MASTER THESIS

Applicability of Comparative Judgement as Alternative to Rubric-Based Assessment for High-Stake Exam Items That Differ in Maximum Score

Jolien Valk
S1986422

Faculty of Behavioural, Management, and Social Sciences
Department of Educational Science and Technology

EXAMINATION COMMITTEE
Prof. dr. ir. B. P. Veldkamp
A. Bijl, MSc

EXTERNAL ORGANISATION
eX:plain

EXTERNAL SUPERVISORS
Dr. S. de Klerk
M. van Puffelen, MSc

Keywords: educational measurement, comparative judgement, pairwise comparisons, rubric-based assessment, reliability

January 28, 2022

Acknowledgement

This thesis is the end product of a half-year internship at eX:plain, where I have had the opportunity to work alongside amazing and bright colleagues. Working on this thesis has confirmed my interest in educational sciences. When I first started, I had minimal knowledge on educational measurement, and thanks to the support of my supervisors and colleagues I have grown to understand a lot more. I am fortunate to have had a lot of support and guidance of so many people, without whom this would not have been possible.

First and foremost, I would like to thank my supervisors Bernard and Aranka for their guidance, encouragement, and feedback which has helped me towards improving and reflecting on my work.

Secondly, I would like to thank my colleagues at eX:plain, especially Sebastiaan and Melvin, for the opportunity to work on this project and for advising me throughout my internship.

I would also like to thank the raters and the colleagues at the Associatie voor Examinering. Your views and experiences have helped me greatly to reflect critically on my work.

A special thanks goes out to my boyfriend, my dad, my sister, my grandparents, and my friends, for their continuous support during the writing of my thesis. The coffee breaks, games, and calls have helped me to reset and regain my motivation and eagerness to continue working on my thesis.

Lastly, a special thanks goes out to my mom who has always believed in me and whose support keeps on motivating me every day. If it wasn’t for you, I would have never recognised my interest in educational sciences.

Abstract

To adequately assess candidates’ competences, candidates must provide proof of performance through answering items on a test. These items are often assessed through rubric-based assessments (RBA) to provide test takers with a score that represents their understanding. However, studies have indicated that RBAs are not sufficiently reliable (< 0.7) for high-stake assessments. An alternative method, comparative judgement (CJ) had been introduced to tackle this reliability issue. CJ is based on the assumption that people are more reliable in comparing items, objects, or performances than in assigning scores to a single item, object, or performance. Previous studies have indicated that CJ is a reliable, effective, and valid form of assessment.

In this study, three exam items with a maximum score of 2, 3, and 4, that had already been assessed through RBA, were assessed through CJ. The results of this study indicated that the reliability of CJ for the item with the maximum score of 2 was the highest (0.86), followed by the reliability of CJ for the item with a maximum of score 4 (0.83). CJ yielded the lowest reliability for the item with a maximum score of 3 (0.77). The mean time investment for raters increased as the maximum score of the items increased. Furthermore, the rank correlations between RBA and CJ were moderately positive, and the correlation increased with the increase of the maximum score.

Although these results indicate that CJ might be applicable for high-stake exam items that vary in maximum score, there are some considerations. First, this study focused on single items that were carefully selected. However, exams exist of multiple items that together form a balanced exam, and it seemed inefficient to apply CJ for every single item. Secondly, the assessment method of CJ does not offer transparency towards the candidates, which is essential for high-stake assessments. Lastly, it seems CJ is biased towards the differences in the candidate pool sample. Taking these ambiguities, it was concluded that CJ is an inefficient method of assessment that lacks transparency when applied to the high-stake assessment as researched within this study’s context.

Contents

List of Tables

List of Figures

Acronyms

RBA	Rubric-based assessment
CJ	Comparative judgement
AvE	Associatie voor Examinering
SSR	Scale separation reliability
RQ	Research question
H	Hypothesis
BIV®	Bestuurlijke Informatievoorziening
D-PAC	Digital platform for assessing competencies
ACJ	Adaptive comparative judgement
CAT	Computerized adaptive testing

Introduction

Knowledge, skills, and abilities are crucial for a successful career, but how can these be assessed properly? Assessment refers to “the process of obtaining information that is used to make educational decisions about students; to give feedback to students about their progress, strengths, weaknesses; to judge instructional effectiveness and curricular adequacy; and to inform policy” (Sanders & Vogel, 1993, p. 41).

Assessment is often performed by one or multiple raters. It is important that raters are consistent in how they assess a candidate’s performance, otherwise, the reliability will decrease. Although for low-stake assessments reliability is important, for high-stake assessments a high reliability is essential. It could be damaging for an organisation if educational decisions such as passing or failing an exam are based on unreliable assessment (Coenen et al., 2018). Therefore, it's important to determine the best method to assess candidate performance in high-stakes assessment reliably.

There are two major approaches in assessment: analytic assessments and holistic assessments. In analytical assessments, candidate work is judged separately on pre-set criteria while in holistic assessments the raters assess the performance of a candidate based on the overall quality (Sadler, 2009). The focus of this study will be on the comparison of the predominantly used analytical assessment method of rubric-based assessment (RBA) and the emerging holistic assessment method of comparative judgement (CJ) when applied to high-stake assessments.

As mentioned, an often-used method of analytical assessment is the traditional RBA method. In this conventional approach, the examination exists of items to which the answers are measured against a rubric that consists of a set of criteria (Nasab, 2015). Jönsson and Panadero (2016) have concluded that a wide implementation of rubrics can have great potential for candidate performance. They described that rubrics support learning by facilitating self-regulated learning as well as facilitating the understanding and use of feedback. However, the use of rubrics has its disadvantages. The preparation, calibration, and monitoring of scores for the RBA method is a time-consuming, cognitively demanding, and resource-intensive process (Steedle & Ferrara, 2016). Furthermore, grades are often inconsistent across different raters resulting in a low reliability (Jones et al., 2019). For low-stake assessments, lower reliability values might be sufficient (Jönsson & Svingby, 2007). However, as already mentioned, for high-stakes assessments reliability is crucial which means that another assessment method might be more suitable.

An alternative to RBA is the holistic assessment method CJ. The concept of CJ is based on the ‘Law of Comparative Judgement’ (Thurstone, 1927), which describes how people are more reliable in comparing items, objects, or performances than in assigning scores to a single item, object, or performance. In CJ, raters compare pairs of work to decide which one of them is better. Based on these holistic judgements, an interval scale ranging from worst to best can be created which shows the relative quality of each candidate's performance (Coertjens et al., 2021; Pollitt, 2012).

Studies have shown that CJ is more advantageous when compared to RBA. First, CJ has an increased validity when compared to RBA. This is because CJ is based on relative judgements which are more accurate than absolute judgements of candidate performance (Steedle & Ferrara, 2016). Secondly, CJ has a high reliability, which refers to the precision of measurement that it can deliver in any context in which subjective judgement is appropriate (Pollitt, 2012). The high reliability is caused by the principle that CJ is based on independent judgements. Raters can be asked to make more judgements, or to rate a particular exam that is close to a grade boundary, thus increasing the reliability. Thirdly, in RBA there are issues regarding raters that do not matter when CJ is applied (McMahon & Jones, 2015; Pollitt, 2012). Some raters can score stricter than others. Furthermore, although a rater may award the same average grades, this rater may discriminate more finely among individual exams, being more favourable to the better ones and harsher to the poorer, or vice versa. Lastly, when CJ is applied as an assessment method, there is a reduction in the time it takes to train raters, as well as a reduction in time spent on assessing (Steedle & Ferrara, 2016).

CJ is found to be most useful when applied to assess competences that are too complex to assess through one aspect such as reflections (Coertjens et al., 2021), essays (Heldsinger & Humphry, 2010), and e-Portfolios (Kimbell, 2012). However, Jones et al. (2015) have also applied CJ to assess a single competence, namely mathematical problem-solving. In this study, CJ was applied to two individual mathematical items with an item maximum score of 4. The results indicate that even for assessing a single competence, CJ is a reliable and valid method of assessment. While this study focused on two items with the same maximum score, tests often consist of multiple items with varying maximum scores. No studies have been found that investigate CJ for items with varying maximum scores. Therefore, this study will contribute to the scientific body concerning the applicability of CJ. This research aims to investigate how the item maximum score influences the reliability and efficiency of CJ.

External Organisation

This research is conducted on behalf of eX:plain. eX:plain is a professional exam institute that focuses on vocational assessment and certification testing for both the industry and education (eX:plain, n.d.). In cooperation with companies, sectors and educational organisations, the labour market needs are determined. Based on these needs, the educational standard is established and examined. Using practice-oriented, multimedia learning resources, eX:plain aims to help people to develop themselves.

eX:plain offers services from training and consultancy to digital learning and test development opportunities to both lower and higher vocational education. Assessment programmes are developed and support for the exams are built in areas varying from public security to salary administration.

Several exam programmes are carried out at eX:plain, including the Exameninstelling Toezicht en Handhaving, the Safety, Health, and Environment Checklist for Contractors, and the Associatie voor Examinering (AvE).

The AvE is one of the largest examination programmes in the Netherlands. They have been the experts in the field of exam development, assessment, and administration for 80 years. This is reflected in the AvE diplomas that are valuable, retain their value, and are developed at every level (Associatie, n.d.).

The AvE offers a variety of exams that combine multiple-choice items as well as open-ended questions. The multiple-choice items have a maximum score of 1 and are automatically assessed through a digital system. The open-ended items differ in pre-determined maximum scores depending on the level of understanding that is needed to answer correctly. These items are currently assessed through RBA by external raters, but as mentioned the reliability of RBA is not sufficient for high-stake examinations. Since studies have indicated that CJ is a reliable and valid method of assessment (Pollitt, 2012; Steedle & Ferrara, 2016), this alternative assessment method of CJ might be more suitable for the high-stake exams of the AvE.

Report Outline

After the introduction, the report will continue with the Theoretical Framework. In this section, an overview of the theoretical perspectives and literature will be discussed. After the Theoretical Framework, the Method chapter will give insights into the followed method. The study design will be covered as well as the used instruments, participants, procedures, and data analysis. The Method chapter is followed by the Results chapter. After this, the Conclusion chapter will critically discuss the results. Moreover, this chapter will cover the limitations of the study, the conclusions that can be drawn from this study, and some recommendations for future research.

Theoretical Framework

This section gives an overview of the theoretical perspectives and literature. Educational measurement is a broad topic with various views. However, to keep the information within the scope of this study, the topics will be discussed as related to high-stake assessments. An elaborate framework of the definitions of educational tests, educational assessments, and test item analysis has been added to Appendix A. In this section, the assessment method of RBA will be discussed as well as the alternative holistic assessment method of CJ.

Rubric-Based Assessment

The Use of Rubrics for Assessment

An often-used method to determine candidate performance is using RBAs. As the name indicates, this method uses rubrics as an instrument to guide raters in judging candidate performance. A rubric can be defined as a series of statements that describe the levels of candidate performance (Almarshoud, 2011). It often includes a set of criteria to which learning outcomes can be linked to assess candidate performance. Based on the thoroughness or accurateness of the candidate's answer, a numerical score is assigned to the answer. The candidate's level of ability in the examination is based on the final cumulative score of each criterion (Kimbell, 2021; Marshall, 2017).

Fundamentals of Rubrics

Since a rubric is an instrument that assists raters in judging the quality of candidate performance, it needs to be well-designed. According to Jönsson and Panadero (2016), rubrics have three fundamental features. First, rubrics must include specific information about what aspects or criteria to look for in candidate performance. This is needed to ensure the correct candidate qualities are identified by raters. Second, the rubric must include descriptions of candidate performance of different levels of quality. With the description of the different levels of quality, it is tried to assist raters in assessing the candidate's understanding. Lastly, a rubric must be accompanied by a scoring strategy. This last step ensures that candidates are rewarded with scores that correspond with the quality of their answer and thus the candidates' level of understanding.

Reliability of Rubric-Based Assessment

Reliability is an important concern in any form of assessment. To determine to what extent RBA is reliable, the assessment scores through different raters as well as through the same rater at different points in time needs to be determined. For intra-rater reliability, Cronbach's alpha is most widely used to estimate rater consistency. For determining the inter-rater reliability, the most frequently used methods are generalization theory and Rasch models (Jönsson & Svingby, 2007).

Stellmack et al. (2009) studied the reliability and validity of a grading rubric applied for rating APA-style introductions. They found that both the intra- and inter-rater reliability values were low but concluded that these low values of reliability were similar to values reported in the literature for comparable research. This is in line with the findings of Jönsson and Svingby (2007), who reviewed research on rubrics regarding their reliability and validity. They found that intra-rater reliability is not a major concern, as most of the studies had reported a Cronbach's alpha larger than 0.7 which is sufficiently reliable. However, for inter-rater reliability, studies often reported an agreement consistency below 0.7. Although this is not sufficiently reliable in theory, studies have claimed these low-reliability values as satisfactory. Jönsson and Svingby (2007) indicated that these claims arise from the difference between low-stake and high-stake assessments. They explain that for high-stake assessments, reliability can be seen as a prerequisite for validity. However, for low-stake assessments, this is not necessarily true since the decisions made based on assessment can be easily changed if they appear wrong. Thus, for low-stake assessments, the low levels of reliability might be accepted as sufficient. For high-stake assessment, on the other hand, reliability is an important aspect and thus needs to be high to be sufficient. In the context of this study, and in line with Ursachi et al. (2015), reliability values of 0.8 and greater are found to be good.

While studies may claim a sufficient reliability, the reliability estimates remain theoretically too low for traditional testing and would thus not suffice for high-stake assessments. Coenen et al. (2018) have stated that the higher the stakes of the assessment, the more damaging an unreliable assessment can be. It would be undesirable for an organisation to base educational decisions such as passing or failing an exam on unreliable assessments.

