7_CI_PARTIAL	*30,771*	*42,521*	*,314*	*,713*
ITEM9_CI_PARTIAL	*29,917*	*43,440*	*,270*	*,717*
ITEM12_CI_PARTIAL	*29,917*	*43,227*	*,345*	*,711*
ITEM13_CI_PARTIAL	*29,958*	*44,296*	*,205*	*,723*
ITEM14_CI_PARTIAL	*30,208*	*41,743*	*,306*	*,714*
ITEM15_CI_PARTIAL	*30,188*	*41,134*	*,367*	*,707*
ITEM17_CI_PARTIAL	*33,375*	*41,686*	*,437*	*,702*
ITEM18_CI_PARTIAL	*30,583*	*40,418*	*,326*	*,713*
ITEM20_CI_PARTIAL	*30,146*	*41,234*	*,395*	*,705*
ITEM21_CI_PARTIAL	*30,438*	*42,975*	*,235*	*,722*
ITEM22_CI_PARTIAL	*30,500*	*41,277*	*,297*	*,716*

Examining the summary report (Table 2), the sample Real Reliability of 0.69 can be initially observed, which means that the test is somewhat reproducible if one gave the same test to the same group of people (Aryadoust et al., 2021; Bond & Fox, 2015). The main difference between real reliability and model reliability is that model reliability is an upper bound of its value and the real reliability is the lower bound of its value, when the persons are filtered by measures, but for this study it will be only analyzed the lower bond which is the real reliability (Aryadoust et al., 2021).

Table 2

Summary Report Persons

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	32,3	16,0	,61	,30	1,00	.1	.99	.1
P.SD	6,8	,0	,58	,06	,26	,8	,29	,8
S.SD	6,9	,0	,58	,06	,26	,8	,29	,8
MAX.	45,0	16,0	2,31	,59	1,64	1,7	1,74	2,1
MIN.	11,0	16,0	-0,98	,26	,41	-1,7	,45	-1,6
REAL RMSE	,32	True SD	,48	Separation	1,50	Person Reliability	,69	
MODEL RMSE	,31	True SD	,49	Separation	1,59	Person Reliability	,72	
S.E. OF PERSON MEAN	= ,08							

Note: Summary of 48 measured person

Person raw score-to-measure correlation = .98

Cronbach alpha (kr-20) person raw score "test" reliability = .73 sem = 3.56

For the sample Real Separation (Table 2), it had a value of 1.50. The test has identified about 1.50 different groups, meaning that the test would not be able to differentiate 2 distinct groups of respondents in the data (Aryadoust et al., 2021; Linacre, 2019). Which goes accordingly to the data since some of the respondents are working and others are not, different genders and predominantly the same age in the with 24/25 years old. Possibly with a wider variety of respondents it could show at least a separation of 2, which can be perceived as better.

In terms of items, the item reliability (Table 3) establishes significantly higher being the real reliability 0.82, which represents great values corresponding a solid amount of different ability levels (Aryadoust et al., 2021; Bond & Fox, 2015). By analyzing the Separation field, one can immediately realize that there is 2.16 degrees of difficulty presented in the items for the real separation, that in contrary to Person, the test has found different difficulty groups of items (Aryadoust et al., 2021; Linacre, 2019).

Table 3

Summary Report Items

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	96,9	48,0	,00	,17	1,03	,0	,99	-,1
P.SD	14,7	,0	,44	,02	,26	1,6	,23	1,3
S.SD	15,2	,0	,45	,02	,27	1,6	,24	1,3
MAX.	125,0	48,0	,61	,24	1,39	2,1	1,32	1,6
MIN.	74,0	48,0	,61	,24	1,39	2,1	1,32	1,6
REAL RMSE	,18	True SD	,40	Separation	2,16	Item Reliability	,82	
MODEL RMSE	,17	True SD	,40	Separation	2,32	Item Reliability	,84	
S.E. OF ITEM MEAN	= ,11							

Note: Summary of 16 measured item

Item raw score-to-measure correlation = -.99

Umean=.0000 uscale=1.0000

In the wright's map (Table 23) one can analyze how the respondents, represented by "X", performed compared to the items delivered in terms of difficulty in a scale from -1 to 3 (Aryadoust et al., 2021; Ferreira et al., 2012).

Immediately one can examine (Table 23) that the items do not fulfill a wide range of difficulty degrees with all the items corresponding a mid-low range of difficulty, not satisfying the needs for the upper half of the respondents above the mean.

Item infits and outfits is a sensitive statistic to inliers and outliers respectively, which catches anomalies and erratic patterns within the data closer to the ability level of respondent (Aryadoust et al., 2021; Linacre, 2019). Since it does not show a normal distribution, the value of both infit/outfit mean square (MNSQ; Table 4), should be between 0.60 and 1.40 recommended for polytomous data (Aryadoust et al., 2021; Wright & Linacre, 1994), which would be productive, useful and accurate analysis where one could rely on the output for. Values outside that range in MNSQ should be further analyzed, and its accuracy and reliability diminish depending on how far it is from the range.

Table 4

Item Statistics: Measure order

Entry Number	*Total Score*	*Total Count*	*Measure*	*Model S.E*	*Infit*		*Outfit*		*PTMEASUR-AL*		*EXACT OBS%*	*MATCH EXP%*	*ITEM*
					MNSQ	*ZSTD*	*MNSQ*	*ZSTD*	*CORR.*	*EXP.*			
1	*74*	*48*	*,61*	*,15*	*,92*	*-,5*	*,90*	*-,5*	*,54*	*,46*	*39,6*	*35,0*	*Item1_DTE*
6	*74*	*48*	*,61*	*,15*	*,81*	*-1.2*	*,86*	*-,8*	*,40*	*,46*	*22,9*	*35,0*	*Item7_C*
13	*83*	*48*	*,40*	*,15*	*1,39*	*2,1*	*1,32*	*1,6*	*,48*	*,46*	*33,3*	*37,3*	*Item18_CI*
3	*85*	*48*	*,35*	*,16*	*,66*	*-2,2*	*,64*	*-2,2*	*,50*	*,46*	*45,8*	*37,3*	*Item1_C*
2	*87*	*48*	*,30*	*,16*	*,48*	*-3,8*	*,52*	*-3,0*	*,35*	*,46*	*68,8*	*37,8*	*Item5_DTE*
16	*87*	*48*	*,30*	*,16*	*1,30*	*1,6*	*1,24*	*1,2*	*,45*	*,46*	*27,1*	*37,8*	*Item22_C*
15	*90*	*48*	*,23*	*,16*	*1,11*	*,7*	*1,09*	*,5*	*,39*	*,45*	*31,3*	*37,9*	*Item21_C*
12	*93*	*48*	*,15*	*,16*	*,65*	*-2,2*	*,62*	*-2,1*	*,54*	*,45*	*60,4*	*37,6*	*Item17_C*
10	*101*	*48*	*-,07*	*,17*	*1,30*	*1,5*	*1,17*	*,8*	*,43*	*,44*	*29,2*	*40,6*	*Item14_C*
11	*102*	*48*	*-,12*	*,17*	*1,20*	*1,0*	*1,17*	*,8*	*,49*	*,43*	*20,8*	*40,6*	*Item15_C*
5	*103*	*48*	*-,15*	*,17*	*1,25*	*1,2*	*1,24*	*1,1*	*,35*	*,43*	*33,3*	*40,6*	*Item6_C*
14	*104*	*48*	*-,45*	*,17*	*1,03*	*,2*	*1,12*	*,6*	*,45*	*,43*	*27,1*	*40,5*	*Item20_CI*
9	*113*	*48*	*-,45*	*,19*	*1,13*	*,6*	*1,05*	*,3*	*,33*	*,40*	*47,9*	*51,0*	*Item13_CI*
7	*115*	*48*	*-,53*	*,20*	*1,14*	*,6*	*1,07*	*,4*	*,37*	*,39*	*47,9*	*51,5*	*Item9_CI*
8	*115*	*48*	*-,53*	*,20*	*1,20*	*,8*	*,98*	*,0*	*,44*	*,39*	*68,8*	*51,5*	*Item12_CI*
4	*125*	*48*	*-,99*	*,24*	*1,20*	*,8*	*,98*	*,0*	*,44*	*,34*	*66,7*	*66,2*	*Item4_C*
MEAN	*96.9*	*48,0*	*.00*	*.17*	*1.03*	*.0*	*.99*	*-.1*			*41.9*	*42.4*	
P.SD	*14.7*	*.0*	*.44*	*.12*	*.26*	*1.6*	*.23*	*1.3*			*16.1*	*8.2*	

For ITEM18_CI (Table 4) it goes nearly to the limit of the desired value, meaning that there is a fit but not on an overwhelming amount.

For items 22_CI, 6_CI and 14_CI (Table 4), they have a somewhat high value but still within the limits. Even though some of the items are under the range mentioned before the only one that stands out is the ITEM5_DT with a 0.48 infit mean square (Table 4), being an underfit and should be explored. This particular case (Table 24) did not have a single respondent choosing the second worst answer and an overwhelming amount of the second-best option with 88% chosen rate and a single person opted for the best answer. After analyzing the item, the best answer can somehow be perceived as being

self-centered that could have thrown of the respondents to the second-best option that is more team centered.

In terms of outfit (Table 4), it leveraged to get overall lower values than the infits, being the only outstanding case the ITEM5_DT with a mean square of 0.52 which is slightly below the desired range and considered an underfit. Again, having what was mentioned previously for the infit in mind, this could be the work of catching anomalies that are far away from the item difficulty. For the remaining items (Table 4), all of them showed values within the range defined previously, having good outfit statistics.

In regard to the correlation between the items (Table 4) in the column PTMEASUR-ALL in CORR., all did as expect varying from 0.33 to 0.54 with having widespread values and none of the items were far off the others, and so passing one of the quality insurances of the model (Aryadoust et al., 2021).

In the Item statistics table (Table 4), since partial scores are the ones being worked on, the scores diverge between 0, for the worst answer, to 3 with the best answer. This is important for the total score column where the most difficult items will show the lowest number since its sum will be lower because of the worst answers, giving less “points”. The Items, ITEM1_DT and ITEM7_CI are the most difficult items, as also was shown on the wright map (Table 23), being out of the standard deviation zone, with fewer best scores (Table 24) with 19% and 17% 3-point answers respectively, and most of the answers were the second best with 44% and 35% respectively.

One odd case in the items is the ITEM5_DT (Table 24), with 88% with the second-best answer and not a single respondent choosing the second worst answer. After this case, all below items in the table show that most of the respondents choose the best answer the most, with an overwhelming amount above 50% on 8 Items, more specifically all the items below ITEM14_CI.