Advantages and Disadvantages

The use of rubrics has several advantages for both candidates and raters. Rubrics offer a clear and detailed framework for assessment, which offers candidates transparency of the assessment process (Schlitz et al., 2009). Rubrics, by definition, include descriptions of candidate performance at different levels of quality and studies have shown that this transparency improved candidate performance by reducing anxiety (Panadero & Jönsson, 2013). Moreover, the specificity of rubrics helps to minimize inaccurate scoring or raters making biased judgements (Gantt, 2010; Shipman et al., 2012). Lastly, rubrics offer a standardized method of grading, which might come in useful when multiple raters assess candidate performance (Knight et al., 2010). However, Steedle and Ferrara (2016) have noted that the preparation, calibration, and monitoring of scores for RBA is a time-consuming, cognitively demanding, and resource-intensive process. Moreover, although the detailed rubric should eliminate bias, studies found that raters have different views regarding what constitutes an acceptable answer, which leads to unreliable scores (e.g., Gantt, 2010; Jones et al., 2019; Knight et al., 2010). Furthermore, as described by Pollitt (2004), when raters assess performance against a rubric, it is likely that raters remember other performances and compare the new performance to them. However, these other performances are unlikely to be truly representative of the rubric and may vary for different raters. Therefore, it might be difficult for raters to assess candidate performance based on RBA.

Considering these ambiguities, an alternative and more reliable form of assessment method might be more suitable for high-stake assessments.

Comparative Judgement

Law of Comparative Judgement

As mentioned shortly in the introduction, CJ is based on Thurstone's Law of Comparative Judgement (Thurstone, 1927). CJ draws on the psychological rationale that humans are better at comparing objects than rating isolated objects. That is because with CJ a rater makes relative judgements instead of absolute judgements. Thurstone (1927) explained that someone perceiving a phenomenon will assign it a 'value'. When comparing this instance to another phenomenon to choose the 'better' one, it is the two values that are compared.

Take the following example as outlined by Kimbell (2012). If one imagines oneself in a house in which the rooms differ in temperature, one would have no difficulty stating that room *x* is relatively warmer than room *y* (assuming that it was). In contrast, when asked to determine the exact temperature on an absolute scale, this would be a lot more difficult. The same goes for educational assessment: a teacher is better at determining candidate A reads better than candidate B than to say candidate A reads at level 4.

For educational measurement purposes, Pollitt (2004) reasoned that when two candidate performances are compared, the standard of the rater will be cancelled out. Kimbell (2021) illustrated this in the following example. Take a lenient rater, when exposed to two candidate products, the rater might think both products are thorough, but one of the two products as better. Similarly, a strict rater might think both products are not thorough, but still, one product will be more thorough than the other. In both instances, and despite personal standards, the more thorough performance will be identified as the better one (Kimbell, 2021). Thus, the personal or internalised standard will be cancelled out (Pollitt, 2004; Pollitt, 2012). It is through this reasoning that CJ can be described as an objective relative measurement method that will construct a true measurement scale depicting the relative true value of the candidate's performances (Pollitt, 2004).

Process of Comparative Judgement

To assess candidate performance with CJ, raters make individual holistic judgements of dichotomous pairs of candidate products. The term product refers to any object that a candidate has submitted as proof of achievement. For each pair, the rater must indicate which candidate product is 'better', which is then translated in a binary decision matrix for each pairing outcome. Each candidate product is compared several times and by multiple raters. If many comparisons are made, this matrix can be fitted in the Bradley-Terry model (Bradley & Terry, 1952), which is an often-used model for pairwise comparisons. The model will generate a rank order scale of candidate products, which will rank from 'worst' to 'best' candidate performance. Because each candidate product is compared by multiple raters, this model will represent a shared consensus (Coertjens et al., 2021; McMahon & Jones, 2015; Pollitt, 2004; Pollitt, 2012; Steedle & Ferrara, 2016). In order to translate the rank order scale to grades, standard setting needs to be applied to determine the pass/fail boundary (Coertjens et al., 2021).

Reliability of Comparative Judgement

One of the most important advantages of CJ over RBA is the increased reliability. Often reliability measures estimate the extent to which differences in the observed scores can be attributed to differences in their true score. However, when CJ is used as an assessment method, each candidate performance is measured once and multiple raters assess each observation (Steedle & Ferrara, 2016). Therefore, the reliability of CJ will reflect the consistency between raters' perceptions of relative quality, in other words, the inter-rater reliability.

CJ uses the scale separation reliability (SSR) which is expressed as:

$$SSR = \frac{\sigma_\beta^2}{\sigma_v^2} \tag{1}$$

where σ_β^2 and σ_v^2 are the variance of the true scores (β) and the observed scores (v), respectively.

From CTT it is known that the variance of the true scores can be estimated from the variance of the observed scores using:

$$\sigma_\beta^2 = \sigma_v^2 - MSE \tag{2}$$

where MSE is the mean squared error. The σ_v^2 and MSE can be calculated from the person parameters and their standard errors of estimation, respectively. Thus, the SSR can be calculated with available software via:

$$SSR = \frac{\sigma_v^2 - RMSE^2}{\sigma_v^2} \tag{3}$$

Applicability of Comparative Judgement

CJ has mostly been applied to assess competences that are too complex to assess through one aspect. However, CJ has also been applied for assessing single competences. In both instances, CJ has been found to be a reliable and valid method of assessment (Jones et al., 2015; Pollitt, 2012; Steedle & Ferrara, 2016). However, there is one thing to consider when applying CJ. Raters might view CJ as a complex assessment method when exposed to products with similar quality (Benton, 2021; Gijsen et al., 2021; Van Daal et al., 2017). In CJ, raters must decide which of the two products is the 'better' one, which is a difficult task if two products are similar or even identical. In their study, Van Daal et al. (2017), found a negative relationship between rank-order distance and experienced complexity for accurate decisions. When the rank order distance between two candidate products increases, the experienced complexity for raters decreases accordingly. Thus, if two products differ more in quality, raters experience less uncertainty in judging the ‘better’ product. Similarly, for products that are similar in quality, it becomes more difficult for raters to select the ‘better’ product.

Research Questions and Hypotheses

Most studies have investigated the use of assessment through CJ compared to the same test or item assessed with RBA. Based on the results, the reliability of CJ was compared to the reliability of RBA and these studies have indicated that CJ is a reliable and valid form of assessment (Jones et al., 2015; Pollitt, 2012; Steedle & Ferrara, 2016). However, no studies have been found that investigated the use of CJ for items with varying maximum scores.

A test often consists of multiple items that form a balanced exam in which some items assess higher levels of understanding and other items that assess lower levels of understanding (Swart, 2010; Ushiro et al., 2008). Combining these items helps raters to gain insights into a candidate's ability. According to Jayakodi et al. (2015), items that assess lower levels of understanding can be considered as easier items than items that assess higher levels of understanding. While lower-level items may be relatively easy to answer, for higher-level items candidates must integrate a wider range of information to correctly answer the item. To account for the differences in complexity, tests often include several items for which the maximum scores differ according to the level of understanding that is needed to answer correctly. To investigate the reliability and efficiency of CJ for items with varying maximum scores, the following research questions (RQ) with hypotheses (H) have been formulated.

RQ 1. How do items with varying maximum scores influence the reliability of CJ?

RQ 2. How do items with varying maximum scores influence the efficiency of CJ?

RQ 3. To what extent are the rank orders of the scores assessed through CJ and RBA correlated for items that vary in maximum scores?

When raters assess a candidate's performance on a test, they must assess answers to an exam that consists of both lower-level items and higher-level items. As the lower-level items can be considered to be easier (Jayakodi et al., 2015), these items assess a smaller range of candidate ability than higher-level items. Newton (1996) has stated that it might be more difficult for raters to discriminate accurately between candidate products when the range of candidates' abilities is lower. This is in line with Jones et al. (2015), who mentioned that when more information is provided by candidates, this might inform raters better on candidates' ability. Since for higher-level items with a higher maximum score, candidates must integrate a wider range of information, it might be easier for raters to assess these answers. Similarly, for lower-level items, a smaller range of ability is assessed thus increasing the difficulty for assessment.

Furthermore, as for lower-level items the range of candidate ability is lower, it may occur for two or more candidates to give the correct answer. While through RBAs, these candidates would all be awarded the same score, in CJ the rater must decide which candidate's answer is the ‘better’ one. This might be difficult, or even impossible if the answers are similar or identical (Gijsen et al., 2021; Van Daal et al., 2017).

H 1. It is hypothesised that the reliability of CJ will be higher for items with a higher maximum score compared to items with a lower maximum score because raters can discriminate more easily for high-level items. It is expected that if raters can discriminate more easily, the raters will choose the same candidate answer as the ‘better’ one, thus increasing the reliability. For lower-level items, raters might experience more difficulty deciding on the ‘better’ answer, which is expected to result in a lower reliability.

H 2. It is hypothesised that the efficiency of CJ differs for items with varying maximum scores, where items with a lower maximum score take less time to assess. It is expected that since high-level items assess a larger range of ability, the candidate answers will be more elaborate. Since the answers include more information, the answers are expected to be longer and thus the time investment for raters to choose the better answer are higher. Furthermore, it is expected to be inefficient to adopt CJ as an assessment method for assessing exams that consist of multiple items. Since raters go through multiple rounds of comparing dichotomous pairs, it is expected that the time investment for assessing candidates’ answers for a single item through CJ is higher than assessing the same answers through RBA.

H 3. It is hypothesised that for high-level items the rank scale orders of CJ are similar to the rank orders of RBA. In contrast, for low-level items, it is expected that the rank orders are less correlated. This expectation is based on the assumption that raters experience more difficulty assessing the lower-level items which might result in rank orders that correlate less with the rank orders of the same products assessed through RBA.

RQ 4. What are the practical implications of assessing high-stake exam items through CJ?

H 4. Despite the promise of CJ, it is expected that it is difficult to adopt CJ as an assessment method for items that vary in maximum scores. When every single item needs to be assessed through CJ, this is expected to be inefficient. Moreover, since the research concerns high-stake exams, it is expected that CJ’s lack of transparency will be problematic. When assessing answers through RBA, raters can refer to the rubric when awarding scores for answers. However, when CJ is applied, these judgements are made compared to other candidates, which may be less transparent.

Method

Research Design

This study had a quasi-experimental design in which the answers to items of the Bestuurlijke Informatievoorziening (BIV®) exam with maximum scores of 2, 3, and 4, were assessed through CJ. The answers to these items had already been assessed through RBA. This study aimed to clarify the efficiency and reliability of CJ for items that vary in maximum scores that can be awarded as compared to that of RBA.

For this study, a selection of three items and $n = 15$ respondents of the BIV® exam were used, made available by the AvE. The BIV® exam consists of items to which the maximum scores vary between 1 to 4 points. CJ could not be applied to the items with a maximum score of 1 since these items were multiple-choice questions. For multiple-choice questions, it follows logically that CJ cannot be applied since an answer is either correct or incorrect. Moreover, the multiple-choice items are automatically assessed, meaning that for multiple-choice items there will be no inconsistencies in rater scoring.

For the items with maximum scores of 2, 3 and 4, three assessments were created in the Digital Platform for Assessing Competencies (D-PAC) tool, where each assessment represented one item. While the items of the BIV® exam were carefully selected as will be elaborated upon later, random sampling was used to select $n = 15$ respondents for each assessment. The respondent's answers, which are also called products, were selected for each of the three items. The products were uploaded in the assessments and raters were asked to go through $n = 57$ comparisons for which they had to select the 'better' product. Afterwards, the Bradley-Terry-Luce model was automatically applied to generate a rank order of the candidate products. The reliability of all three assessments was calculated using the SSR, and the rank correlation between RBA and CJ was analysed using the Kendall Rank correlation.

Instrumentation

Various instruments were used in this study: the BIV® exam, specific items from the BIV® exam, and the D-PAC tool for assessment through CJ.

BIV® Exam

As the BIV® exam was designed to assess whether candidates possess the skills to properly inform management about internal affairs, reliability of management information plays an important role. Incomplete or incorrect information might lead to wrong policy decisions. Therefore, insight into the risks that threaten reliable management information is an important skill to possess.

For the BIV® exam, no preliminary education is needed. The BIV® is an online examination that can be taken at multiple locations in the Netherlands. It is an exam that consists of both multiple-choice and open-ended items. The time allocated to finish the exam is 120 minutes.

In previous years the BIV® exam has been taken by $n = 148$ candidates from the time that the examinations started in January 2018 up until October 2021. From these candidates, a total of $n = 117$ candidates passed the exam which entails a success rate of 79%. For the BIV exam, candidates need to obtain at least a 5.5 to be able to pass the exam which corresponds to obtaining 57% of the total number of points.

Items of BIV®

For this experiment, three items were selected from the BIV® exam. The item bank of the BIV® exam consists of a total of $n = 81$ items. Within this item bank, the items with a maximum score of 1 denote 43%, items with a maximum score of 2 denote 6%, items with a maximum score of 3 denote 16%, and items with a maximum score of 4 denote 35%.

To evaluate the quality of an item, the p-value, as well as the Rit-value, must be considered. The p-value denotes the proportion of candidate's that answer an item correctly. The Rit-score indicates how the score of an item is related to the total exam score. An elaborate framework on item quality can be found in Appendix A. Since candidates must obtain 57% of total points to reach the passing grade of 5.5, the item's ideal p-value lies around 0.57. Additionally, the higher the Rit-value, the better the item can discriminate between candidate's ability. Based on these considerations, the items as outlined in Table 1 have been selected from the BIV® exam.