According to the infit and outfit standardized deviation (ZSTD; Table 4), since the sample used is below 250 the recommended range is between -1.96 and 1.96 (Aryadoust et al., 2021; Linacre, 2002). Also due to the low sample, the ZSTD often becomes more useful, because the data cannot get as inflated then with a larger sample above 250 (Aryadoust et al., 2021). The values of the items, for the majority, are within the range mentioned, but in the cases of the items 18_CI, 1_CI, 5_DT and 17_CI, they go outside of the range. For the item 18_CI, it has a good outfit value but in the infit it goes slightly over the limit with a value of 2.10, which is desired but still not very problematic. For Item 1_CI, both the infit and outfit values go slightly below the wanted values with a value of -2.20, that leads to the same conclusion as the previous item. For item 5_DT, in contrast to the other items, it goes below the limit for a large amount indicating infit and outfit values of -3.80 and -3.00 respectively, meaning that the item does not fit perfectly in the model (Aryadoust et al., 2021), and the item should be explored and potentially changed or removed from the model. Lastly, for item 17_CI has similar values has item 1_CI, with an infit and outfit of -2.20 and -2.10 respectively that can also be deduced the same way.

In terms of measure (Table 4), it shows the severity or difficulty of the items which do not deviate a lot from each other, there so representing similar levels of difficulty. The 2 hardest items with a measure of 0.61 were ITEM1_DT and ITEM7_CI with a 0.21 difference from the third hardest item. One irregularity that distinguishes itself from the others is ITEM4_CI with a very low degree of measure being easier than the other items with a -0.99, being the closest -0.53 and it translates in the total score.

The standard error of measurement or Model S.E in the table (Table 4), is the margin of error of the measurement of the items in this table and was shown on the wright map, which one can interpret the precision of the items. In this case, the model S.E., has a standard error of 0.15 to 0.24, that are quite low, and a good level of precision (Aryadoust et al., 2021).

In terms of reliability, all of the predictors had different spectrums of reliability according to Cronbach's alpha. Starting at the lowest end, divergent thinking composed of 6 items, in the reliability analysis (Table 25) had a Cronbach's alpha value of 0.582 which is considered poor (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011). According to Item-Total Statistics (Table 26), in the "Cronbach's Alpha if item deleted" tab, this predictor cannot be improved.

Thereafter, factor OE with 3 items, in the reliability analysis (table 27) has a higher Cronbach's value of 0.811 that can be interpreted has a good internal consistency (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011). This predictor is fully improved and by verifying Item-Total Statistics (table 28) it can be concluded the latter with no room for improvement.

Finally, the factor IM with 4 items, is the strongest factor according to Reliability Statistics (table 29), with a Cronbach's alpha of 0.932 that indicates excellent internal consistency (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011). The Item-Total Statistics (table 30), indicates that the predictor does not get any higher with the removal of other items, thus being the optimal solution.

Before examining the correlations, it will be discussed the descriptive statistics (Table 5) of all the factors stated previously, taking into account the literature (Fisher & Marshall, 2009; Haden, 2019; Rendón-Macías et al., 2016).

Table 5

Descriptive Statistics of the factors + Correlation matrix between factors

	COMP_DT	*COMP_IM*	*COMP_OE*	*COMP_FACTOR*
MEAN	*3,594*	*5,260*	*6,063*	*2,020*
STD. DEVIATION	*,561*	*1,325*	*,906*	*,430*
N	*48*	*48*	*48*	*48*

*Note: * Correlation is significant at the 0.05 level (1-tailed)*

*** Correlation is significant at the 0.01 level (1-tailed)*

The items of divergent thinking have a Likert scale ranging 5 possible answers from “strongly disagree” to “strongly agree”, and as one can observe in the descriptive statistics table (Table 5) Comp_DT had a mean of approximately 3.594 (SD ≈ 0.561). For the items of Intrinsic Motivation and Openness to Experience also have a Likert scale, however it ranges 7 possible options from “strongly disagree” to “strongly agree”.

Regarding Comp_IM had a mean of approximately 5.260 (Table 5), where the average answer is “somewhat agree” with a standard deviation about 1.325 widening the span of answers to “neither agree or disagree” and “agree”. As for, Comp_OE it had a mean of approximately 6.063 (SD ≈ 0.906; Table 5), very central on the option “agree”, and with standard deviation ranging to “somewhat agree” and “strongly agree”.

To conclude, the factor, as mentioned before, has a partial credit model option ranging from 0 to 3, being 0 the worst answer and 3 the best answer. Comp_Factor had a mean of approximately 2.020 (SD ≈ 0.430; Table 5), thus the respondents on average choose the second-best answer most often, which could indicate that the participants performed above the average.

Moreover, the correlation between all the predictors factors mentioned before and Comp_Factor in a correlation matrix will be verified. In the correlation matrix (Table 6), it can be examined that the Comp_Factor in terms of Pearson Correlation, has a weak positive correlation with Comp_DT of 0.208, nonexistence relationship with Comp_IM of 0.005, and a moderate positive relationship with Comp_OE of 0.411 (Dancey & Reidy, 2007; Schober et al., 2018; Taylor, 1990).

Table 6

Descriptive Statistics of the factors + Correlation matrix between factors

		COMP_DT	COMP_IM	COMP_OE	COMP_FACTOR
COMP_DT	Pearson Correlation	1	,108	,263*	,208
	Sig. (1-tailed)		,232	,036	,078
	N	48	48	48	48
COMP_IM	Pearson Correlation	,108	1	-,023	,005
	Sig. (1-tailed)	,232		,439	,485
	N	48	48	48	48
COMP_OE	Pearson Correlation	,263*	-,263	1	,411**
	Sig. (1-tailed)	,036	,439		,002
	N	48	48	48	48
COMP_FACTOR	Pearson Correlation	,208	,005	,411**	1
	Sig. (1-tailed)	,078	,485	,002	
	N	48	48	48	48
MEAN		3,594	5,260	6,063	2,020
STD. DEVIATION		,561	1,325	,906	,430
N		48	48	48	48

*Note: * Correlation is significant at the 0.05 level (1-tailed)*
*** Correlation is significant at the 0.01 level (1-tailed)*

Additionally in the correlation matrix (Table 6), considering an α of 0.050, supports the correlation where the only significant correlation, where the null hypothesis can be rejected is between the factors Comp_Factor with Comp_OE with a value of 0.002 (Perreault et al., 2022). All the other factors did not have a significant correlation with Comp_Factor and one can not reject the null hypothesis.

3.1. Comparation with previous studies

In a brief analysis, comparing the gamification factors obtained and the factors developed by the researcher Trigo (2021), the Factor developed through the gamification has 16 items and a Cronbach's alpha of 0.728 (Table 22), on the other hand for the traditional SJT "Fator 1" has 4 items and a Cronbach's alpha of 0.620, "Fator 2" with 6 items has a Cronbach's alpha of 0.650, and "Fator 3" has 3 items and a Cronbach's alpha of 0.570, displaying a greater reliability for the factor developed in the gamification (Trigo, 2021). The traditional SJT factors, when reproduced with the gamification data, showed low reliability with "Fator 1", "Fator 2", and "Fator 3", having a Cronbach's alpha of 0.190 (Table 7), 0.558 (Table 8) and 0.408 (Table 9) respectively (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011). Additionally, in a factor analysis "Fator 1" did not have a significant score in the Bartlett's test, has low communalities and a total variance indicating 2 factors, and the same can be applied to "Fator 3", with the exception of the total variance indicating only 1

factor, in terms of “Fator 2”, has a significant value of Bartlett’s, still low communalities and a total variance explained indicating 2 factors (Hooper, 2012; Shrestha, 2021). However, nothing can be concluded due to the sample differences.

Table 7

“Fator 1” Reliability

CRONBACH’S ALPHA	***N OF ITEMS***
,190	*4*

Table 8

“Fator 2” Reliability

CRONBACH’S ALPHA	***N OF ITEMS***
,558	*6*

Table 9

“Fator 3” Reliability

CRONBACH’S ALPHA	***N OF ITEMS***
,408	*3*

CHAPTER 4

4. Discussion

Throughout the development of this study, it has been followed the Hinkin guidelines for scale development (Hinkin 1995 and 1998). These guidelines in a simplified manner consist of 7 major steps. Item generation, content adequacy, questionnaire administration, factor analysis, internal consistency assessment, construct validity and replication, all of which (except replication) were applied and adherent along the development of the study.

4.1. The Game

As one arrives on the floor, where the whole gameplay will be on, the player exits the elevator and has an arrow above the characters model, indicating where the nearest interaction is, and on the top left two icons (Figure 4) indicating how the game can be paused with a tab containing the games instructions, as well as an icon for consulting a map of the floor with the location of the tasks marked. After interacting with the first non-playable character (NPC) and completing its tasks, it will unlock a mission on the top right (Figure 4) indicating how many NPC's/objects they will need to interact and complete the tasks to finish the game. Each interaction does not have a fixed number of tasks, taking some interactions longer than others, and the tasks are the items from the traditional SJT converted to dialogue (Figure 5), where it has a question with 4 possible answers that the respondent has to choose, and at the end of the interaction the arrow will no longer point at that position and the mission counter will increase. This interactions in order to be diversified and appealing are taken in different contexts, for instance in a job interview (Figure 1), in a remote call (Figure 3), and even a few with objects without tasks (Figure 2) that does not increment the mission counter and are there to give a more immersive experience. After completing the mission, the arrow will now point at the other elevator, that now will be unlocked so the player can leave, and an end-game screen will be displayed indicating that the respondent completed the game.

When completing the gamification, the population had to put some discrete information to relate the gameplay to the questionnaire and after submitting, it transitions into the game scene and the game starts. Inside the game with a theme of an office environment, the player starts in the elevator, where the respondent gets a brief introduction of the company they are working in, the character's position within the company, its aspirations and goals, as well as a recap of some of the tasks they will do later.

4.2. Theorical contributions

It was recoded the items results into partial credit model and dichotomous Rasch model to enable data validation, and tagged has “itemX_X_Partial” and “itemX_X_Dichotomous” respectively. In terms of dichotomous items, it was recorded in a binary way where the best answer was considered 1, and the remaining 0. For the partial items, since all of the items had 4 possible solutions, it went from 3 has the highest score, to 0 has the lowest score, based on answers best to worst scores on the traditional SJT developed by Beatriz Trigo (Trigo, 2021). In order to decide on which of the 2 scores models is more suitable for this study, both of the scores went through a factor analysis with all of the items, where it was compared the results and evaluated which is more promising for an upcoming analysis. First, a factor analysis was done on all of the items with a partial score that displayed immediately good results. The partial credit model scale reveals to be a more promising and consistent than the dichotomous Rasch model scale in this study, with the data being more suitable for data reduction, the items with higher communalities, where each item overall contains more information, and allegedly less factors into consideration. This difference between both scales, could be explained by the items containing 4 possible answers which the partial credit model options translates well into scores from 0 to 3, where the dichotomous scale recoded into a binary number of responses, in order words wrong or right, shifts the answers results. Therefore, the items that will be used from now on, are from the partial credit model, discarding the dichotomous Rasch model.