Table 1

Selected Items of the BIV® Exam

Item	Max. Score	P-value	Rit-value
452532.1	2	0.63	0.41
452495.1	3	0.56	0.53
452515.1	4	0.57	0.68

D-PAC Tool

For the assessment of candidates' exams through CJ, the web-based tool D-PAC was used. To assign the comparisons to the raters, the D-PAC tool used a quasi-random method. The algorithm of the D-PAC tool has two checks. First, to ensure each candidate is compared an equal number of times, the algorithm identifies the candidates that have been least compared and draws a candidate randomly from this group. Secondly, the algorithm selects the second candidate from the candidates that the first candidate has not yet been compared to, to ensure no duplicate pairs will be compared.

The D-PAC tool keeps track of the comparisons per rater as well as how many products have been compared. This way, the evolution of the SSR could be calculated in Rstudio

(RStudio Team, 2021). Moreover, the time spent can be used to determine how items with varying maximum scores influence the efficiency.

According to Comproved (n.d.), the guideline for a reliable assessment is 15 to 20 assessments per product. As there was a selection of $n = 15$ exams; this entails a total of $15 * 15 = 225$ products had to be assessed. Since the examinations could be assessed in pairs, this entailed $225/2 = 112.5$ comparisons. Lastly, as there were two raters, each rater had to assess $n = 57$ pairs.

Although the D-PAC tool has a function that as for raters to provide feedback for the comparisons, this option was disabled. The reasoning behind being that if raters would have to provide feedback for each of the $n = 57$ comparisons, this would slow down the process of assessment.

Participants

The participants that were involved in this research were two raters. These raters assessed the candidate products through CJ for each of the three items that differ in maximum scores. The raters are the same raters that assess the current exams of BIV® through RBA. This ensured that the raters are familiar with the competences that are assessed through the BIV® exam.

Procedure

As this research concerned human participants, the research was submitted to the Ethics Committee. This research has been approved by the ethics departments of the Behavioural, Management, and Social Sciences (BMS) of the University of Twente under request number 211180.

Prior to the research, all raters had been informed about all aspects of the research to enable them to make an informed decision on their willingness to participate. Therefore, all raters had to sign an informed consent form. In this consent form, the purpose of the research was described as well as the risks of participating. Moreover, it was stated that participation is voluntary, and participants can withdraw at any given moment. To protect the participant's privacy, personal data has been anonymised using pseudonyms such as rater1. Lastly, the consent form included the contact details of the researcher as well as the contact details of the BMS Ethics Committee of the University of Twente.

Due to the Covid-19 regulations, this experiment took place online. A first online meeting was conducted to explain CJ, the research, and the D-PAC tool. Hereafter, raters were exposed to a test assignment as a means of training. This training was implemented to ensure raters are familiar with the D-PAC system and the rationale of CJ. Before the actual assessment started, the raters first had to look at the guidelines that candidates received for taking the BIV® exam. Raters additionally received the competence description, detailing the same dimensions as the rubric used for the RBAs. Each rater had a period of two weeks to complete the assessments. This timeframe allowed for the raters to pause and continue the assessments whenever convenient. During the assessments, each rater had to keep in mind 'which answer is better?'.

After the experiment, a second online meeting was planned. During this meeting, time was allocated for a discussion among the raters on their experiences regarding CJ, D-PAC tool, and the possible practical implications CJ for the AvE.

In addition to the discussion among raters on their experience with CJ, an interview was conducted with the programme director of the AvE to gain insights into the applicability of CJ and its practical implications.

Data Analysis

The D-PAC tool analysed the comparisons using the Bradley-Terry-Luce model, which is a frequently used model for pair comparisons. This model generated a rank order of the CJ assessments from lowest to the highest score on the item.

The reliability of CJ was determined by calculating the SSR. The SSR was calculated each time a comparison is made. The higher the SSR, the higher the level of agreement of the raters on the position of the candidate's answer in the rank order. In addition to calculating the SSR, the evolution of the SSR was analysed. The D-PAC tool records the time spent per rater and the time spent per comparison. This allowed for data analysis to create an evolution graph of the SSR for each round of comparisons.

Lastly, both CJ and RBA measure candidate performance on the items. Therefore, it was relevant to examine the correlation between both rank orders. The Kendall rank correlation, or Kendall's τ coefficient, is a rank correlation coefficient that evaluates the degree of similarity between two sets of ranks. The Kendall's τ coefficient can be calculated via:

$$Kendall's\,\tau = \frac{C - D}{C + D} \tag{4}$$

in which C represents the concordant pairs, and D represents the discordant pairs. A concordant pair resembles the number of observed ranks below a particular rank that are *larger* than that particular rank and a discordant pair represents the number of observed ranks below a particular rank, that are *smaller* than that particular rank (Nelsen, 2002).

The Kendall rank correlation coefficient was used to measure the ordinal scale association between RBA and assessment through CJ to evaluate the degree of concordance between the two assessment methods.

Results

This section will elaborate on the results of the experiment. For each section, the results of the three items with maximum scores 2, 3, and 4 will be shown to allow for easy comparison. The colour palettes used in the graphs have been specifically selected as they are colour-blind friendly.

Rubric-Based Assessments Rank Order

For each of the three assessments, the rank order scale of the candidate products assessed through RBA has been determined. In selecting the candidate products, random sampling was applied to ensure the $n = 15$ products form a representative candidate pool. Although the candidate products for each assessment are represented by letters A through O, these candidate products are not related. Random sampling was applied for each of the three assessments, so the letters A through O denote different candidate products for each assessment.

Within the product pool, multiple candidates had received the same score through RBA. Therefore, the order of the products with identical scores has been randomly determined.

Figure 1 visualizes the rank order scale of the candidate products assessed through RBA for the item with a maximum score of 2. In contrast to CJ where a candidate's score is based on the rank order, the score for candidates assessed through RBA is identical to the number of points awarded to the candidates. For this assessment, seven candidates received the maximum score, four candidates received one point, and four candidates received zero points.

Figure 1

RBA Candidate Rank Order for Item With Maximum Score 2

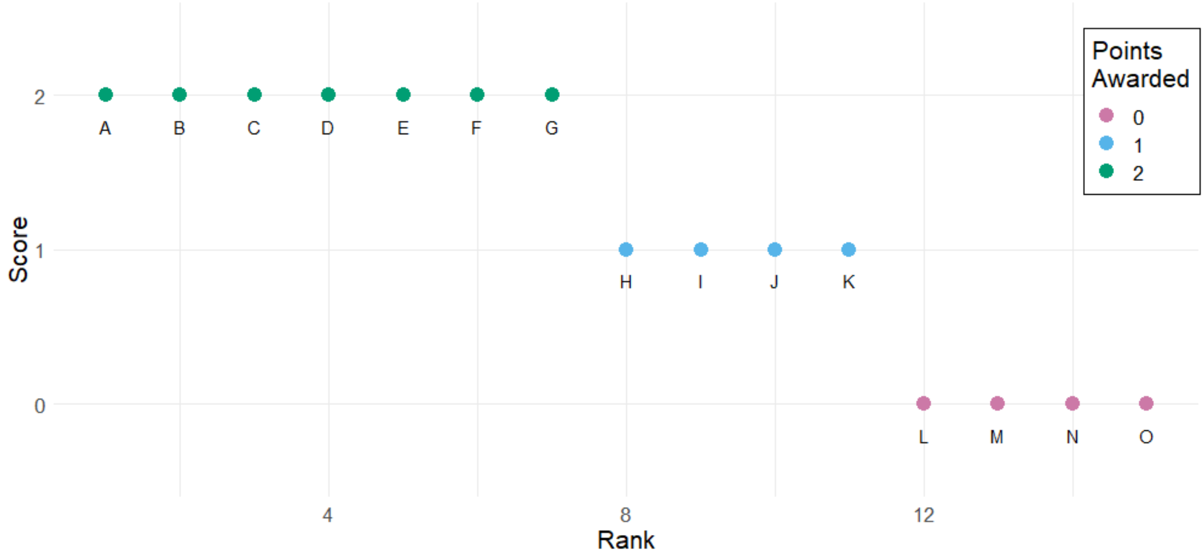
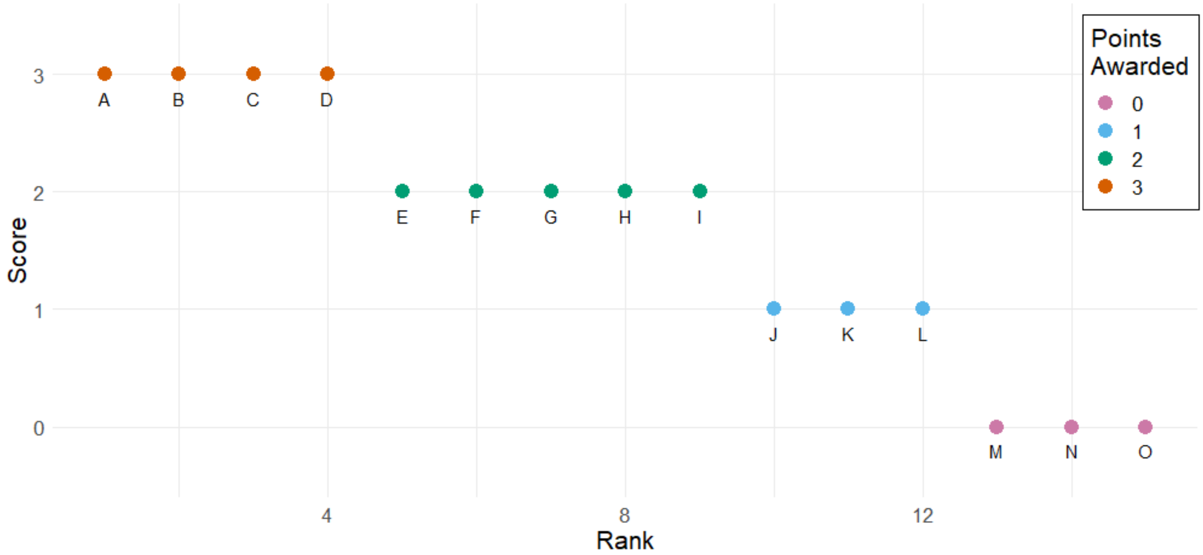


Figure 2 visualizes the rank order scale of the candidate products assessed through RBA for the item with a maximum score of 3. With RBA, four candidates have received the maximum score, five candidates received two points, three candidates received one point, and three candidates received zero points.

Figure 2

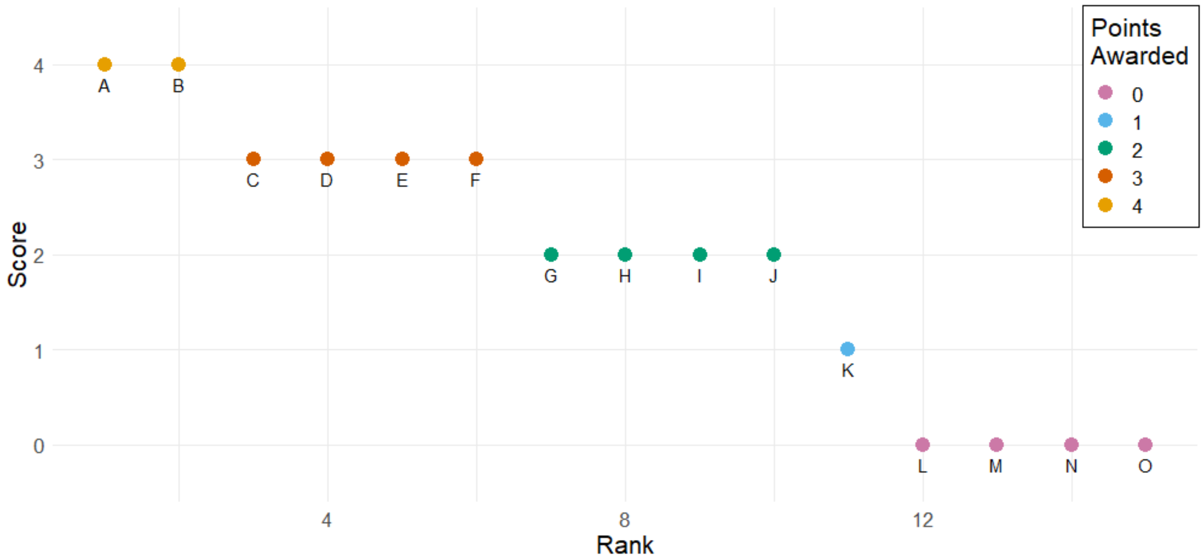
RBA Candidate Rank Order for Item With Maximum Score 3



Lastly, Figure 3 visualizes the rank order scale of the candidate products assessed through RBA for the item with a maximum score of 4. Although random sampling had been applied to form a representable candidate pool, only one candidate (i.e., candidate K) had been selected that received one point when assessed through RBA. Moreover, only two candidates received a total of four points (i.e., candidates A and B).

Figure 3

RBA Candidate Rank Order for Item With Maximum Score 4



Bradley-Terry-Luce Model

The CJ rank order scale of the candidate products for each assessment has been created using the Bradley-Terry-Luce model. This model produced a rank order from 'best to 'worst' candidate for the $n = 15$ products for each assessment. Hereafter, based on the rank order, the score for each product could be assigned. This script that the D-PAC tool used, needed a certain level of freedom. It was recommended to select two products: one product on 1/3 and one product on 2/3 of the rank order scale. The score's minimum was set to zero, and the maximum score was set corresponding to the item's maximum score. Lastly, based on the rank scale position and corresponding score, the number of points the candidate would receive was determined. The points were rounded up or rounded down to the nearest whole integer.