As for the Kaiser-Meyer-Olkin (KMO) adequacy of the sample, it has a score of 0.400 (Table 16) indicating that the sample used is not adequate, as expected due to the low number of responses per item in this study (Hooper, 2012; Shrestha, 2021). For the Bartlett's Test of Sphericity, it’s a hypothesis test that will tell whether the variables are orthogonal and there is enough correlation between the items, and with a sig of <0.001 (Table 16), which is lower than α (0.05), and so one can reject the null hypothesis and conclude that the variables are not orthogonal and the data is suitable for data reduction (Hooper, 2012; Shrestha, 2021).

However, due to a low sample, it was not possible to conduct an exploratory factor analysis, and so there was a necessity to calculate the Cronbach alpha until obtaining a construct with an internal consistency over 0.700 (Bland & Altman, 1997; Bujang et al., 2018; Gliem & Gliem, 2003; Tavakol & Dennick, 2011) creating a Factor with 16 items and a Cronbach’s alpha of 0.726 that was verified in the Item-Total Statistics that it could not be more optimized (Table 22 and 14).

Afterwards, an Item Response Theory (IRT) was adopted as an alternative to the Classical Test Theory (CTT) to check the factor unidimensionality. One of the limitations of CTT that IRT overcomes is the limitation that data and parameters are sample and test dependent, that due to the low sample presented in this study, can be a valuable alternative. Another limitation that the literature defends is

that it is presumed that the values across the range of the test are uniform for measurement precision, which is incorrect for various tests. Research has also proved that IRT, in some cases, provided reliable results where the CTT does not, for example in human memory (Ferreira et al., 2011).

In order to test the unidimensionality of the factor, the IRT was used, that despite the Classical Test Theory (CTT), this approach is not sample sensitive to get a more creditable information (Ferreira et al., 2011/2012). It was conducted the IRT analysis in the WINSTEPS application, where the sample was labeled "Person" for clarification throughout the remaining analysis of the IRT.

Initially it was analyzed the factor's reliability and separation (Table 3), where the items had a good real reliability of 0.82 corresponding to solid amount of ability levels. In the separation field (Table 3), that indicates the degrees of difficulty presented on the model, the items had a real separation of 2.16 which is solid but could be improved and should be more diversified with items that rather spreads the difficulty levels ranges wider.

In terms of Persons, even though the sample is low and the IRT is not as dependent on sample, is has a somewhat acceptable reliability of 0.69, and a real separation of 1.50 indicating that the test would not be able to differentiate two different groups, which goes to some extent accordingly to the data, since some of the respondents are working and others are not, different genders and predominantly the same age in the with 24/25 years old. Possibly with a wider variety of respondents it could show at least a separation of 2, which can be perceived as better.

Examining the wright map (Table 23), one has confirmation on what was speculated before with the items being concentrated in a mid-low field of difficulty, not meeting the demands of respondents with higher scores.

However, the items are well distributed on the range, they do not leave large gaps, or rather jumps, in difficulty that could ramp up the complexity and leave room to some degrees of difficulty to be fulfilled. Only 6 of the items are beyond the standard deviation range, 4 of them being lower the value and 2 over the value, and of those 6 only 1 item is outer the double standard deviation value, in the lower end, leaving a large gap to the next item. In terms of the sample (Table 23), in other words people have a normal distribution with the mode being slightly over the mean value and has 1 inlier and 1 outlier outside of the double standard deviation zone. Overall, the person's distribution is over the item's values indicating, as stated previously, that the degree of difficulty of the items are lower than the population scored and should target a wider range of complexity to get a deeper analysis and in gamification wise bring more engagement and an enjoyable experience.

Concerning items infit and outfit statistics (Table 4), due to the use of polytomous data, it was used the conventional fit range of the mean square between 0.60 to 1.40 (Aryadoust et al., 2021; Wright et al., 1994) and with a sample of less than 250, the recommended value of standardized deviation (ZSTD) is between -1.96 and 1.96 (Aryadoust et al., 2021). In terms of infits and outfits mean

square (Table 4), most of the items are within the range stipulated in the latter with the exception of Item5_DT that goes under the value in both infit and outfit and needs to be further examined.

The Item18_CI it's within the desired values, but not far from the limits. This could be explained (Table 24) due to the two most answers being the best one with 40%, and the worst one with 29%, going into extremely opposite directions of answers and the phrasing of the answers forward should be reconsidered, even though its within fit ranges. Another point that could explain this high value is respondents that overall did not perform as well on other answers often choose the best answer.
Also, the Items 22_CI, 6_CI and 14_CI present high values that are not concerning. This case is different from the latter with the second-best option being ahead (Table 24) from the worst option by 2% for ITEM22_CI, and 10% for the ITEM14_CI with a more distributed range of answers. For ITEM6_CI (Table 4) it's on another case with values potentially less problematic, being the second most chosen the answer the second worst response with 25% (Table 24).

Further, while exploring the Item5_DT it was out of bounds, it was found that the item has an irregularity span of answers (Table 24), being one of the options never chosen and the second-best answer has been selected it an overwhelming amount of 88% of the respondents, the item should be reviewed or rephrased to mitigate the disparities. Exploring the standardized deviation besides (Table 4), Item5_DT that goes far from the desired range, only 3 items deviated slightly of desired value with the farthest being 0.24 off and so being a concern but not a major problem in term of item fits. After analyzing the item, the best answer can somehow be perceived as being self-centered that could have thrown of the respondents to the second-best option that is more team centered.

The outfits in general presented lower values than the infits, this could be the work of catching anomalies that are far away from the item difficulty.

While exploring the items, after ITEM5_DT (Table 24), all below items in the table show that most of the respondents choose the best answer the most, with an overwhelming amount above 50% on 8 Items, more specifically all the items below ITEM14_CI.

In sum, the model has shown potential with good values in terms of item fit, and reliability, however the item 5_DT, should be closely analyzed, changed or even removed with underfit values in various aspects, and the degrees of difficulty should be widened and well spread.

In sum, the model has shown potential with good values in terms of item fit, and reliability, however the item 5_DT, should be closely analyzed, changed or even removed with underfit values in various aspects, and the degrees of difficulty should be widened and well spread.

After the unidimensionality of the factor was tested, in order to verify the factors validity, it is needed to correlate them to the creativity major predictors, being them as mentioned previously, Divergent Thinking (Plucker et al., et al., 2006; Runco & Chand, 1995; Runco et al., 2001; Sen, 2016, Intrinsic Motivation (Chan et al., 2018; Georgiou & Lievens, 2022; Runco & Chand, 1995; Salikutluk et

al., 2019), and Openness to Experience (McCrae, 1987; Tan et al., 2019; Woo et al., 2013). However, before the correlations one needs to check each predictor reliability in the questionnaire delivered to each respondent to evaluate if there is a solid foundation behind the items, even though these are researched proven tests (Gamboa et al., 2013; Runco et al., 2001; Vasconcellos & Hutz, 2008).

In brief, after doing a factor analysis on the predictors there was a data reduction in every predictor. For divergent thinking, the item Q19 was removed, for intrinsic motivation the item Q29 was deleted and in Openness to Experience items Q10, Q11, Q12, Q13, Q14 and Q15 were deleted, additionally the factors were named Comp_DT, Comp_IM, and Comp_OE respectively.

Afterwards it was done a factor analysis onto the Creativity predictors divergent thinking (Runco et al., 2001), intrinsic motivation (Gamboa et al., 2013) and openness to experience (Vasconcellos & Hutz, 2008) with the items from the questionnaire that was adapted from researched Portuguese adaption of each one, to later check the factors validity for the creation of a model. After the removal of some redundant items, it was generated Comp_DT, Comp_IM, Comp_OE, and the Comp_Factor as the principal factor.

Moreover, it was analyzed the predictors and factor descriptive statistics, where it was able to draw the following conclusions. For the divergent thinking with a mean of approximately 3.59 (SD ≈ 0.56) in a scale of 1 to 5, that could indicate that the average respondent often would select the options “neither agree or disagree” or “somewhat agree”, identifying themselves as someone who somewhat or neutrally thinks outside the box or has innovative ideas. In terms of intrinsic motivation, the mean is approximately 5.26 (SD ≈ 1.32) in a scale of 1 to 7, where the average answer is “somewhat agree” with a standard deviation about 1.32 widening the span of answers to “neither agree or disagree” and “agree”, meaning that the average person felt somewhat motivated and engaged by the gamification. Next, the openness to experience with a mean of approximately 6.06 (SD ≈ 0.91) in a scale of 1 to 7, very central on the option “agree”, and with standard deviation ranging to “somewhat agree” and “strongly agree”, with the purpose of the typical respondent identifying themselves as someone open to experience new things. Finally, the factor with a mean of approximately 2.02 (SD ≈ 0.43) in a scale of 0 to 3, thus the respondents on average choose the second-best answer most often, which could indicate that the participants performed above the average.

Furthermore, it was verified the factors validity through the correlation with the predictors, and initially it was analyzed (Table 6) that the Comp_Factor had Pearsons’s correlations of low correlation with Comp_DT and Comp_IM, however having a moderate correlation with factors Comp_OE (Dancey & Reidy, 2007; Schober et al., 2018; Taylor, 1990). In terms of significance (Table 6), the factor only had a significant correlation with Comp_OE, having a nonsignificant correlation with the other predictors. The researchers Schönbrodt & Perugini (2013) has proven, that with low samples the correlations are not stable yet, and so with the increase of sample the correlations will deviate to

their true value, which could increase the factors correlations, additionally nothing can be concluded due to a type II error in the sample size (Akobeng, 2016).

In conclusion, the model has shown potential assessing creativity with good item fit, reliability values, internal consistency and psychometric properties, but there is still room for improvement and some of the items used in the model could be reviewed, as well as a proper sample.

In comparation to the traditional SJT, the gamification presented different results in various aspects. For instance, the factors developed in both ends were composed of different items and delivered very different results. For the factor development Trigo (2021), tried to generate factors with the DT items and CI items separately in 2 different studies, which was concluded unsuccessfully for the DT items and create 3 different factors for the CI items, while on my end with the gamification it was used a different approach where the two types of items were combined to create one factor to explain creativity.

In comparison with the traditional SJT (Trigo, 2021) where there were 12 comments out of 26 (46.25%) from the respondents, that mentioned the tool is either “too long”, “boring”, “exhausting”, while in the gamification 14.28% (N = 5, out of 35; Figure 8), of the comments were negative with the majority complaining of being “too long” (Table 15). Through the complement of the intrinsic motivation descriptive statistics mentioned previously, the average responses were “somewhat agreed” with feelings of motivation towards the gamification that might indicate the gamification was more engaging and immersive thus having fewer negative comments than the traditional SJT, yet it is not possible to conclude with only this amount of information (Chan et al., 2018; Georgiou & Lievens, 2022; Gkorezis et al., 2020; Salikutluk et al., 2019).