Figure 4 visualizes the rank order scale of the candidate products assessed through CJ for the item with a maximum score of 2. As can be seen, the method of CJ has yielded nine candidates that were awarded the maximum score of two points, while only one candidate received zero points (i.e., candidate O).

Figure 4

CJ Candidate Rank Order for Item With Maximum Score 2

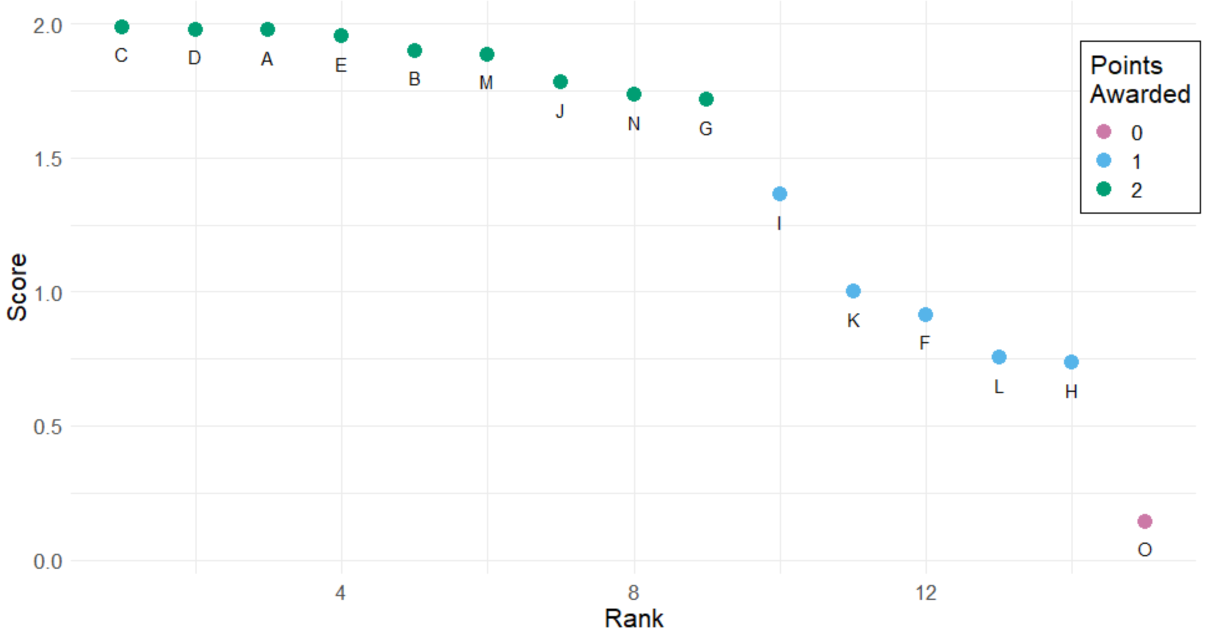


Figure 5 visualizes the rank order scale of the candidate products assessed through CJ for the item with a maximum score of 3. In this graph, ten candidates have been awarded the maximum score of three points, two candidates received two points (i.e., candidates H and M), one candidate received one point (i.e., candidate J), and two candidates received zero points (i.e., candidates O and N).

Figure 5

CJ Candidate Rank Order for Item With Maximum Score 3

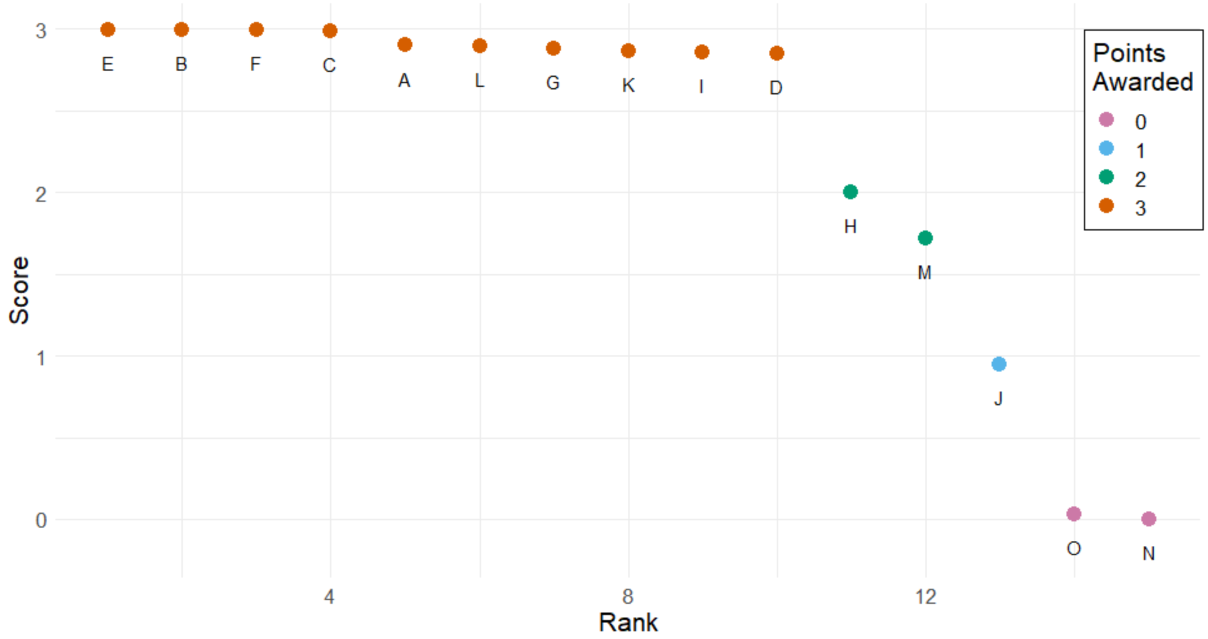
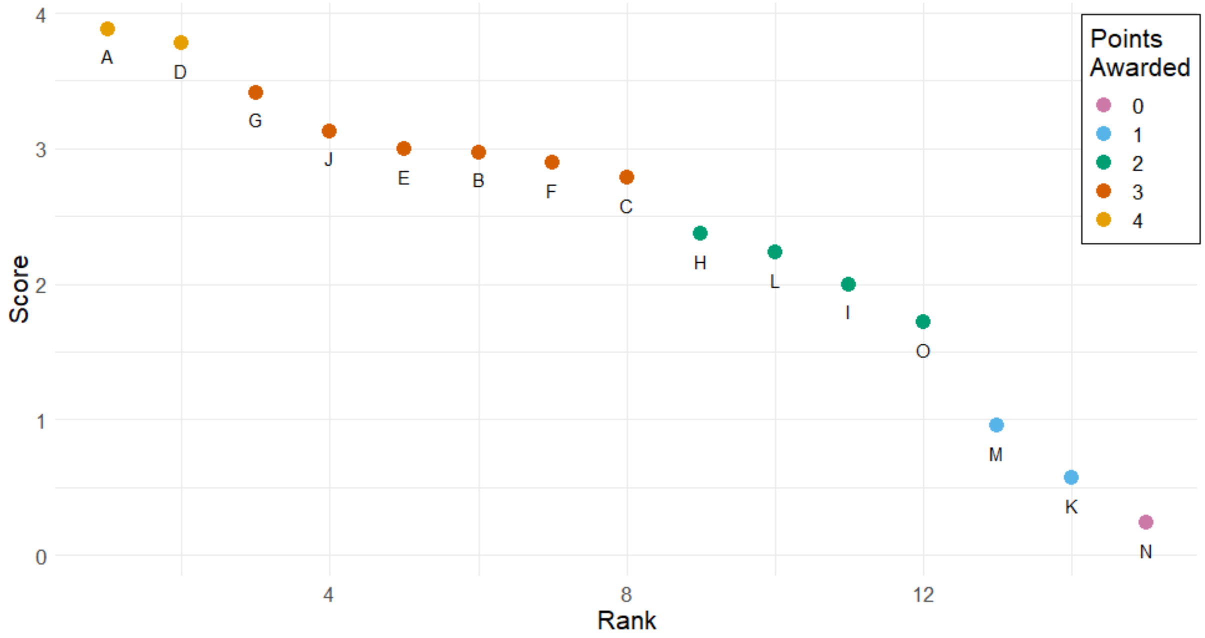


Figure 6 visualizes the rank order scale of the candidate products assessed through CJ for the item with a maximum score of 4. In this graph, the scores seem to be balanced throughout the rank order. However, there is only one candidate (i.e., candidate N) that had received zero points for their answer.

Figure 6

CJ Candidate Rank Order for Item With Maximum Score 4



Kendall Rank Correlation

In this section, the correlation between rank orders of candidates assessed through RBA and CJ is analysed.

Figure 7 visualizes the correlation between points awarded through RBA and CJ for the item with a maximum score of 2. The assessment methods have yielded the same number of points to be awarded for 66.7% of the candidates. However, for three candidates the two assessment methods differ one point from each other (i.e., candidates F, J, and L), while for two candidates the methods differ two points (i.e., candidates M and N). These differences led to a Kendall Rank correlation of $\tau = 0.524$ with $z\text{-}score = 2.722$.

Figure 7

Correlation RBA and CJ Points Awarded to Candidates for Item With Maximum Score 2

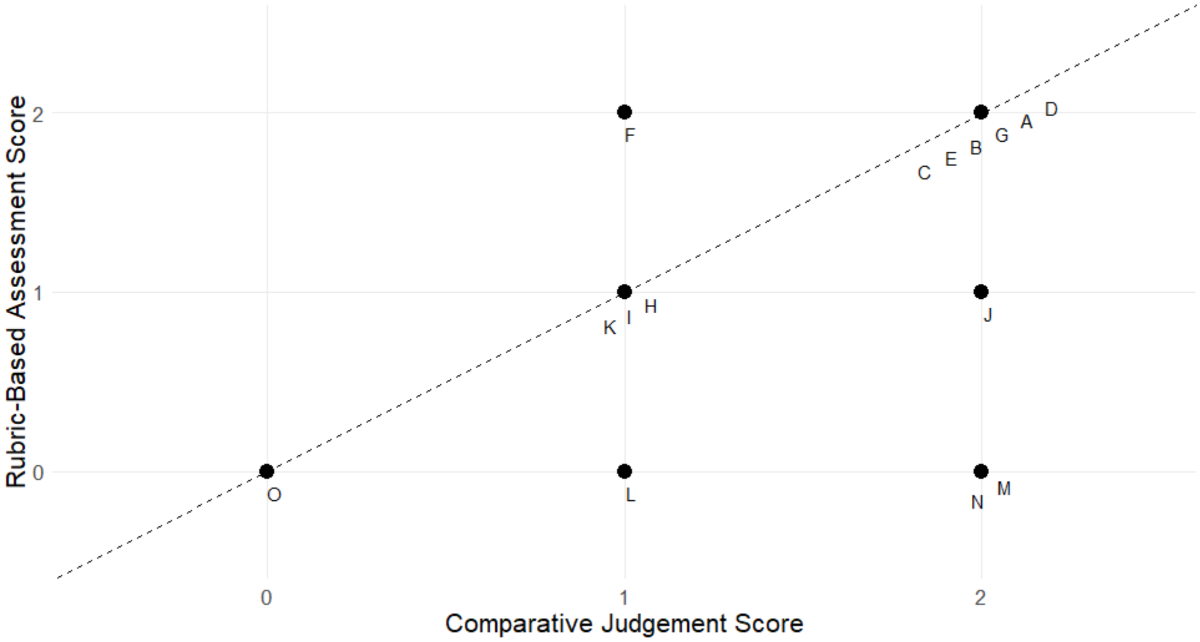
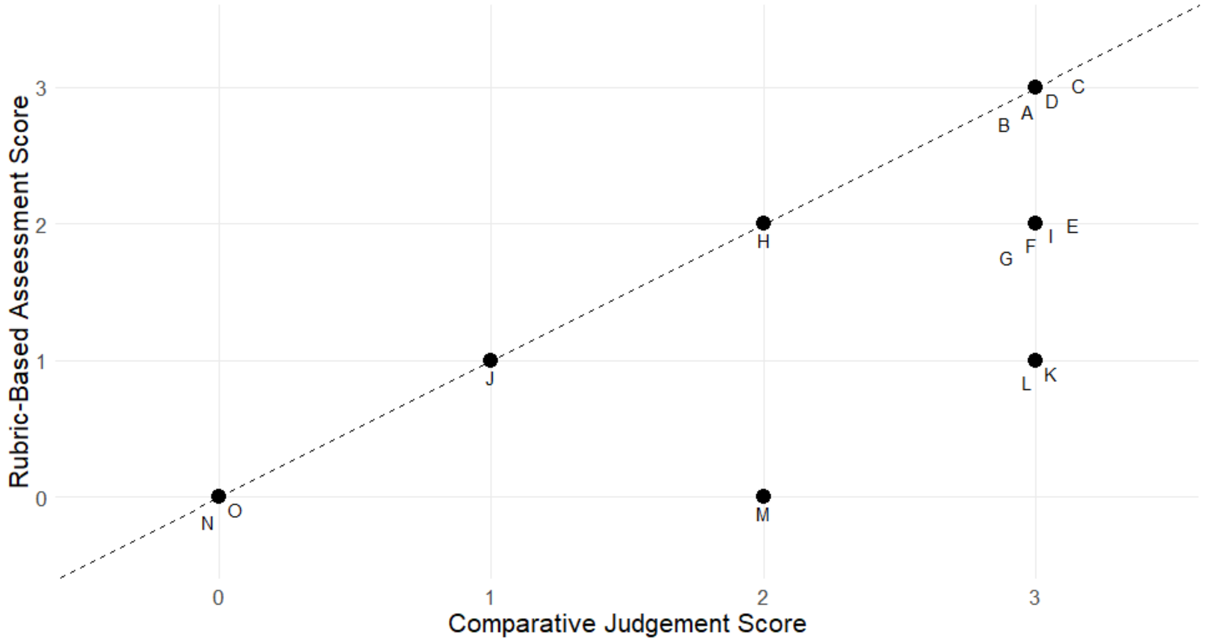


Figure 8 visualizes the correlation between points awarded through RBA and CJ for the item with a maximum score of 3. The assessment methods have yielded the same number of points to be awarded for 53.3% of the candidates. For four candidates, the methods differ one point from each other (i.e., candidates E, F, G, and I) and for three candidates the difference is two points (i.e., candidates K, L, and M). This resulted in a Kendall Rank correlation of $\tau = 0.543$ with $z\text{-}score = 2.821$.

Figure 8

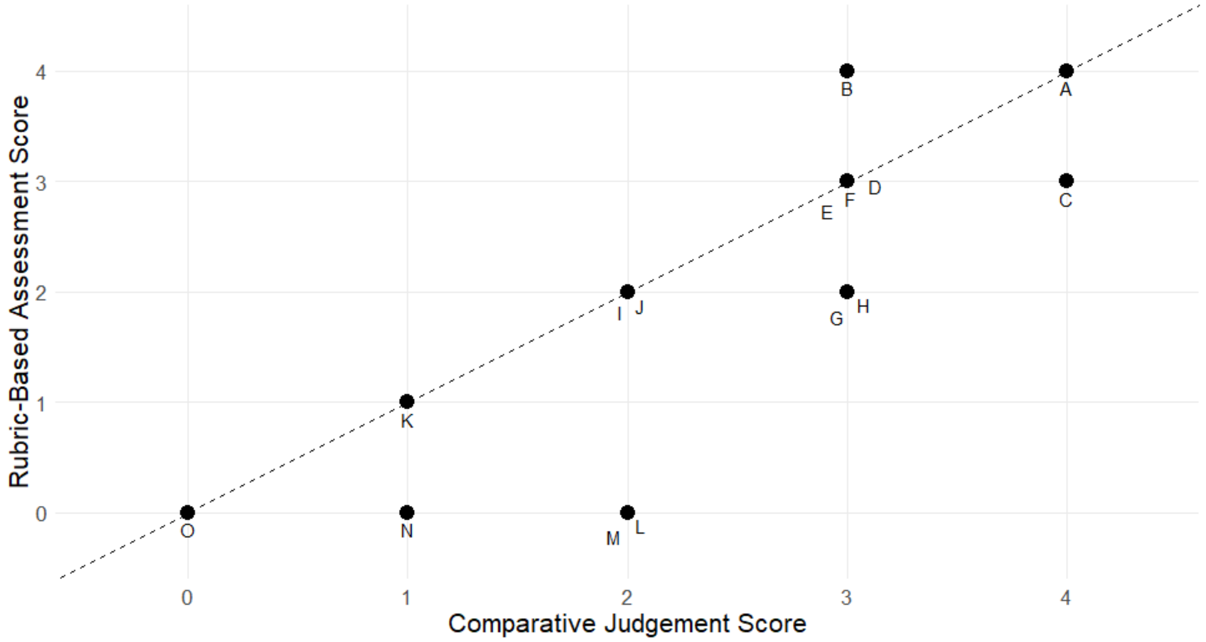
Correlation RBA and CJ Points Awarded to Candidates for Item With Maximum Score 3



Lastly, Figure 9 visualizes the correlation between points awarded through RBA and CJ for the item with a maximum score of 4. The assessment methods have yielded the same number of points to be awarded for 53.3% of the candidates. For five candidates, the methods differ one point (i.e., candidates B, D, G, J, and M) and for two candidates the methods differ two points (i.e., candidates L and O). This resulted in a Kendall Rank correlation of $\tau = 0.615$ with $z\text{-}score = 3.197$.

Figure 9

Correlation RBA and CJ Points Awarded to Candidates for Item With Maximum Score 4



For all three assessments, it can be noted that the maximum difference in points that are awarded to candidates is two points. Moreover, when differences in points awarded occur, CJ yields a higher score than RBA for 88.2% of the candidates.

Evolution SSR

In this section, the evolution of the reliability for each assessment will be analysed. The reliability has been calculated by the SSR each time all products have been compared one additional time. Thus, the first round entails that all products have been compared once by the raters, the second round means that all products have been compared twice and so on. In the first few rounds for each assessment, the reliability was below 0. This is because during these rounds not enough data has been collected to result in a positive reliability. However, since reliability is expressed as a value between zero and one, and to increase interpretability, these graphs focus on the reliability values above zero.

Figure 10 visualizes the evolution of the reliability for the item with a maximum score of 2. In round 4, the reliability exceeded zero and increased to a value of $SSR = 0.86$.

Figure 10

Evolution of the Reliability of Item With Maximum Score 2

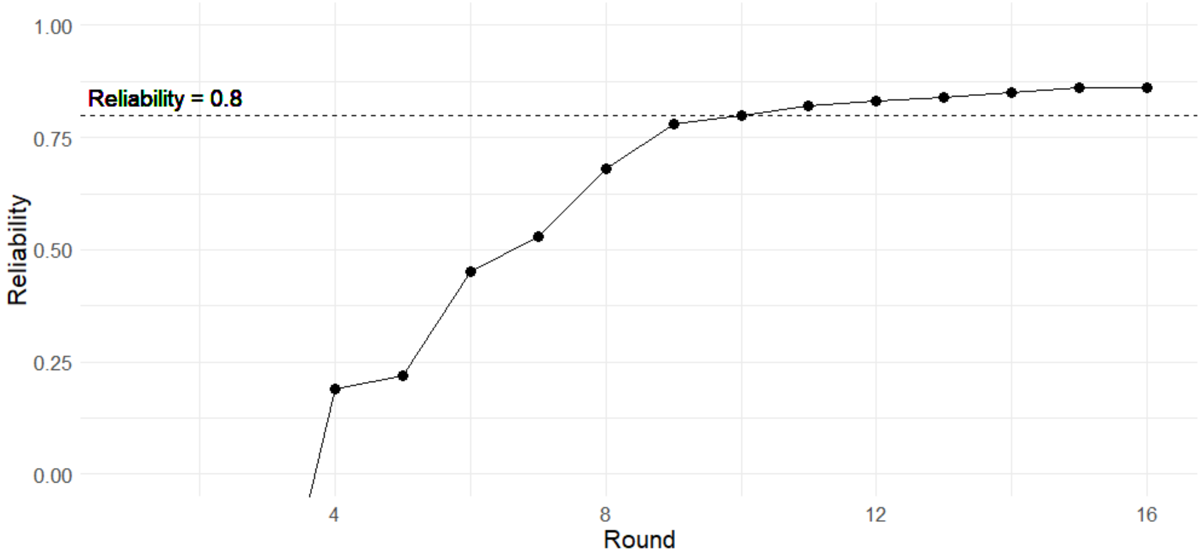


Figure 11 visualizes the evolution of the reliability for the item with a maximum score of 3. The reliability of this assessment exceeded zero in the fifth round, and the reliability had been increasing in value until round 10, when it dropped significantly. Hereafter, the reliability slowly regained in value until $SSR = 0.77$.

Figure 11

Evolution of the Reliability of Item With Maximum Score 3

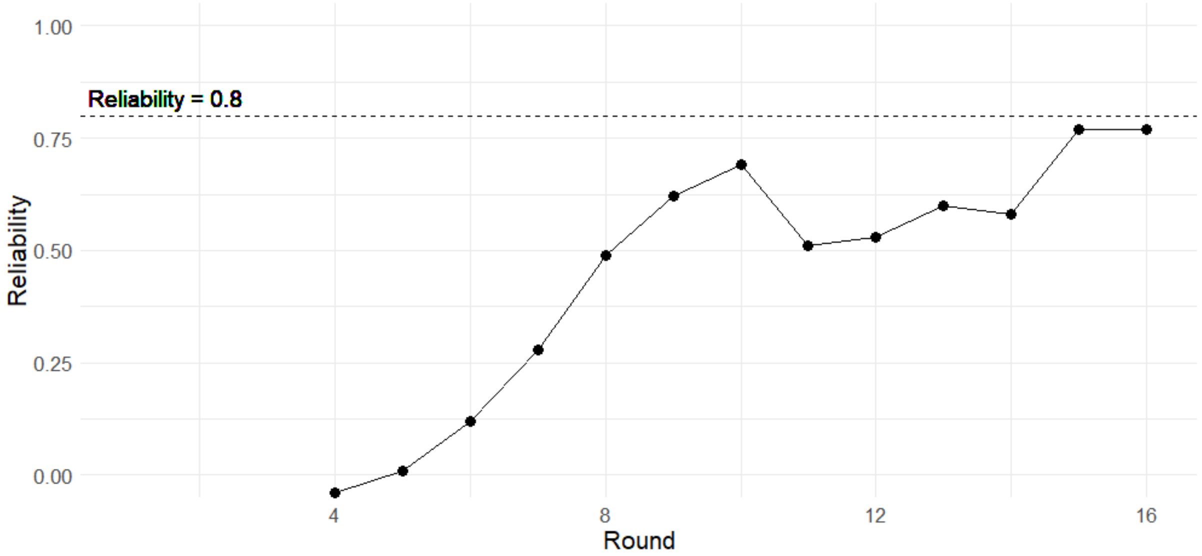
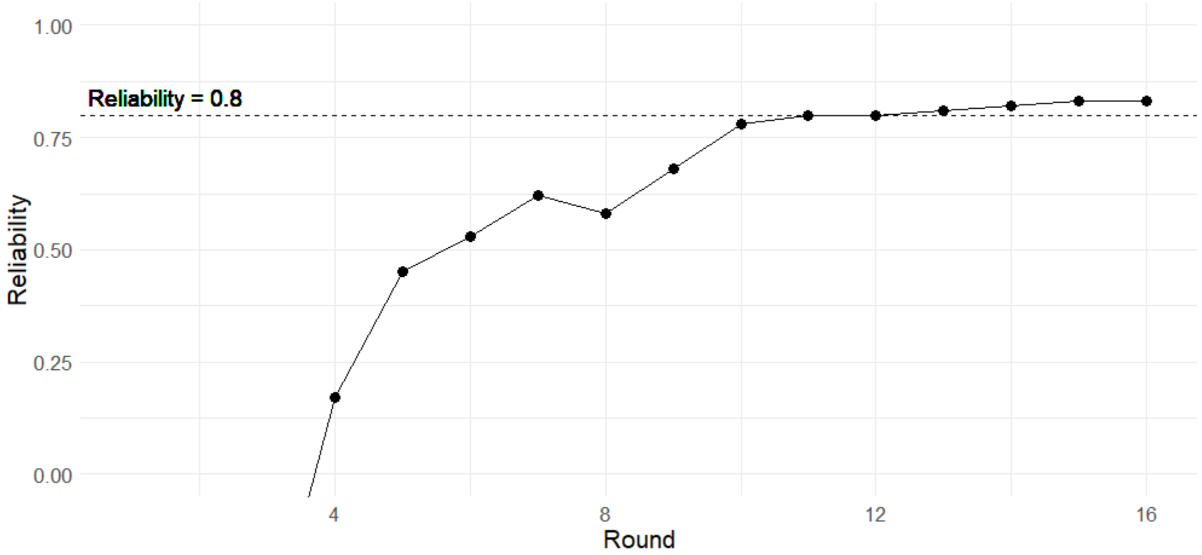


Figure 12 visualizes the evolution of the reliability for the item with a maximum score of 4. For this assessment, the reliability exceeded zero in round four and encountered a small drop in reliability after round seven. Hereafter, the reliability increased again until it reached $SSR = 0.83$.

Figure 12

Evolution of the Reliability of Item With Maximum Score 4



As visualized in Figure 10 and Figure 12, the reliability curve flattens out the more rounds the raters go through for the item with maximum a score of 2 and the item with a maximum score of 4. This entails that the gain for the SSR for these last rounds is small, meaning that if raters would have been exposed to more comparisons, this would not result in a much higher reliability. In contrast, as represented in Figure 11, the reliability curve does not flatten out as clearly for the item with maximum score 3. It is unclear if more comparisons would have resulted in a higher reliability.

Time Investment

In this section, the time investment for the comparisons for each assessment is analysed. Since each rater made $n = 57$ comparisons in each of the three assessments, the data comprised a total of 342 data points. In the data, there were five major outliers where raters took more than 300 seconds to rate the comparisons. This may have been caused by 1) taking small breaks during comparisons, 2) starting for the first time with comparisons and needing to read the rubric, and 3) by pausing the comparisons and resuming the next day. These outliers have been removed.

Table 2 depicts the mean time investments and standard deviation of each rater as well as the average time investment for each assessment. These results indicate that rater1 invested more time than rater2 for assessing the products in each of the three assessments. Moreover, the results indicate that the least time was invested in assessing the candidate products for the item with maximum score 2. The time investment seems to only increase for rater2 and on average when the maximum assessment score increases. That is because the mean time investment for rater1 on assessing the products for the item with maximum score 3 is slightly higher than for the item with maximum score 4.