The development of this gamified SJT could contribute to SJT’s theory by showing how a gamified assessment can be developed through my steps and my setbacks, having in mind that there is room for improvement. Also, the validation of the gamification showed potential to this of this format as an alternative to the written, video-format and other types of SJT and should be further explored (Gkorezis et al., 2020; Lievens & Chapman, 2019).

When one speaks of creativity, this research shows the theoretical contributions using divergent thinking (Plucker et al., et al., 2006; Runco & Chand, 1995; Runco et al., 2001; Sen, 2016), intrinsic motivation (Chan et al., 2018; Georgiou & Lievens, 2022; Runco & Chand, 1995; Salikutluk et al., 2019), and openness to experience (McCrae, 1987; Tan et al., 2019; Woo et al., 2013) to validate the latter. The scales contemplated the different items used and can be used for further research in a Portuguese population, even though nothing could be concluded due to a type II error (Akobeng, 2016; Gamboa et al., 2013; Runco et al., 2001; Vasconcellos & Hutz, 2008). This research also contributes to the development of creative models.

Despite its reputation, the use of IRT to validate unidimensionality and how to interpret the results can be valuable for research, especially in cases of inadequate samples that bring the most value to this theory (Aryadoust et al., 2021; Ferreira et al., 2011/2012).

4.3. Practical contributions

Applied to practical examples, this study contributed to better understanding the importance of creativity in recruitment and overall business and due to its variability, it can be very challenging to spot over traditional selection and recruitment processes (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020). The use of gamification with SJT's to evaluate a psychometric property of a candidate could lead to more positive feelings towards the recruiter, and the overall recruitment and selection experience, enhancing the employer branding and how the company is perceived (Chan et al., 2018; Georgiou & Lievens., 2022; Gkorezis et al., 2020; Oostrom et al., 2019; Salikutluk et al., 2019). Becoming more enjoyable for the recruiter to select, retain the best candidates for a better sustainable competitive advantage to the business (Anderson et al., 2014; Elidemir et al., 2020; Klepić et al., 2020).

4.4. Limitations and direction for future research

Like any piece of research, this one had its challenges and limitations. Due to the original questionnaire being too long, in game bugs, and there is not a mobile version of the game, the sample is smaller than initially expected despite the efforts. Overall, between March 5th and May 27th, the game had a total of 204 views and 149 browser plays but with only 48 completed attempts, due to the latter reasons stated in incomplete questionnaires since the game has not been finished to complete the data. Additionally, due to the gamification being developed by a single individual, it used large amounts of time and resources that could have been directed for example, to gather a larger sample. In future research, it would be recommended to use a wider sample of 5 to 10 respondents per item (Akobeng, 2016).

Another limitation regarding the sample, is that initially it was stipulated to use the same sample that was used on the traditional SJT, but unfortunately no one of the participants from the original sample replied or did the gamification, that was originally targeted to IT workers and in this version a part of respondents were students and the majority of the working population, works in the service sector. Additionally, due to the latter it was not possible to survey the population of the traditional SJT to find the main differences between it and the gamification, as well as if it was more enjoyable, more engaging, more immersive and so on, that would have been valuable information for this study.

To host the gamification online, so it would be available for the respondents was difficult due to the most convenient options where too expensive, or with a high degree of complexity to adopted.

The free, or more affordable versions did not support the original version of the game, and therefore, it was necessary to convert the meshes into smaller portions which affected negatively the visual quality of the game and took a couple of days with a large amount of meshes. Moreover, on the sample used, when bugs occurred, at the initial stage the gamified SJT would crash the game and leave a pitch-black image, which was problematic since the respondents would refuse to replay the gamification affecting very negatively the sample and it did not happen when played locally making hard to fix them. Overall, the items used in this study present an adequate fit with somewhat good results, in future studies, its recommended to explore the items that do not have good infit and outfit statistics. As well as increasing several levels of difficulty due to the items present to evaluate similar degrees of difficulty, limiting the respondent's ability to feel challenged and distinguish themselves from others.

The personalization of the assets, all characters were based on four models of a cowboy themed where clothing, textures and colors, hairstyle/color, eye color, facial hair, skin tone had to be manipulated so the cast would look more business-like dressed in business attire. Creating 15 unique and diversified characters (Figure 1, 4 and 5). In terms of the gamification itself and giving a more game-a-like experience, there are a few recommendations that were not implemented due to scarce time or high levels of complexity, that could be explored and improved in future research.

The creation of 4 different main characters which the respondent can choose the one they identify the most to create a more immersive experience. Second, the addition of voice lines to the characters to get an overall better experience. The addition of a "report a bug" button or text box, to help the developer to understand where the problem is, as well as deeper settings options in terms of sound, graphics, controls and so on.

Having into consideration the comments (Table 15), there should be caution in the future with the gamification not being too long to conclude, and in cases like this study where the gamification is based on a question-answer basis, make the decisions more impactful and meaningful throughout the game to keep the respondents engaged and immersed by feeling that the option they will pick will have consequences.

5. Conclusion

What started as a comparation between traditional SJT and gamification, quickly became a model development, and the differences in the development and in the sample did not permit us to come to a fair conclusion between them. However, it was possible to have some interesting comparations among the two, and this dissertation gathered valuable information on the development of a gamification and creativity model, as well as in SJT's, creativity, and selection and recruitment that could help in the future developments in these areas.

6. References

Akobeng, A. K. (2016). Understanding type I and type II errors, statistical power and sample size. *Acta Paediatrica, 105*(6), 605-609.

Anderson, N., Potočnik, K., & Zhou, J. (2014). Innovation and Creativity in Organizations. A state-of-the-science review, prospective commentary, and guiding framework. *Journal of Management, 40*(5), 1297–1333. https://doi.org/10.1177/0149206314527128

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

App, S., Merk, J., & Büttgen, M. (2012). Employer Branding: Sustainable HRM as a Competitive Advantage in the Market for High-Quality Employees. *Management Revue*, 23(3), 262–278. http://www.jstor.org/stable/41783721

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40.

Auger, P., & Woodman, R. W. (2016). Creativity and Intrinsic Motivation. *The Journal of Applied Behavioral Science, 52*(3), 342–366. https://doi.org/10.1177/0021886316656973

Batey, M. D. (2007). *A psychometric investigation of everyday creativity*. University of London, University College London (United Kingdom).

Bennett, N., & Lemoine, J. (2014). What VUCA really means for you. *Harvard business review, 92*(1/2).

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj, 314*(7080), 572.

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model Fundamental Measurement in the Human Sciences (3rd ed.). *Mahwah, NJ L. Erlbaum.*

Brown, M. I., Speer, A. B., Tenbrink, A.P., & Chabris, C. F. (2022). Using game-like animations of geometric shapes to simulate social interactions: An evaluation of group score differences.

International Journal of Selection and Assessment, 30, 167–181. https://doi.org/10.1111/ijsa.12375

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. *The Malaysian journal of medical sciences: MJMS, 25*(6), 85.

Chan, E., Nah, F. F. H., Liu, Q., & Lu, Z. (2018). Effect of gamification on intrinsic motivation. In *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 5* (pp. 445-454). Springer International Publishing.

Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2018). Reopening Openness to Experience: *A Network Analysis of Four Openness to Experience Inventories. Journal of Personality Assessment, 101*(6), 574–588. https://doi.org/10.1080/00223891.2018.1467428

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Ellison, L.J., Johnson, T.M., Tomczak, D., Siemson, A. and Gonzalez, M.F. (2020), "Game on! Exploring reactions to game-based selection assessments", *Journal of Managerial Psychology*, Vol. 35 No. 4, pp. 241-254.

Ferreira, A. I., Almeida, L. S., & Prieto, G. (2011). The role of processes and contents in human memory: An item response theory approach. *Journal of Cognitive Psychology, 23*(7), 873-885.

Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian critical care*, 22(2), 93-97.

Gamboa, V., Valadas, S. T., & Paixão, O. (2013). Validação da versão portuguesa da Situational Motivation Scale (SIMS) em contextos académicos. In *Atas do XII Congresso Galego-Português de Psicopedagogia, 11-13 de setembro de 2013* (pp. 4868-4882). Universidade do Minho.

García, L. F., Aluja, A., García, Ó., & Cuevas, L. (2005). Is openness to experience an independent

personality dimension? Convergent and discriminant validity of the openness domain and its NEO-PI-R facets. *Journal of Individual Differences, 26*(3), 132-138.

Georgiou, K., & Lievens, F. (2022). Gamifying an assessment method: what signals are organizations sending to applicants? *Journal of Managerial Psychology, 37*(6), 559–574. https://doi.org/10.1108/JMP-12-2020-0653

Gillet, N., Vallerand, R. J., Lafreniere, M. A. K., & Bureau, J. S. (2013). The mediating role of positive and negative affect in the situational motivation-performance relationship. *Motivation and Emotion, 37*, 465-479.

Gkorezis, P., Georgiou, K., Nikolaou, I. and Kyriazati, A. (2020), "Gamified or situational judgement test? A moderated mediation model of recommendation intentions", *European Journal of Work and Organizational Psychology, 30*(2), pp. 240-250.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.

Govendo, J. A. (2005). Workforce, diversity and corporate creativity. *Handbook of Business Strategy*, *6*(1), 213-218.

Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). *Motivation and emotion, 24,* 175-213.

Haden, P. (2019). Descriptive statistics. *The Cambridge Handbook of Computing Education Research*, 102-132.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate data analysis. Upper Saddle River, New Jersey: Prentice Hall.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of management, 21*(5), 967-988.

Hinkin, T. R. (1998). A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires. *Organizational Research Methods, 1*(1), 104–121. https://doi.org/10.1177/109442819800100106

Hocevar, D. (1981). Measurement of Creativity: Review and Critique. *Journal of Personality Assessment, 45*(5), 450–464. https://doi.org/10.1207/s15327752jpa4505_1

Holtrop, D., Born, M. Ph., de Vries, A., & de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences, 68*, 234–240. https://doi.org/10.1016/j.paid.2014.04.029

Hooper, D. (2012). Exploratory factor analysis.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Klepić, I., Mabić, M., & Madžar, D. (2020, August). Recruitment and selection of human resources and organizational creativity in small and medium-sized enterprises. In *BH Ekonomski forum* (Vol. 12, No. 1, pp. 11-34).

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection & Assessment, 30*(1), 1–13. https://doi.org/10.1111/ijsa.12376

Lievens, F., & Chapman, D. (2019). Recruitment and selection. *The SAGE handbook of human resource management*, 123-150.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*(4), 426-441. https://doi.org/10.1108/00483480810877598

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? R*asch Measurement Transactions, 16* (2), 878.

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions, 19*(3), 1032.

Linacre, J. M. (2019). A User's Guide to WINSTEPS® Rasch-Model Computer Programs: Program Manual 4.4. 6. *Mesa-Press, Chicago, IL.*

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The international Journal of Educational and Psychological assessment, 1*(1), 1-11.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/bf02296272

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of personality and social psychology, 52*(6), 1258.