Table 2

Mean Time Investment (in s) and Standard Deviation per Rater for Each Assessment

Statistic	Rater	Assessment maximum score		
		2	3	4
Mean	Rater1	33.6	55.1	53.0
	Rater2	14.3	24.6	29.2
	Average	24.1	39.8	41.2
Standard deviation	Rater1	22.8	34.9	34.4
	Rater2	9.36	26.8	23.5
	Average	19.9	34.6	31.7

Conclusion

This study explored the applicability of CJ as an alternative to RBA for assessing high-stake exam items that differ in the maximum score. A study was conducted in which two raters assessed candidate products for three items with different maximum scores through CJ. Based on the results, a rank order scale was produced which correlated moderately with the rank order of RBA. These results, in combination with high-reliability scores, suggest that CJ could be a feasible method for assessing candidate products. However, there are some practical implications to consider. In this section, the results of the study will be discussed as related to the RQs. Moreover, the limitations of this study will be elaborated upon as well as some suggestions for future research.

Discussion

This section will elaborate on the results of the study as related to the RQs.

Reliability

Regarding the first research objective, it was hypothesised that the reliability would be higher for items with a higher maximum score compared to items with a lower maximum score. To determine the reliability, the SSR had been calculated for each round for the three assessments. The reliability values for CJ are in line with the reliability values as reported in other studies (e.g., Coertjens et al., 2021; Jones et al., 2015; Keppens et al., 2019; McMahon & Jones, 2015; Van Daal et al., 2017). However, in contrast with the hypothesis, the results indicated that for the item with the highest maximum score the reliability ($SSR = 0.83$) was lower than for the item with the lowest maximum score ($SSR = 0.87$). Moreover, in contrast with the expectations, the item with a maximum score of 3 yielded the lowest reliability ($SSR = 0.77$).

It is unsure as to why the item with a maximum score of 3 yielded the lowest reliability. However, it might be because this item was the first assessment for which the raters made the comparisons. When looking at the data, in round four, the items with a maximum score of 2 and 4 ($SSR = 0.19$, $SSR = 0.17,$ respectively) have already exceeded zero, while the reliability of the item with a maximum score of 3 ($SSR = -0.04$) had not. Although raters went through an example as means of training, it might be that the raters were still unfamiliar with the CJ tool during the first rounds of the first assessment.

Furthermore, the reliability of the item with a maximum score of 3 had a decrease of reliability (0.18) at round 11. It is unsure what caused this drop in reliability. It might be that if the raters had made additional comparisons, the reliability would have increased to similar reliability levels as the other two assessments.

Lastly, it is uncertain what caused the difference in reliability where the item with a maximum score of 2 yielded a higher reliability than the item with a maximum score of 4. While it was expected that it would be more difficult for raters to discriminate between candidate products when lower-level items assess lower ability (Newton, 1996), it is encouraging that the reliability of both assessments is sufficiently high. The difference in reliability might have been caused by a difference in the candidate product sample. While the candidate products as sampled for the item with a maximum score of 2 formed a balanced sample in which candidates received zero, one, and two points, the candidate sample for the item with a maximum score of 4 was disproportionate. Although random sampling had been applied, there was only one candidate that had received one point and only two candidates that had received the maximum score of 4 points. More research is necessary to address the difference in reliability for low-level and high-level items.

It can be concluded that although the reliability values are in line with CJ reliability values as presented in other studies, an increase in item maximum score does not necessarily result in an increase in reliability.

Efficiency

For the second research objective, it was hypothesised that the time investment would be higher for an item with a higher maximum score than for items with a lower maximum score. The time that raters invested in comparing the candidate products was tracked in the D-PAC tool. Based on the results, it can be determined that, in line with the hypothesis, the average time investment increases when the item maximum score increases. However, while the mean time investment for the average as well as for rater2 increases when the maximum score increases, for rater1 the mean time investment for the item with a maximum score of 3 was slightly higher (55.1) than for the item with a maximum of score 4 (53.0). When analysing the time investment data of rater1, it can be concluded that rater1 divided the comparisons over two moments. The first moment, rater1 compared 26.3% of the comparisons with an average time spent of 103.7 seconds. The second moment rater1 completed the remaining 73.7% of comparisons with a lower average of 43.9 seconds. As already mentioned, the item with a maximum score of 3 was for both raters the first assessment in which candidate products had been compared. Thus, the difference in average time investment between the two moments for rater1 may have been caused by inexperience with the D-PAC tool at the first moment, thus increasing the average time spent for comparing candidate products for the item with a maximum score of 3.

It must be noted that to ensure a high reliability, raters had to assess $n = 57$ pairs of candidate products. If more raters would have assessed the candidate products, each rater would have had to compare fewer items, thus decreasing individual time investment. However, even with more raters, it would still have been inefficient to assess candidate products for a single item, which will be elaborated upon later. An additional consideration is that while the tool tracked the time investment of raters, time was also invested regarding the preparation and analysis of the comparisons. The candidate products had to be exported to a usable format, then uploaded in the tool, and based on the candidate rank order the item scores needed to be assigned. When applying RBA, this time investment is not needed.

Therefore, it can be concluded that the time investment for assessing the candidate products increases if the maximum score of the item increases.

Rank Order Correlation

The goal of the third research question was to examine the rank order correlation between CJ and RBA for items that differ in maximum score. It was hypothesised that the correlation for lower-level items would be lower than the correlation for higher-level items. The rank orders for each of the three assessments have been calculated using Kendall's τ rank correlation. The results indicate that the correlation for lower-level items is indeed lower than for higher-level items, where the item with a maximum score of 2 resulted in the lowest correlation ($\tau = 0.524$), and the item with a maximum score of 4 resulted in the highest correlation ($\tau = 0.615$). Each of the three assessments shows a moderate positive correlation, for which Kendall's τ increases slightly for each item with a higher maximum score.

As mentioned, since multiple candidates received the same score through RBA, the order of these candidate products in the rank had been randomly determined. The analysis of the results revealed that the rank correlation would have been higher if responses of CJ had been paired more optimally with the responses of RBA. By pairing the responses more optimally, the rank order of the products assessed through RBA would have been determined based on the lowest difference in rank order when compared with the products assessed through CJ. When calculating Kendall's τ rank correlation, the difference in concordant pairs and discordant pairs is calculated. Thus, when rearranging the RBA rank order to pair more optimally with the CJ order, the correlation value will increase. The differences in correlation between random pairing and optimal pairing has been added to Table 3. The results of optimal pairing are *not* in line with the expectation, since the item with a maximum score of 3 has yielded the lowest correlation ($\tau = 0.676$), and the item with a maximum score of 4 yielded the highest correlation ($\tau = 0.733$).

Table 3

Comparison of the Correlation between Random Pairing and Optimal Pairing

	Kendall's τ		Z-Score	
Assessment	Random pairing	Optimal pairing	Random pairing	Optimal pairing
Maximum score 2	0.524	0.712	2.722	3.697
Maximum score 3	0.543	0.676	2.821	3.514
Maximum score 4	0.615	0.733	3.197	3.811

The optimal correlation values are in line with other correlation values as reported in literature such as Coertjens et al. (2021) who have reported a correlation of $\tau = 0.74$. However, other studies have reported similar or even higher correlation values by reporting the Spearman ρ correlation (e.g., Jones et al., 2015; McMahon & Jones, 2015). The Spearman and Kendall rank correlations are both suitable for ordinal data. However, as concluded by Puth et al. (2015), Kendall's τ correlation produces narrower confidence intervals and might be preferred over Spearman's ρ when there are no ties in the data.

It must be noted that the small sample size could have affected the correlation. For small sample sizes, the correlation values can be substantially affected by any changes in scores (Goodwin & Leech, 2006). If the sample size would have been larger, the differences in rank position would have had less impact on the correlation, thus resulting in higher correlation values.

Due to the differences between random pairing and optimal pairing, and the small sample size, no definite conclusion can be drawn regarding the rank order correlation between RBA and CJ for items with varying maximum scores. This should be investigated in further research.

Practical Implications

Besides determining the reliability, efficiency, and rank order correlation, the last research objective was to clarify the practical implications of CJ for assessing high-stake exam items. This study revealed that in addition to the expected inefficiency of assessing single items through CJ and the issues with transparency, the candidate pools of low-ability candidates, difficulties with similar candidate products, and the minimum number of candidates are shortcomings of CJ for high-stake item assessment as well. This section aims to elaborate on these limitations.

Assessing Single Items. It was already expected that applying CJ to assess a candidate's performance on single items was inefficient. Although the results have indicated that with regards to reliability and rank order correlation CJ would be a feasible assessment method, the total time investment for assessing answers through CJ was experienced as too high. In contrast to RBA, where raters go through a candidate's answer once, in CJ multiple comparisons need to be made to produce a reliable rank order. The raters have indicated that CJ will take about two to three times longer than assessing the exam through RBA (Rater1, Rater2, personal communication, December 14, 2021). Therefore, applying CJ for assessing candidate's performance on single items would be impractical.

Transparency. Another limitation of CJ for high-stake assessments is the lack of transparency towards candidates. If the items are assessed through CJ, it is difficult for raters to inform candidates as to how their score has been determined. In contrast to RBA where raters can refer to the rubric dimensions, in CJ candidate performance has been determined holistically and in comparison with other candidates, which makes it difficult for raters to argue how the score of the candidate has been determined. It is not practical for raters to go through all comparisons in order to elaborate how a candidate has performed compared to other candidates, and it is also not practical to eliminate the candidates' right to appeal their scores. Thus, for high-stake assessments, CJ lacks transparency and RBA is better suitable.

Candidate Pools. In addition to assessing candidates' answers to single items, and CJ's lack of transparency, another limitation is the possibility of having a low ability candidate pool. While this study applied random sampling to form a representable candidate pool, in practice it might occur for the candidate sample to consist of low ability candidates. In that case, the best performing candidate will have the highest rank position, resulting in receiving the maximum score of the item. However, if this candidate would have been assessed through RBA, this candidate might have received a much lower score. Thus, in CJ the maximum score is awarded to the best-ranked candidate of the pool, which not necessarily means the product has a high quality. Rather, the product is ranked highest relative to all products in the candidate pool. In contrast, in RBA the maximum score is awarded *only* when the candidate has included all dimensions of the rubric.

Similarly, a candidate pool can consist of high-ability candidates. Then, a candidate might receive no points because the candidate's answer is 'worse' relatively to the other candidates. In RBA the candidate may have received the same number of points as other candidates, but in CJ the candidate's score is dependent on how other candidates have performed. Thus, CJ is biased towards the differences in the candidate pool sample.

To counteract this limitation, it is necessary to have a large candidate pool or a candidate pool that is representative of all candidates. However, the simplest solution might be to assess candidate products individually, not based on the performance of other candidates.

Similar Candidate Products. As elaborated upon in the theoretical framework, previous studies have described that CJ is a difficult assessment method when raters are exposed to products with similar quality (Benton, 2021; Gijsen et al., 2021; Van Daal et al., 2017). This is in line with the opinions of the raters, who have indicated that during the CJ assessment they often had to choose the 'better' candidate between two candidates with similar answers (Rater1, Rater2, personal communication, December 14, 2021). Although it is difficult to see it in the results, raters explained it was especially difficult to choose the better candidate between two products to which the raters would have awarded zero points when assessed through RBA. When looking at the results, it can be seen that CJ yielded far fewer candidates that were awarded zero points. Raters pointed out that it would be desirable if there was an option to select both answers or neither one as the better one, in case of similar answers.

Number of Candidates. The last practical shortcoming for applying CJ for high-stake assessments is with regards to the number of candidate products that are needed for a reliable assessment. Since with CJ candidate products must be compared in dichotomous pairs, a minimal selection of $n = 15$ products should be considered. However, the exams of the AvE can be taken at any given time. With RBA, the exam answers can be sent directly to the rater, who then can assess the answers. However, when applying CJ, the raters must wait until the pool of products consists of at least 15 candidate products. Furthermore, in the specific case of the BIV® exam, there are roughly $n = 8$ candidates that participate each month, which would entail that a candidate must wait two months before receiving their grade. Since this is undesirable, the current method of RBA is better suited for the current settings of the AvE.

Additionally, the items for the exam are randomly assigned to each candidate from the item bank. It might be that in a period of two months the candidates are exposed to different items, making it difficult to reach the minimum number of candidate products that are needed to apply CJ as an assessment method.

Limitations

For this study, two limitations have to be mentioned. The first limitation is the small number of raters that participated in this study. Since this study had as goal to obtain reliable results, a total of $n = 113$ comparisons had to be made. Since only two raters were willing to participate in the study, the total comparisons had to be divided over two raters. However, if more raters would have participated, each rater would have had to make fewer comparisons, which would have led to a lower individual total time investment. Moreover, CJ represents the combined raters' perception of product quality, so if more raters would have participated the rank order would have considered the perceptions of all raters. However, it must be noted that even with more raters, assessing single exam items would still have been inefficient.

Due to Covid-19, the study was not standardized. The raters worked from home, using their personal electronics, in their own environment. Although this environment is the same as when raters assess the candidate exams through RBA, the raters may have used a different approach to assess the candidate products through CJ which may have an unknown effect on the results. For instance, as mentioned, rater1 divided assessing the candidate products for the item with a maximum score of 3 over two moments, while rater2 did not. It is unsure what the impact of these kinds of differences in approach might have on the results.

Suggestions for Future Research

Although it may not be practical to adopt CJ as an assessment method for the AvE's high-stake exams, there may be other applications of CJ. This section will elaborate on suggestions for future research on CJ as well as suggestions for CJ within the context of the AvE.