McCrae, R. R., & Greenberg, D. M. (2014). Openness to Experience. *The Wiley Handbook of Genius*, 222–243. https://doi.org/10.1002/9781118367377.ch12

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-reliability simulation. *Journal of Applied Psychology, 75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Nasifoglu Elidemir, S., Ozturen, A., & Bayighomog, S. W. (2020). Innovative Behaviors, Employee Creativity, and Sustainable Competitive Advantage: A Moderated Mediation. *Sustainability, 12*(8), 3295. https://doi.org/10.3390/su12083295

Nikolaou, I., Georgiou, K. and Kotsasarlidou, V. (2019), "Exploring the relationship of a gamified assessment with performance", *The Spanish Journal of Psychology, 22*, E6.

Olaru G, Burrus J, MacCann C, Zaromb FM, Wilhelm O, Roberts RD (2019) Situational Judgment Tests as a method for measuring personality: Development and validity evidence for a test of Dependability. *PLoS ONE 14*(2): e0211884. https://doi.org/10.1371/journal.pone.0211884

Oostrom, J. K., de Vries, R. E., & De Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance, 32*(1), 1-29. DOI: 10.1080/08959285.2018.1539856

Ostrow, K. S., & Heffernan, N. T. (2018). Testing the validity and reliability of intrinsic motivation inventory subscales within ASSISTments. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19* (pp. 381-394). Springer International Publishing. https://doi.org/10.1007/978-3-319-93843-1_28

Perreault, S., Nešlehová, J. G., & Duchesne, T. (2022). Hypothesis tests for structured rank correlation matrices. *Journal of the American Statistical Association,* 1-12.

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research. *Educational Psychologist, 39*(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1

Plucker, J. A., Runco, M. A., & Lim, W. (2006). Predicting Ideational Behavior from Divergent Thinking and Discretionary Time on Task. *Creativity Research Journal, 18*(1), 55–63. https://doi.org/10.1207/s15326934crj1801_7

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. *University of Chicago Press*.

Ratelle, C. F., Baldwin, M. W., & Vallerand, R. J. (2005). On the cued activation of situational motivation. *Journal of Experimental Social Psychology, 41*(5), 482-487.

Rendón-Macías, M. E., Villasís-Keever, M. Á., & Miranda-Novales, M. G. (2016). Descriptive statistics. *Revista Alergia de Mexico, 63*(4).

Runco, M. A., & Chand, I. (1995). Cognition and creativity. *Educational Psychology Review, 7*(3), 243–267. https://doi.org/10.1007/bf02213373

Runco, M. A., Plucker, J. A., & Lim, W. (2001). Development and Psychometric Integrity of a Measure of Ideational Behavior. *Creativity Research Journal, 13*(3-4), 393–400. https://doi.org/10.1207/s15326934crj1334_16

Ryan, A. M., Inceoglu, I., Bartram, D., Golubovich, J., Grand, J., Reeder, M., ... & Yao, X. (2015). Trends in testing: Highlights of a global survey. *Employee Recruitment, Selection, and Assessment*, 136-153.

Salikutluk, V., Kampling, H., & Niehaves, B. (2019). The Influence of Situational Factors and Gamification on Intrinsic Motivation and Learning.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5), 1763-1768.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality, 47*(5), 609-612.

Sen, S. (2016). Applying the Mixed Rasch Model to the Runco Ideational Behavior Scale. *Creativity Research Journal, 28*(4), 426–434. https://doi.org/10.1080/10400419.2016.1229985

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), 4-11.

Silvia, P. J., Kaufman, J. C., & Pretz, J. E. (2009). Is creativity domain-specific? Latent class models of creative accomplishments and creative self-descriptions. *Psychology of Aesthetics, Creativity, and the Arts, 3*(3), 139.

Stoeffler, K., Rosen, Y., Bolsinova, M., & von Davier, A. A. (2020). Gamified performance assessmentof collaborative problem solving skills. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2019.05.033

Tan, C. S., Lau, X. S., Kung, Y. T., & Kailsan, R. A. L. (2019). Openness to experience enhances creativity: The mediating role of intrinsic motivation and the creative process engagement. *The Journal of Creative Behavior, 53*(1), 109-119.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education, 2*, 53.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1), 35-39.

Teng, C.-I. (2017). Impact of avatar identification on online gamer loyalty: Perspectives of social identity and social capital theories. *International Journal of Information Management, 37*(6), 601-610.

Trigo, B. (2021). Desenvolvimento e validação de Testes de Julgamento Situacional para medição do potencial criativo e inovador em talentos de TI e Estudo do reconhecimento da importância desse potencial por profissionais de GRH [Review of Desenvolvimento e validação de Testes de Julgamento Situacional para medição do potencial criativo e inovador em talentos de TI e Estudo do reconhecimento da importância desse potencial por profissionais de GRH].

Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. In *Advances in experimental social psychology* (Vol. 29, pp. 271-360). Academic Press.

Vasconcellos, S. J. L., & Hutz, C. S. (2008). Construção e validação de uma escala de abertura à experiência. *Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment, 7*(2), 135-141.

von Stumm, S., Chung, A., & Furnham, A. (2011). Creative ability, creative ideation and latent classes of creative achievement: What is the role of personality? *Psychology of Aesthetics, Creativity, and the Arts, 5*(2), 107–114. https://doi.org/10.1037/a0020499

Woo, S. E., Chernyshenko, O. S., Longley, A., Zhang, Z.-X., Chiu, C.-Y., & Stark, S. E. (2013). Openness to Experience: Its Lower Level Structure, Measurement, and Cross-Cultural Equivalence. *Journal of Personality Assessment, 96*(1), 29–45. https://doi.org/10.1080/00223891.2013.806328

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology, 29*, 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

Wright, B. D., & Stone, M. H. (1979). Best test design.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse. C*omparing classical test theory*.

7. Annexes

Annex A:

Gamification link: https://gnpsa-iscteiul.itch.io/creative-gamification

Gamification Password: r6C)8#%QD6EBPWs7

7.1. Figures

Figure 9
Research Calendar

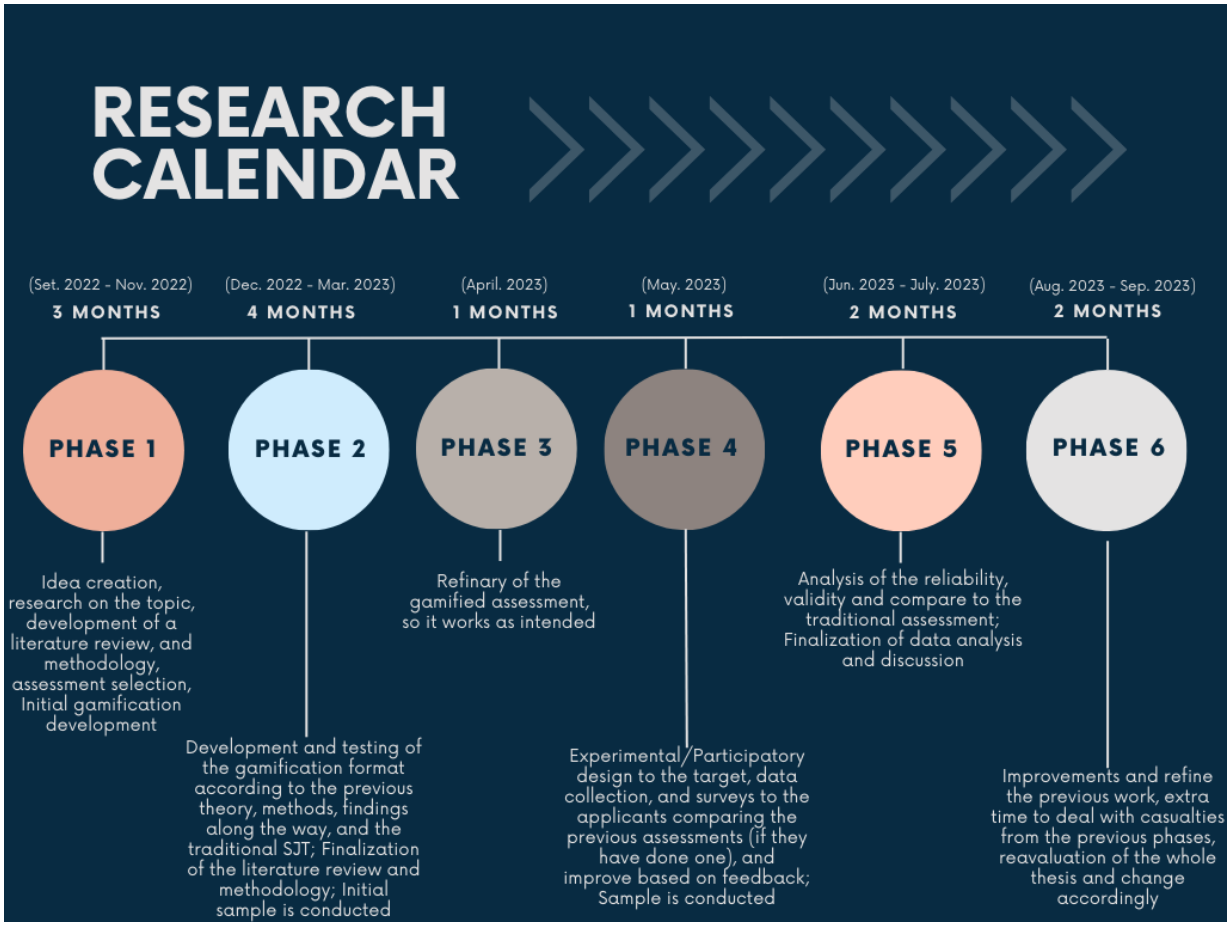


Figure 10
Gender Histogram

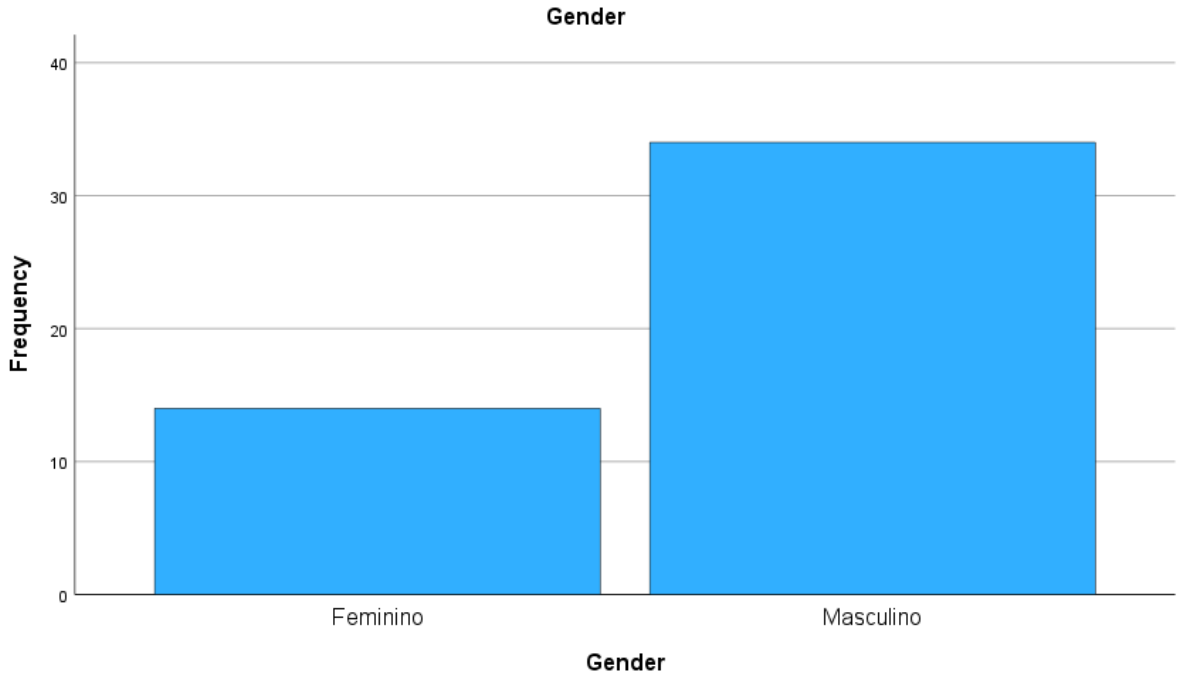


Figure 11
Histogram of the Age frequency

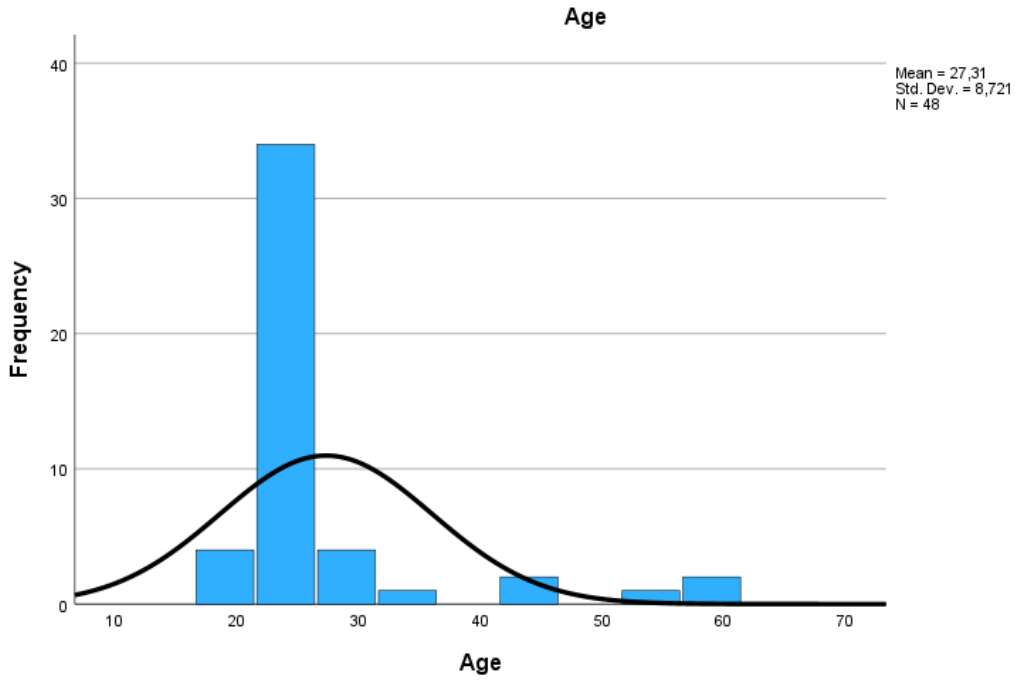
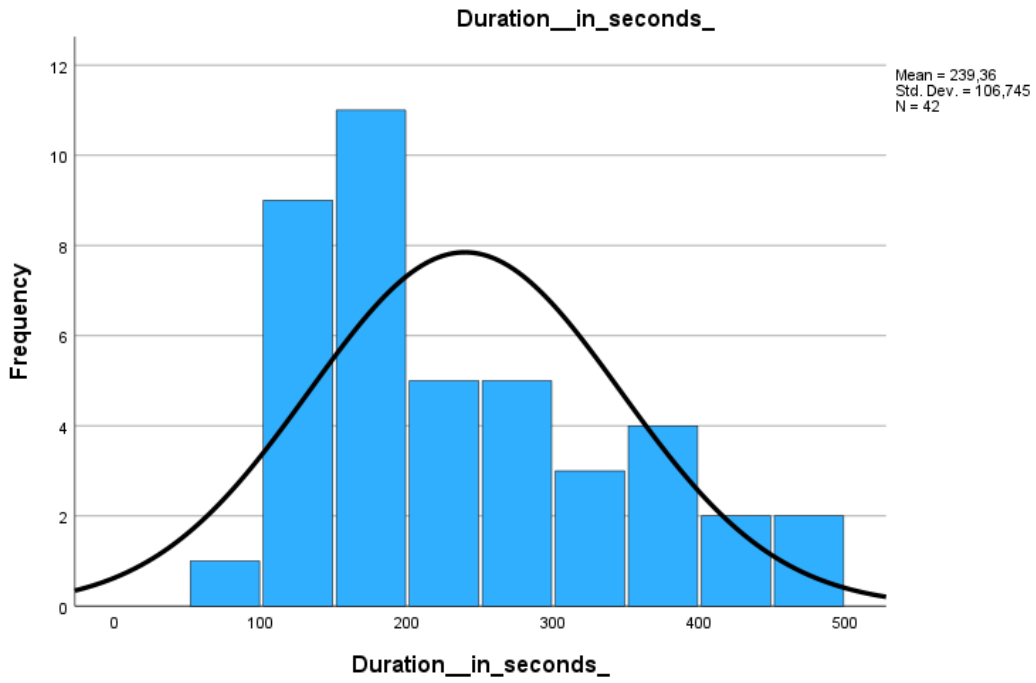


Figure 12

Histogram of the questionnaire duration



7.2. Tables

Table 10

Gender Frequency

	FREQUENCY	PERCENT	VALID PERCENT	CUMULATIVE PERCENT
FEMALE	14	29.17	29.17	29.2
MALE	34	70.83	70.83	100.0
TOTAL	48	100.0	100.0	

Table 11

Questionnaire duration and Age statistics

	DURATION IN SECONDS	AGE
VALID	42	48
MISSING	6	0
MEAN	239,36	27,31
STD. ERROR OF MEAN	16,47	1,26
MEDIAN	207,50	24,00
MODE	164[a]	24
STD. DEVIATION	106,74	8,72
VARIENCE	11394,43	76,05
SKEWNESS	,75	2,63
STD. ERROR OF SKEWNESS	,37	,34
KURTOSIS	-,42	6,24
STD. ERROR OF KURTOSIS	,72	,67
RANGE	403	39
MINIMUM	80	19
MAXIMUM	483	58
SUM	10053	1311
PERTENCIL 25	156,25	24,00
50	207,50	24,00
75	311,25	25,75

Note: Multiple modes exist. The smallest value is shown

Table 12

Age frequency

	FREQUENCY	PERCENT	VALID PERCENT	CUMULATIVE PERCENT
19	1	2,08	2,08	2,08
21	3	6,25	6,25	8,33
23	5	10,42	10,42	18,75
24	16	33,33	33,33	52,08
25	11	22,92	22,92	75,00
26	2	4,17	4,17	79,17
28	2	4,17	4,17	83,33
29	1	2,08	2,08	85,42
30	1	2,08	2,08	87,50
33	1	2,08	2,08	89,58
42	1	2,08	2,08	91,67
45	1	2,08	2,08	93,75
53	1	2,08	2,08	95,83
57	1	2,08	2,08	97,92
58	1	2,08	2,08	100,00
TOTAL	48	100,00	100,00	

Table 13

Job occupation Frequency

	FREQUENCY	PERCENT	VALID PERCENT	CUMULATIVE PERCENT
1	24	50,00	50,00	50,00
2	23	47,92	47,92	97,92
3	1	2,08	2,08	100.0
TOTAL	48	100.0	100.0	

Table 14

Job specialization Frequency

	FREQUENCY	PERCENT	VALID PERCENT	CUMULATIVE PERCENT
	27	56.3	56.3	56.3
ADMINISTRATIVE	1	2,1	2,1	58,3
AUTOMOBILE	2	4,2	4,2	60,4
CONSULTANCY	2	4,2	4,2	64,6
PRODUCT DESIGN	1	2,1	2,1	66,7
FINANCE	1	2,1	2,1	68,8
HOSPITALITY	3	6,3	6,3	75,0
REAL ESTATE	1	2,1	2,1	77,1
INVESTIGATION	1	2,1	2,1	81,3
SHOPKEEPER	1	2,1	2,1	83,4
INTERNAL ADMINISTRATION MINISTRY	1	2,1	2,1	85,5
MUSIC	1	2,1	2,1	87,5
POOL	1	2,1	2,1	89,6
PSYCHOLOGY	1	2,1	2,1	91,7
HUMAN RESOURCES	1	2,1	2,1	93,8
SERVICES	1	2,1	2,1	95,9
RENT A CAR	1	2,1	2,1	97,9
TRANSPORTS	1	2,08	2,08	100,00
TOTAL	48	100,00	100,00	

Table 15

Comments

Topics	Comments
Positive comments	"Feito em código, está um jogo bem responsivo, com boas capacidades." "Estava excelente." "O jogo está bem idealizado parabéns." "Tudo ok." "Para uma tese está bastante interessante, só dar double check pa ver se tem bugs." "Gostei da história do jogo e acho que devia de haver algumas melhorias mas físicas do jogo." "Gostei de toda a dinâmica das perguntas e respostas." "Gostei dos gráficos do jogo e do facto de ter muitas personagens diferentes." "Jogo demasiado bom para o propósito que tem." "Gostei da variedade de possibilidade de resposta."