Determining Item Maximum Score

First, CJ can be applied in the context of the AvE to determine the maximum score of an item. Items are created by external item writers that are knowledgeable in the concerning domain. Currently, the writers have to create an item with a pre-determined maximum score. The maximum score of an item is dependent on the difficulty of the item, and the amount of information that candidates must integrate to answer correctly (Jayakodi et al., 2015). Instead of creating items based on the maximum score, it may be possible to use CJ to determine the maximum score of a new item. If item writers create a variety of items with varying levels of difficulty, these items can be uploaded into the D-PAC tool. Other writers and raters can go through pairs of comparisons to select the most difficult item. Based on these judgements, a rank order will be created that ranks the items on difficulty from 'easy' to 'hard'. Similar to awarding scores to candidate products, based on the rank order the maximum scores of the items can be determined.

Nevertheless, this application of CJ has a limitation. Although no issues with regards to transparency will arise, CJ will still be biased towards the sample of items. Similar to having a low or high ability candidate pool, a writer can have created multiple items that assess the same level of understanding. Although these items should theoretically have the same maximum score, one item might be ranked higher resulting in a higher maximum score whilst assessing the same level of difficulty. Research is necessary to determine to what extent this application of CJ is feasible within the AvE.

Formative Application of CJ

Lastly, as it may not be possible to adopt CJ for high-stake summative assessments, it may be possible to apply CJ formatively within the AvE. Bartholomew and Yoshikawa (2018) have stated that the use of CJ as tool in formative settings has shown potential in terms of candidate learning and achievement. Within the context of the AvE, the formative application of CJ might be applied in the form of a practice exam to which a variety of responses have been added. Candidates can take the practice exam and go through comparisons of responses while choosing the 'better' answer. This way, candidates may familiarize themselves with what a 'good' answer entails, which in turn might make it easier for them to articulate their own constructs of quality and apply this to their own work during the actual exam.

The programme director of the AvE has stated that this application of CJ might be useful for candidates (Programme Director AvE, personal communication, January 26, 2022). She elaborates that currently candidates have access to a correction form that describes correct answers to the items on the practice exam. However, candidates are often under the impression that they have to answer in the exact same way. By including multiple responses to the item, candidates may get an idea of what must be included in an answer to be correct.

An important aspect to include is the function of showing the correct response. This ensures candidates know whether their choice of the 'better' answer actually was the better answer. Otherwise, candidates may familiarize themselves with 'good' answers whilst unknowingly choosing the 'worst' answer.

However, there are two limitations that must be noted. First, the same issue with inefficiency of rating answers for a single item will occur. Secondly, a possible limitation of this application of CJ might be that candidates might get confused with the variety of correct answers to the item. Since multiple responses should be uploaded, it may occur that multiple answers are correct. As research has already indicated that it is difficult for raters to choose the better item when exposed to similar or equal answers (Benton, 2021; Gijsen et al., 2021; Van Daal et al., 2017), this might make it even more difficult for candidates to choose the correct answer. Therefore, research is needed to explore the applicability and shortcomings of this formative application of CJ.

Adaptive Comparative Judgement

Lastly, an emerging improvement on the CJ process is adaptive comparative judgement (ACJ). Through adaptivity, raters will have to compare less dichotomous pairs while maintaining a high reliability, thus increasing the efficiency (e.g., Newhouse, 2014; Whitehouse & Pollitt, 2012). Adaptivity refers to the choice of the dichotomous pairs to be compared being based on the outcomes of the judgements made previously. Its rationale is comparable to Computerized Adaptive Testing (CAT), where the choice of the next item in an exam is based on the candidate’s correct or incorrect answer to items. In CAT, the adaptiveness allows for better targeting of items to test takers and for a similar level of precision of a candidate’s estimated ability to be obtained through fewer items than a fixed-length test (Bramley & Vitello, 2019). In CJ, the adaptiveness would entail that the products that have ‘won’ most of the comparisons will become increasingly unlikely to be paired with products that have ‘lost’ most of the comparisons. This way, by pairing the to be compared products more efficiently, the reliability can theoretically be increased whilst having fewer raters (Bramley & Vitello, 2019).

However, ACJ might still not be applicable to high-stake exams since the same issue with transparency remains.

Concluding Remarks

The current research has investigated the applicability of CJ for high-stake exam items that differ in the maximum score. CJ has been compared as an alternative to the conventional approach of RBA because RBA yields lower reliability values than desired for high-stake examinations. Candidate products, previously assessed through RBA, were added to three assessments that represented items with a different maximum score. The results indicate a high reliability and rank orders similar to RBA rank orders. However, CJ has some practical implications including lack of transparency and inefficiency in applying CJ for single items, that lead to the conclusion that CJ is not suitable for assessing high-stake exams that combine multiple items. Since the reliability of RBA is still not desirable for high-stake assessments, an alternative method might be better suitable. More research is necessary to determine which assessment method has a high reliability, validity, efficiency, and transparency that is needed for assessing high-stake exams.

References

AERA, APA, & NCME. (2014). Validity. In *Standards for Educational and Psychological Testing* (pp. 11-32). American Educational Research Association.

Alagumalai, S., & Curtis, D. D. (2005). Classical Test Theory. In R. Maclean, R. Watanabe, R. Baker, Boediono, Y. C. Cheng, W. Duncan, J. Keeves, Z. Mansheng, C. Power, J. S. Rajput, K. H. Thaman, S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars* (pp. 1-14). Springer Netherlands. https://doi.org/10.1007/1-4020-3076-2_1

Almarshoud, A. (2011). Developing a Rubric-Based Framework for Measuring the ABET Outcomes Achieved by Students of Electric Machinery Courses. *International Journal of Engineering Education, 27*, 1-8.

Anderson, H. M., Anaya, G., Bird, E., & Moore, D. L. (2005). A Review of Educational Assessment. *American Journal of Pharmaceutical Education, 69*(1), 84-100. https://doi.org/10.5688/aj690112

Associatie. (n.d.). *Associatie voor Examinering. Dé Experts in Examinering.* Retrieved June 9, 2021 from https://associatie.nl/

Bartholomew, S., & Yoshikawa, E. (2018). A Systematic Review of Research Around Adaptive Comparative Judgement (ACJ) in K-16 Education. *1*, 1-28. https://doi.org/10.21061/ctete-rms.v1.c.1

Benton, T. (2021). Comparative Judgement for Linking Two Existing Scales. *Frontiers in Education, 6*(524). https://doi.org/10.3389/feduc.2021.775203

Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I The Method of Paired Comparisons. *Biometrika, 39*(3/4), 324-345. https://doi.org/https://doi.org/10.2307/2334029

Bramley, T., & Vitello, S. (2019). The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 26*(1), 43-58. https://doi.org/10.1080/0969594X.2017.1418734

Cito. (2018). *De Waarde van de Ritwaarde*. Stichting Cito, Instituut voor Toets- en Examenontwikkeling. https://www.cito.nl/-/media/files/nieuws/vakinhoudelijke-publicaties/de_waarde_van_de_rit-waarde.pdf?la=nl-NL

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., & De Maeyer, S. (2018). An Information System Design Theory for the Comparative Judgement of Competences. *European Journal of Information Systems, 27*(2), 248-261. https://doi.org/10.1080/0960085X.2018.1445461

Coertjens, L., Lesterhuis, M., De Winter, B. Y., Goossens, M., De Maeyer, S., & Michels, N. R. M. (2021). Improving Self-Reflection Assessment Practices: Comparative Judgment as an Alternative to Rubrics. *Teaching and Learning in Medicine, 11*, 1-11. https://doi.org/10.1080/10401334.2021.1877709

Comproved. (n.d.). *Beoordeel Beter en Makkelijker*. Retrieved July 1, 2021 from https://comproved.com/

Cronbach, L. J. (1960). *Essentials of Psychological Testing* (2nd ed.). Harper.

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care, 44*(11), 50-59. https://doi.org/https://doi.org/10.1097/01.mlr.0000245426.10853.30

eX:plain. (n.d.). *About eX:plain*. Retrieved September 8, 2021, from https://www.explain.nl/about-explain

Furr, R., & Bacharach, V. (2013a). Item Response Theory and Rasch Models. In *Psychometrics: An Introduction* (2nd ed., pp. 385-412). SAGE Publications, Inc.

Furr, R., & Bacharach, V. (2013b). Validity: Conceptual Bias. In *Psychometrics: An Introduction* (2nd ed., pp. 197-220). SAGE Publications, Inc.

Gantt, L. T. (2010). Using the Clark Simulation Evaluation Rubric With Associate Degree and Baccalaureate Nursing Students. *Nursing Education Perspectives, 31*(2), 101-105. https://doi.org/https://doi.org/10.1043/1536-5026-31.2.101

Gijsen, M., Van Daal, T., Lesterhuis, M., Gijbels, D., & De Maeyer, S. (2021). The Complexity of Comparative Judgments in Assessing Argumentative Writing: An Eye Tracking Study. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.582800

Goodwin, L. D., & Leech, N. L. (2006). Understanding Correlation: Factors That Affect the Size of r. *The Journal of Experimental Education, 74*(3), 249-266. https://doi.org/10.3200/JEXE.74.3.249-266

Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educational Measurement: Issues and Practice, 8*(1), 35-41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Heldsinger, S., & Humphry, S. (2010). Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments. *The Australian Educational Researcher, 37*(2), 1-19. https://doi.org/10.1007/BF03216919

Huhta, A. (2008). Diagnostic and Formative Assessment. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 469-482). Blackwell Publishing Ltd.

Jayakodi, K., Bandara, M., & Perera, I. (2015). An Automatic Classifier for Exam Questions in Engineering: A Process for Bloom's Taxonomy. 2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE),

Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring Conceptual Understanding in Randomised Controlled Trials: Can Comparative Judgement Help? *British Educational Research Journal, 45*(3), 662-680. https://doi.org/10.1002/berj.3519

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing Mathematical Problem Solving Using Comparative Judgement. *International Journal of Science and Mathematics Education, 13*(1). https://doi.org/10.1007/s10763-013-9497-6

Jönsson, A., & Panadero, E. (2016). The Use and Design of Rubrics to Support Assessment for Learning. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up Assessment for Learning in Higher Education* (1st ed., pp. 99-111). Springer Singapore. https://doi.org/10.1007/978-981-10-3045-1_7

Jönsson, A., & Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Praeger Publishers.

Kellaghan, T., & Greaney, V. (2001). Assessment, Quality, Standards, and Accountability. In *Using Assessment to Improve the Quality of Education* (pp. 19-28). UNESCO: International Institute for Educational Planning.

Keppens, K., Consuegra, E., Goossens, M., De Maeyer, S., & Vanderlinde, R. (2019). Measuring Pre-Service Teachers' Professional Vision of Inclusive Classrooms: A Video-Based Comparative Judgement Instrument. *Teaching and Teacher Education, 78*, 1-14. https://doi.org/https://doi.org/10.1016/j.tate.2018.10.007

Kimbell, R. (2012). Evolving Project E-scape for National Assessment. *International Journal of Technology and Design Education, 22*(2), 135-155. https://doi.org/10.1007/s10798-011-9190-4

Kimbell, R. (2021). Examining the Reliability of Adaptive Comparative Judgement (ACJ) as an Assessment Tool in Educational Settings. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-021-09654-w

Knight, J., Allen, S., & Tracy, D. L. (2010). Using Six Sigma Methods to Evaluate the Reliability of a Teaching Assessment Rubric. *The Business Review, Cambridge, 15*(1), 1-6.

Li, H. (2003). The Resolution of Some Paradoxes Related to Reliability and Validity. *Journal of Educational and Behavioral Statistics, 28*(2), 89-95. https://doi.org/10.3102/10769986028002089

Marshall, B. (2017). The Politics of Testing. *English in Education, 51*(1), 27-43. https://doi.org/10.1111/eie.12110

McMahon, S., & Jones, I. (2015). A Comparative Judgement Approach to Teacher Assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 368-389. https://doi.org/10.1080/0969594X.2014.978839

Mehta, G., & Mokhasi, V. R. (2014). Item Analysis of Multiple Choice Questions- An Assessment of the Assessment Tool. *International Journal of Health Sciences and Research, 4*, 197-202.

Mellenbergh, G. J. (2011a). Introduction. In *A Conceptual Introduction to Psychometrics*. Eleven International Publishing.

Mellenbergh, G. J. (2011b). Principles of Item Response Theory. In *A Conceptual Introduction to Psychometrics*. Eleven International Publishing.

Moskal, B. M., Leydens, J. A., & Pavelich, M. J. (2002). Validity, Reliability and the Assessment of Engineering Education. *Journal of Engineering Education, 91*(3), 351-354. https://doi.org/10.1002/j.2168-9830.2002.tb00714.x

Muntinga, J. H., & Schuil, H. A. (2007). Effects of Automatic Item Eliminations Based on Item Test Analysis. *Advances in Physiololgy Education, 31*(3), 247-252. https://doi.org/10.1152/advan.00019.2007

Nasab, F. G. (2015). Alternative Versus Traditional Assessment. *Journal of Applied Linguistics and Language Research, 2*(6), 165-178.

Nelsen, R. B. (2002). Concordance and Copulas: A Survey. In *Distributions With Given Marginals and Statistical Modelling* (pp. 169-177). Springer. https://doi.org/10.1007/978-94-017-0061-0_18

Newhouse, C. P. (2014). Using Digital Representations of Practical Production Work for Summative Assessment. *Assessment in Education: Principles, Policy & Practice, 21*(2), 205-220. https://doi.org/10.1080/0969594X.2013.868341

Newton, P. E. (1996). The Reliability of Marking of General Certificate of Secondary Education Scripts: Mathematics and English. *British Educational Research Journal, 22*(4), 405-420. https://doi.org/10.1080/0141192960220403

Newton, P. E. (2007). Clarifying the Purposes of Educational Assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170. https://doi.org/10.1080/09695940701478321

Olmuş, H., Nazman, E., & Erbaş, S. (2017). An Evaluation of the Two Parameter (2-PL) IRT Models Through a Simulation Study. *Gazi University Journal of Science, 30*(1), 235-249.