	"Achei interessante para perceber a capacidade e o interesse no trabalho." "Gostei do bom desenvolvimento geral e raciocínio por trás da sequência de perguntas. Acho que no tamanho da letra e dimensão das frases de texto poderia ser diminuído a fim de facilitar a leitura e torná-la mais prática." "Gostei do jogo no geral." "Gostei da forma como a gamificação dá uma contextualização e visualização do tipo de situação que a personagem estava a vivenciar, o que ajuda a dar uma melhor resposta com base na conjuntura existente. Em termos de otimização provavelmente seria os gráficos/design."
Negative comments	"A leitura das perguntas não foi muito agradável." "Acho que o jogo deveria ter menos texto e perguntas mais diretas. Às vezes sentia dificuldade em manter a minha atenção e tinha de reler as perguntas e as respostas várias vezes." "Não gostei dos longos percursos com diversos obstáculos." "Tinha que ser mais intuitivo, mais rápido. Demorei algum tempo a perceber para onde me dirigir, o que me tirou a vontade de ler as perguntas." "Ao utilizar dois monitores por vezes ocorreu um bug se eu clicar no ecrã em que não está o jogo, levando a que não conseguisse selecionar o "continuar a meio dos diálogos". Resolvi o problema abrindo o menu do jogo e utilizando aí o rato para clicar no "continuar". "
Suggestions	"Gráficos, navegabilidade." "Um objetivo final diferente tendo em conta a diferença nas respostas dadas ao longo do jogo." "O texto deveria ser mais resumido." "Simplificação da linguagem e dos desafios - necessidade de maior enquadramento sobre quem sou eu e qual é o objetivo da minha personagem." "Menos pessoas." "Acho que deveria melhorar os gráficos do jogo e a câmara esta muito próxima." "Colocar setas no chão com os percursos sugeridos para mais facilmente chegar ao destino e eliminar algumas barreiras arquitetónicas/mobiliário." "Não preciso controlar a personagem porque cada interação é sequencial, mais valia saltar de uma frame para a outra." "O gameplay e a história." "Os bugs já foram comunicados ao programador."

	"As diferentes escolhas possuírem impacto/diferentes reações nos npcs." "Haver uma opção para mudar as definições gráficas, o jogo estava com FPS muito baixos. Usar vozes em vez de texto, torna-se mais fácil tomar atenção (ou os dois juntos). A seta de orientação apontava para a direção do NPC no mapa, talvez direcionar no chão seria mais fácil. Mas de resto a experiência foi agradável, mas algumas melhorias podia criar o jogo mais dinâmico. Bom trabalho." "Mais opções diálogo ou questões com percentagens/probabilidades de algo acontecer. (Exemplo: probabilidade de contratar X indivíduo de 0 a100%)." "Em algumas situações escolhemos uma solução e depois temos de ficar à espera que esse mesmo texto aparece novamente no ecrã como fala da personagem. Se já escolhemos essa opção, não seria necessário mostrar novamente ou, a mostrar, deveria aparecer rapidamente e não palavra a palavra." "O controlo da camara, os dois shifts deveriam funcionar." "É um jogo algo monótono. Compreendo a ideia e tem potencial, no entanto acredito que poderia ser mais interativo. Além de que estive muito tempo a tentar descobrir a folha de atividades, o que acabou por se tornar frustrante."

Table 16

KMO and Bartlett Test of the partial credit model

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	**,400**
Bartlett Test of Sphericity. Approx. Chi-Square	445,292
DF	351
SIG.	<,001

Table 17

Item's Communalities for the partial credit model

	INITIAL
ITEM1_DTE_PARTIAL	*,645*
ITEM2_DTE_PARTIAL	*,761*
ITEM3_DTE_PARTIAL	*,497*
ITEM4_DTE_PARTIAL	*,729*
ITEM5_DTE_PARTIAL	*,619*
ITEM1_CI_PARTIAL	*,627*
ITEM2_CI_PARTIAL	*,632*
ITEM3_CI_PARTIAL	*,621*
ITEM4_CI_PARTIAL	*,674*
ITEM5_CI_PARTIAL	*,543*
ITEM6_CI_PARTIAL	*,525*
ITEM7_CI_PARTIAL	*,415*
ITEM8_CI_PARTIAL	*,715*
ITEM9_CI_PARTIAL	*,554*
ITEM10_CI_PARTIAL	*,410*
ITEM11_CI_PARTIAL	*,399*
ITEM12_CI_PARTIAL	*,715*
ITEM13_CI_PARTIAL	*,679*
ITEM14_CI_PARTIAL	*,624*
ITEM15_CI_PARTIAL	*,642*
ITEM16_CI_PARTIAL	*,707*
ITEM17_CI_PARTIAL	*,657*
ITEM18_CI_PARTIAL	*,697*
ITEM19_CI_PARTIAL	*,642*
ITEM20_CI_PARTIAL	*,684*
ITEM21_CI_PARTIAL	*,447*
ITEM22_CI_PARTIAL	*,549*

Note: Extraction Method: Maximum Likelihood

a. *One or more communalitiy estimates greater than 1 were encountered during iterations. The resulting solution should be interpreted with caution.*

Table 18

Total Variance Explained of the items for the partial credit model.

FACTOR	*Initial Eigenvalues*			*Rotation Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	*3,701*	*13,709*	*13,709*	*1,988*	*7,364*	*7,364*
2	*2,802*	*10,376*	*24,085*	*1,931*	*7,153*	*14,517*
3	*2,298*	*8,512*	*32,597*	*1,845*	*6,834*	*21,351*
4	*2,135*	*7,907*	*40,504*	*1,652*	*6,118*	*27,470*
5	*1,699*	*6,294*	*46,798*	*1,524*	*5,643*	*33,113*
6	*1,635*	*6,057*	*52,854*	*1,508*	*5,585*	*38,697*
7	*1,577*	*5,842*	*58,697*	*1,411*	*5,227*	*43,924*
8	*1,247*	*4,620*	*63,317*	*1,409*	*5,217*	*49,141*
9	*1,097*	*4,064*	*67,381*	*1,316*	*4,873*	*54,014*
10	*1,074*	*3,977*	*71,358*	*1,206*	*4,468*	*58,482*
11	*0,985*	*3,648*	*75,006*			
12	*0,936*	*3,468*	*78,474*			
13	*0,824*	*3,053*	*81,528*			
14	*0,725*	*2,686*	*84,213*			
15	*0,601*	*2,226*	*86,439*			
16	*0,583*	*2,161*	*88,600*			
17	*0,556*	*2,060*	*90,660*			
18	*0,508*	*1,881*	*92,541*			
19	*0,460*	*1,702*	*94,243*			
20	*0,363*	*1,346*	*95,588*			
21	*0,282*	*1,043*	*96,632*			
22	*0,247*	*0,915*	*97,547*			
23	*0,197*	*0,731*	*98,278*			
24	*0,164*	*0,606*	*98,884*			
25	*0,122*	*0,451*	*99,334*			
26	*0,110*	*0,408*	*99,742*			
27	*0,070*	*0,258*	*100,000*			

EXTRACTION METHOD: MAXIMUM LIKELIHOOD.

Table 19

KMO and Bartlett Test of the dichotomous Rasch model

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	**,389**
Bartlett Test of Sphericity. Approx. Chi-Square	354,593
DF	351
SIG.	,436

Table 20

Item's Communalities for the dichotomous Rasch model

	INITIAL
ITEM1_DTE_DICHOTOMOUS	*0,584*
ITEM2_DTE_DICHOTOMOUS	*0,567*
ITEM3_DTE_DICHOTOMOUS	*0,470*
ITEM4_DTE_DICHOTOMOUS	*0,551*
ITEM5_DTE_DICHOTOMOUS	*0,630*
ITEM1_CI_DICHOTOMOUS	*0,560*
ITEM2_CI_DICHOTOMOUS	*0,545*
ITEM3_CI_DICHOTOMOUS	*0,473*
ITEM4_CI_DICHOTOMOUS	*0,638*
ITEM5_CI_DICHOTOMOUS	*0,545*
ITEM6_CI_DICHOTOMOUS	*0,495*
ITEM7_CI_DICHOTOMOUS	*0,500*
ITEM8_CI_DICHOTOMOUS	*0,568*
ITEM9_CI_DICHOTOMOUS	*0,538*
ITEM10_CI_DICHOTOMOUS	*0,449*
ITEM11_CI_DICHOTOMOUS	*0,533*
ITEM12_CI_DICHOTOMOUS	*0,716*
ITEM13_CI_DICHOTOMOUS	*0,550*
ITEM14_CI_DICHOTOMOUS	*0,371*
ITEM15_CI_DICHOTOMOUS	*0,600*
ITEM16_CI_DICHOTOMOUS	*0,524*
ITEM17_CI_DICHOTOMOUS	*0,465*
ITEM18_CI_DICHOTOMOUS	*0,538*
ITEM19_CI_DICHOTOMOUS	*0,453*
ITEM20_CI_DICHOTOMOUS	*0,571*
ITEM21_CI_DICHOTOMOUS	*0,497*
ITEM22_CI_DICHOTOMOUS	*0,359*

Note: Extraction Method: Maximum Likelihood.
a. One or more communalitiy estimates greater than 1 were encountered during iterations. The resulting solution should be interpreted with caution.

Table 21

Total Variance Explained of the items for the dichotomous Rasch model

TOTAL VARIANCE EXPLAINED

FACTOR	*Initial Eigenvalues*			*Rotation Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	*2,822*	*10,453*	*10,453*	*1,665*	*6,167*	*6,167*
2	*2,438*	*9,030*	*19,483*	*1,637*	*6,063*	*12,231*
3	*2,185*	*8,094*	*27,577*	*1,549*	*5,736*	*17,966*
4	*2,020*	*7,480*	*35,057*	*1,480*	*5,481*	*23,447*
5	*1,877*	*6,953*	*42,010*	*1,477*	*5,471*	*28,918*
6	*1,845*	*6,832*	*48,842*	*1,438*	*5,327*	*34,245*
7	*1,582*	*5,860*	*54,701*	*1,375*	*5,093*	*39,338*
8	*1,400*	*5,184*	*59,885*	*1,313*	*4,863*	*44,201*
9	*1,330*	*4,924*	*64,810*	*1,261*	*4,672*	*48,873*
10	*1,105*	*4,091*	*68,901*	*1,252*	*4,639*	*53,512*
11	*1,019*	*3,772*	*72,673*	*1,220*	*4,517*	*58,029*
12	*0,908*	*3,361*	*76,034*			
13	*0,824*	*3,053*	*79,087*			
14	*0,762*	*2,822*	*81,909*			
15	*0,710*	*2,628*	*84,537*			
16	*0,695*	*2,575*	*87,112*			
17	*0,622*	*2,304*	*89,416*			
18	*0,498*	*1,843*	*91,258*			
19	*0,476*	*1,762*	*93,020*			
20	*0,411*	*1,523*	*94,543*			
21	*0,334*	*1,236*	*95,779*			
22	*0,272*	*1,007*	*96,786*			
23	*0,228*	*0,843*	*97,629*			
24	*0,204*	*0,757*	*98,386*			
25	*0,178*	*0,661*	*99,047*			
26	*0,150*	*0,556*	*99,603*			
27	*0,107*	*0,397*	*100,000*			

EXTRACTION METHOD: MAXIMUM LIKELIHOOD.

Table 22

Reliability Statistics of the Factor

CRONBACH'S ALPHA	***N OF ITEMS***
,726	*16*

Table 23

Wright Map

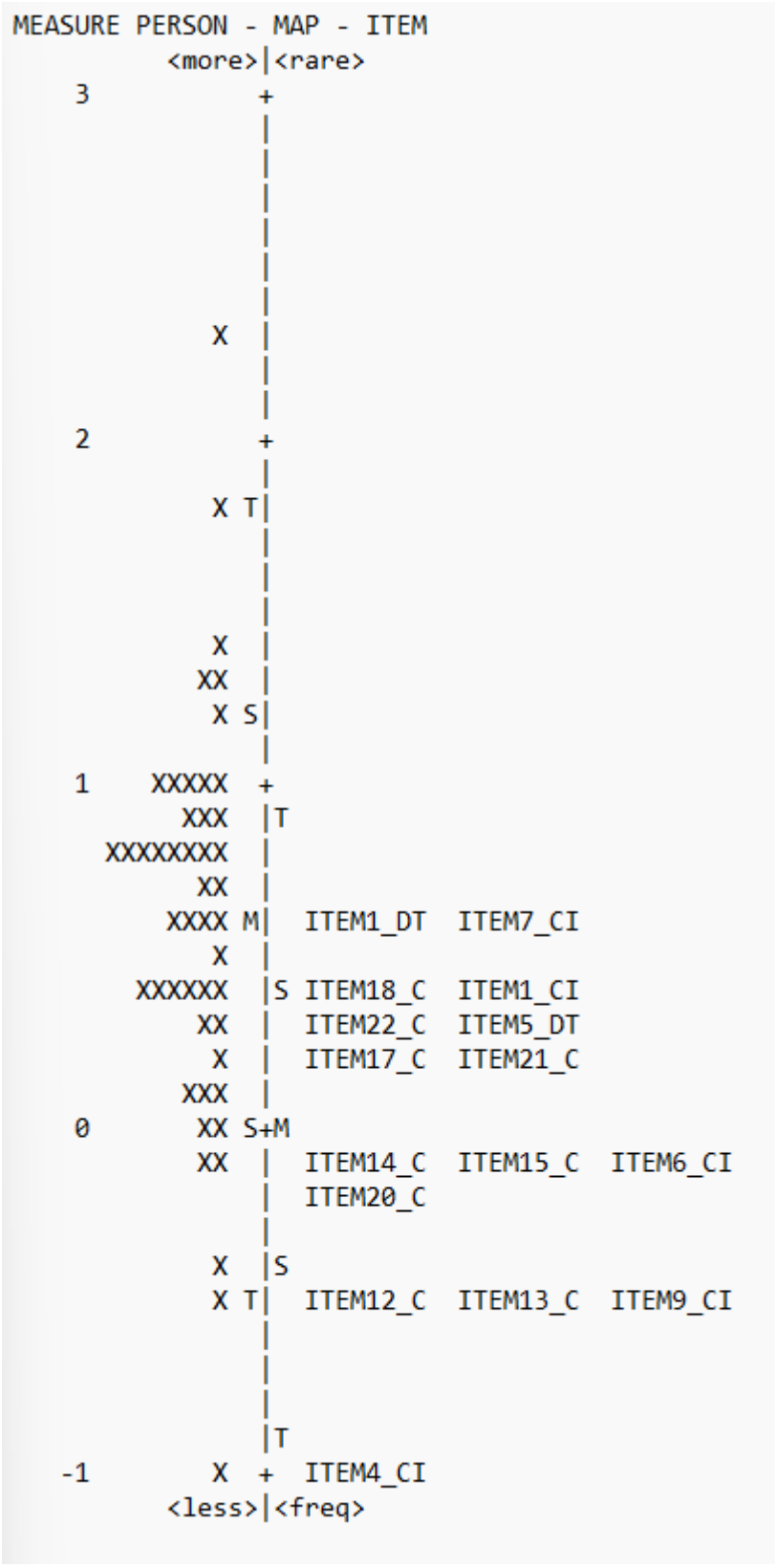


Table 24

Item category/option/distractor frequencies: Measure order

ITEM CATEGORY/OPTION/DISTRACTOR FREQUENCIES: MEASURE ORDER

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	DATA %	ABILITY MEAN	ABILITY P.SD	S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR.	ITEM
1	0	0	13	27	.25	.56	.16	1.1	1.0	-.37	ITEM1_DT
	1	1	5	10	.45	.45	.22	.8	.9	-.09	
	2	2	21	44	.57	.37	.08	1.2	1.1	-.06	
	3	3	9	19	1.29	.46	.16	.6	.6	.57	
6	0	0	7	15	-.03	.51	.21	.7	.7	-.45	ITEM7_CI
	1	1	16	33	.70	.53	.14	1.7	1.8	.12	
	2	2	17	35	.60*	.55	.14	1.4	1.4	-.01	
	3	3	8	17	.98	.20	.08	.8	.8	.29	
13	0	0	14	29	.28	.55	.15	1.2	1.2	-.37	ITEM18_C
	1	1	4	8	.02*	.40	.23	.3	.3	-.31	
	2	2	11	23	.77	.30	.10	.4	.3	.15	
	3	3	19	40	.88	.54	.13	1.0	1.0	.38	
3	0	0	4	8	-.07	.31	.18	.6	.6	-.36	ITEM1_CI
	1	1	13	27	.39	.58	.17	1.1	1.0	-.23	
	2	2	21	44	.68	.40	.09	.8	.6	.12	
	3	3	10	21	1.00	.61	.20	1.0	.9	.35	
2	0	0	5	10	-.04	.69	.35	1.0	1.0	-.38	ITEM5_DT
	2	2	42	88	.69	.51	.08	1.0	1.1	.38	
	3	3	1	2	.36*	.00		1.4	1.5	-.06	
16	0	0	11	23	.17	.44	.14	.9	.9	-.41	ITEM22_C
	1	1	6	13	.38	.44	.20	.8	.9	-.15	
	2	2	12	25	.79	.56	.17	1.3	1.3	.19	
	3	3	19	40	.81	.54	.13	1.1	1.2	.29	
15	0	0	6	13	.00	.60	.27	.9	.9	-.40	ITEM21_C
	1	1	11	23	.58	.30	.10	1.2	1.1	-.02	
	2	2	14	29	.63	.47	.13	1.0	.9	.03	
	3	3	17	35	.82	.63	.16	1.2	1.2	.27	
12	0	0	5	10	-.02	.67	.33	1.0	1.0	-.37	ITEM17_C
	1	1	4	8	.08	.42	.25	.5	.4	-.27	
	2	2	28	58	.64	.36	.07	.6	.5	.07	
	3	3	11	23	.99	.66	.21	1.0	.9	.37	
10	0	0	7	15	-.10	.45	.19	.8	.7	-.50	ITEM14_C
	1	1	5	10	.58	.32	.16	1.4	1.3	-.01	
	2	2	12	25	.77	.36	.11	.9	.9	.16	
	3	3	24	50	.74*	.59	.12	1.2	1.2	.22	
11	0	0	3	6	.14	.17	.12	1.0	.9	-.21	ITEM15_C
	1	1	15	31	.26	.46	.12	.8	.8	-.41	
	2	2	3	6	.75	.25	.18	.8	.7	.06	
	3	3	27	56	.84	.56	.11	1.1	1.1	.45	
5	0	0	3	6	.24	.28	.20	1.2	1.1	-.17	ITEM6_CI
	1	1	12	25	.33	.65	.20	1.2	1.2	-.28	
	2	2	8	17	.64	.34	.13	.7	.7	.03	
	3	3	25	52	.77	.55	.11	1.1	1.1	.30	
14	0	0	4	8	-.07	.70	.40	1.0	1.0	-.35	ITEM20_C
	1	1	8	17	.12	.38	.14	.6	.6	-.38	
	2	2	12	25	.91	.63	.19	1.7	2.7	.31	
	3	3	24	50	.73*	.35	.07	.9	1.0	.21	
9	0	0	2	4	.23	.14	.14	1.4	1.2	-.14	ITEM13_C
	1	1	5	10	.26	.51	.25	1.0	1.0	-.20	
	2	2	15	31	.49	.54	.14	1.2	1.0	-.14	
	3	3	26	54	.77	.57	.11	1.1	1.0	.31	
7	0	0	3	6	-.35	.59	.42	.7	.7	-.43	ITEM9_CI
	1	1	3	6	.60	.20	.14	1.9	1.4	.00	
	2	2	14	29	.59*	.42	.12	1.1	1.0	-.02	
	3	3	28	58	.72	.57	.11	1.1	1.1	.23	
8	0	0	1	2	.76	.00		3.1	2.4	.04	ITEM12_C
	1	1	5	10	.09*	.38	.19	.8	.7	-.31	
	2	2	16	33	.32*	.39	.10	.5	.5	-.35	
	3	3	26	54	.88	.57	.11	1.1	1.0	.51	
4	0	0	2	4	-.31	.20	.20	.9	.7	-.33	ITEM4_CI
	1	1	2	4	-.31*	.67	.67	.5	.6	-.33	
	2	2	9	19	.61	.42	.15	1.4	1.2	.00	
	3	3	35	73	.71	.53	.09	1.0	1.0	.30	

* Average ability does not ascend with category score

Table 25

Reliability Statistics divergent thinking

CRONBACH'S ALPHA	*N OF ITEMS*
,582	*6*

Table 26

Item-Total Statistics divergent thinking

	SCALE OF MEAN IF ITEM DELETED	*SCALE OF VARIANCE IF ITEM DELETED*	*CORRECTED ITEM-TOTAL CORRELATION*	*CRONBACH'S ALPHA IF ITEM DELETED*
Q20	*17,458*	*9,785*	*,318*	*,551*
Q21	*17,458*	*9,998*	*,211*	*,578*
Q22	*17,813*	*9,177*	*,301*	*,547*
Q23	*18,00*	*6,979*	*,514*	*,435*
Q24	*18,792*	*8,296*	*,244*	*,579*
Q25	*18,292*	*6,848*	*,396*	*,503*

Table 27

Reliability Statistics OE

CRONBACH'S ALPHA	*N OF ITEMS*
,811	*3*

Table 28

Item-Total Statistics OE

	SCALE OF MEAN IF ITEM DELETED	*SCALE OF VARIANCE IF ITEM DELETED*	*CORRECTED ITEM-TOTAL CORRELATION*	*CRONBACH'S ALPHA IF ITEM DELETED*
Q7	*12,417*	*2,759*	*,709*	*,717*
Q8	*11,896*	*3,712*	*,683*	*,719*
Q9	*12,063*	*4,315*	*,646*	*,773*

Table 29

Reliability Statistics intrinsic motivation

CRONBACH'S ALPHA	*N OF ITEMS*
,932	*4*

Table 30

Item-Total Statistics intrinsic motivation

	SCALE OF MEAN IF ITEM DELETED	*SCALE OF VARIANCE IF ITEM DELETED*	*CORRECTED ITEM-TOTAL CORRELATION*	*CRONBACH'S ALPHA IF ITEM DELETED*
Q17	*15,208*	*18,126*	*,780*	*,931*
Q26	*15,917*	*16,163*	*,864*	*,903*
Q27	*15,958*	*14,934*	*,849*	*,910*
Q28	*16,042*	*15,402*	*,884*	*,896*