Panadero, E., & Jönsson, A. (2013). The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review. *Educational Research Review, 9*, 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Picardi, C. A., & Masick, K. D. (2014). Reliability. In *Research Methods: Designing and Conducting Research With a Real-World Focus* (pp. 43-53). SAGE Publications.

Pollitt, A. (2004). Let's Stop Marking Exams. *IAEA Conference*. https://www.researchgate.net/publication/241197532

Pollitt, A. (2012). Comparative Judgement for Assessment. *International Journal of Technology and Design Education, 22*(2), 157-170. https://doi.org/10.1007/s10798-011-9189-x

Puth, M.-T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's Correlation Coefficients for Association Between Two Measured Traits. *Animal Behaviour, 102*, 77-84. https://doi.org/https://doi.org/10.1016/j.anbehav.2015.01.010

Rao, C., Prasad, H. K., Sajitha, K., Permi, H., & Shetty, J. (2016). Item Analysis of Multiple Choice Questions: Assessing an Assessment Tool in Medical Students. *International Journal of Educational and Psychological Researches, 2*(4), 201-204. https://doi.org/10.4103/2395-2296.189670

RStudio Team. (2021). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Bosten, MA. http://www.rstudio.com/

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*(2), 119-144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (2009). Indeterminacy in the Use of Preset Criteria for Assessment and Grading. *Assessment and Evaluation in Higher Education, 34*(2), 159-179. https://doi.org/10.1080/02602930801956059

Sanders, J. R., & Vogel, S. R. (1993). The Development of Standards for Teacher Competence in Educational Assessment of Students. In S. L. Wise (Ed.), *Teacher Training in Measurement and Assessment Skills* (pp. 234). Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Schlitz, S. A., O'Connor, M., Pang, Y., Stryker, D., Markell, S., Krupp, E., Byers, C., Jones, S. D., & Redfern, A. K. (2009). Developing a Culture of Assessment Through a Faculty Learning Community: A Case Study. *International Journal of Teaching and Learning in Higher Education, 21*(1), 133-147.

Shipman, D., Roa, M., Hooten, J., & Wang, Z. J. (2012). Using the Analytic Rubric as an Evaluation Tool in Nursing Education: The Positive and the Negative. *Nurse Education Today, 32*(3), 246-249. https://doi.org/https://doi.org/10.1016/j.nedt.2011.04.007

Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education, 29*(3), 211-223. https://doi.org/10.1080/08957347.2016.1171769

Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions. *Teaching of Psychology, 36*(2), 102-107. https://doi.org/10.1080/00986280902739776

Swart, A. (2010). Evaluation of Final Examination Papers in Engineering: A Case Study Using Bloom's Taxonomy. *IEEE Transactions on Education, 53*(2), 257-264. https://doi.org/10.1109/TE.2009.2014221

Tavakol, M., & Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International journal of medical education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thissen, D., & Orlando, M. (2001). Item Response Theory for Items Scored in Two Categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 73-140). Routledge.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychological Review, 34*(4), 273-286. https://doi.org/10.1037/h0070288

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators. *Procedia Economics and Finance, 20*, 679-686. https://doi.org/https://doi.org/10.1016/S2212-5671(15)00123-9

Ushiro, Y., Nakagawa, C., Morimoto, Y., Hijikata, Y., Watanabe, F., & Kai, A. (2008). Effects of Question Types on Item Difficulty in Two Reading Test Formats : Open-Ended and Multiple-Choice. *Annual Review of English Language Education in Japan, 19*, 201-210. https://doi.org/10.20581/arele.19.0_201

Van Daal, T., Lesterhuis, M., Coertjens, L., Kamp, M.-T., Donche, V., & De Maeyer, S. (2017). The Complexity of Assessing Student Work Using Comparative Judgment: The Moderating Role of Decision Accuracy. *Frontiers in Education, 2*. https://doi.org/10.3389/feduc.2017.00044

Webber, K. L. (2012). The Use of Learner-Centered Assessment in US Colleges and Universities. *Research in Higher Education, 53*(2), 201-228. https://doi.org/10.1007/s11162-011-9245-0

Whitehouse, C., & Pollitt, A. (2012). Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment.

Appendix A

Definitions Educational Tests, Educational Assessment, and Test Item Analysis

Educational Tests

An educational test refers to “the instrument for the measurement of a person’s performance under standardized conditions, where the performance is assumed to reflect one or more latent attributes” (Mellenbergh, 2011a, p. 19). This definition includes several statements.

First, tests are instruments. This entails that the test is in itself a measurement instrument, other uses such as predicting or analysing a candidate’s performance, are applications of the instrument.

Secondly, a test measures a candidate’s performance. Two types of performance can be distinguished: maximum and typical performance. Cronbach (1960) used these concepts to distinguish between ability performance and personality measures. He stated that maximum performance refers to what a person knows, or what a person can mentally do. In maximum performance, the correctness can vary as answers can be correct, partly correct, or incorrect. Typical performance, on the other hand, refers to a person’s affective traits or characteristics. Since these personality measures typify a person, a person’s typical performance cannot be evaluated on correctness. In this study, test performance will refer to the maximum performance of a candidate.

Thirdly, a test should be applied under standardized conditions. Performances between candidates and occasions must be comparable, therefore the test should be applied under standard conditions to ensure fair comparisons (Mellenbergh, 2011a).

Lastly, the definition states the assumption of the test performance reflecting one or more latent attributes of a candidate. As the name suggests, latent attributes, or latent traits, cannot be observed directly. Latent traits are assumed to have an effect on the candidate’s item response (Mellenbergh, 2011b). An item refers to the smallest possible subset of a test. A well-constructed item can accurately discriminate between candidates with a high or low ability. If a candidate has a higher ability, it is more likely that the candidate will respond to an item correctly (Furr & Bacharach, 2013a).

Educational Assessment

After an educational test has been conducted, raters can assess the candidate’s responses. This section will elaborate on the definition of assessment and the type of assessment that is applied at the AvE.

Definition

Although assessment is a term that is often referred to, multiple definitions exist that encompass different views. Webber (2012) referred to assessment as a method of evaluating a candidate’s comprehension and achievement. Although this definition is the basis of assessment, there are more aspects to assessment making this definition oversimplified. Erwin (1991, as cited in Anderson et al., 2005) defined assessment as “the systematic basis for making inferences about the learning and development of candidates. More specifically, assessment is the process of defining, selecting, designing, collecting, analysing, interpreting, and using information to increase candidates' learning and development”. This definition is more specific but mostly focuses on the formative goal of assessment. As this research focuses on the summative application of assessment, this report will refer to assessment as “the process of obtaining information that is used to make educational decisions about students; to give feedback to students about their progress, strengths, weaknesses; to judge instructional efficiency and curricular adequacy; and to inform policy” (Sanders & Vogel, 1993, p. 41). This definition of assessment encompasses the broad purposes of assessment.

Summative Versus Formative Assessment

As mentioned, this research focuses on the summative goal of assessment rather than formative assessment. Summative assessment is currently the most dominant method of assessment in education and is also referred to as assessment *of* learning. As the name implies, summative assessment aims to summarize the achievement of a candidate (Sadler, 1989). Summative assessment occurs at the end of learning and aims to determine the outcome of an instructional program or of individual learners (Huhta, 2008). In contrast, formative assessment, which is also referred to as assessment *for* learning, aims for candidates to improve their learning, and for teachers to improve their teaching (Huhta, 2008).

While formative assessment aims to provide feedback to improve, the main goal of summative assessment is comparing a candidate’s performance against a standard or benchmark. Therefore, summative assessment is benchmark-referenced. Summative assessments often have high stakes that determine a candidate’s grade for the purpose of certification (Sadler, 1989).

Assessment Versus Evaluation

Although related and sometimes used interchangeably, assessment differs from evaluation. Assessment focuses on whether candidates have displayed skills, abilities, or competencies towards stated educational outcomes. It highlights the relationship between educational programs and candidate learning and performance. The data collected during assessment can be used towards the evaluation process which helps to make judgements about quality or efficiency. Thus, evaluation goes beyond candidate performance to determine the impact of programs or curriculum (Anderson et al., 2005).

Assessment Standards

All forms of assessment need to adhere to certain standards. An assessment needs to be reliable as well as valid.

Reliability

In its simplest definition, reliability refers to the consistency of assessment scores. If a test or measurement instrument is reliable, this would mean that a candidate would attain the same score regardless of where the test was taken, when the scores were obtained, or who assessed the test (Moskal et al., 2002). The current study concerns two specific types of reliability which is the inter- and intra-rater reliability and internal consistency.

Inter- and Intra-Rater Reliability. The rater reliability refers to the consistency of assessment scores through two aspects. First, inter-rater reliability refers to the consistency of the scores of candidates across two or more independent raters. Secondly, intra-rater reliability refers to the consistency of a single rater to one assessment at different points in time (Moskal et al., 2002; Picardi & Masick, 2014).

Internal Consistency. The internal consistency is most widely expressed as Cronbach's alpha, or the coefficient alpha (α). Cronbach's alpha was introduced to determine the reliability of a measurement instrument at the test level. It can be defined as "a unitless index assigned to a measurement instrument, taking its value on the interval [0,1], for the extent to which the measurement instrument is free from error, with the values 0 and 1 corresponding the extreme cases of "pure error" and "no error" respectively (Li, 2003, p. 91). If test items are correlated, Cronbach's alpha will increase. Thus, if a higher value of Cronbach's alpha is needed, more correlated items need to be added. Generally, a Cronbach's alpha value of higher than 0,7 would be sufficient, while a Cronbach's alpha of higher than 0.8 would be considered great. However, it is important to note that if Cronbach's alpha is too high, this might indicate redundancy in items (Tavakol & Dennick, 2011).

Validity

Although it is often said that a test is valid, this formulation is confusing. That is because a measure as such is neither valid nor invalid. Rather, validity concerns how the test scores are used and interpreted (Furr & Bacharach, 2013b). In the domain of measurement, validity can be used in two ways. First, validity can be used to show how an interpretation is justified. In this sense, evidence should be collected that supports the interpretation. Secondly, validity can be used to evaluate an interpretation's overall plausibility. Then, it should be evaluated to what extent the proposed interpretations are plausible and appropriate (Kane, 2006). In both definitions, the validity of a test's interpretations should be based on evidence and theory.

In this report, validity will refer to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" source (AERA, APA, & NCME, 2014, p. 11).

Assessment Purpose

Educational assessment can serve different purposes on different levels. In the context of this study, the purposes concern the candidate level and the organisational level.

Candidate Level

On the level of the candidate, assessment can serve the following purposes. First of all, assessment describes a candidate's understanding, and it diagnoses learning problems. Candidates can use this information to help them learn. Moreover, assessment can serve as a motivation to achieve goals. Additionally, assessment can serve to certify one's competences, which in turn can help candidates to select a job or to select the next level of the educational system (Kellaghan & Greaney, 2001; Newton, 2007).

Organisational Level

On the organisational level, assessment can serve as a judgement about the efficiency of an institution, and to reach a judgement about the adequacy of the performance of an educational system (Kellaghan & Greaney, 2001; Newton, 2007). The assessment results of candidates can be used to decide whether the standards of the organisation are rising or falling over time. Moreover, when item analysis is applied, organisations can gain insights into the quality of items and thus the test. Additionally, this information can identify needs which can help administrators to decide how to allocate resources.

Item Analysis

As mentioned, assessments are useful on the organisational level. Organisations can apply item analysis to gain insights into candidates' responses to each test item. Based on these results, decisions can be made regarding the quality of an item. Two approaches to item analysis are Item Response Theory (IRT) and Classical Test Theory (CTT). IRT attempts to model the relationship between a respondent's latent ability called theta (θ) and the probability of the candidate correctly answering a test item (Harris, 1989; Olmuş et al., 2017). Although IRT is a more realistic approach than CTT, it involves large scale test development and scoring (Thissen & Orlando, 2001). Since the AvE exams do not involve large scale testing, CTT is more suitable for item analysis. Therefore, this section will elaborate more on item analysis through CTT.

CTT constitutes a series of concepts and related techniques that forms the base for various measurement theories and approaches. It concerns using observable information such as test scores, to gain insights into unobservable information such as a candidate's ability or the quality of items (Alagumalai & Curtis, 2005; DeVellis, 2006). To determine the quality of items, CTT uses the difficulty of an item as well as the item discrimination.

Item Difficulty

To determine the difficulty of an item in CTT, the p-value is used. The p-value is the proportion of candidate's that answer an item correctly. Generally, an item with a p-value between 0.3 and 0.7 is considered to be acceptable. Within this range, the items with a p-value between 0.5 and 0.6 are considered to be ideal (Mehta & Mokhasi, 2014; Rao et al., 2016). A high p-value, more than 0.7, indicates that many candidates answered the item correctly, which suggests that the item may be too easy. Similarly, a low p-value, lower than 0.3, suggest that the item is too difficult (Alagumalai & Curtis, 2005; DeVellis, 2006).

Item Discrimination

The Rit-score indicates how the score of an item is related to the total exam score. In other words, the Rit-score is the correlation (r) between an item score (i) and the total score (t). It indicates to what extent an item can discriminate between high ability candidates and low ability candidates. If the Rit-score of an item is high, this means the item is strong and positively related to the rest of an exam. So, an item with a high Rit-score indicates that candidates who answered the item correctly, also did well on the total exam (Alagumalai & Curtis, 2005; Cito, 2018; DeVellis, 2006; Muntinga & Schuil, 2007). Generally, a Rit-value between 0.2 and 0.4 is considered as good while values higher than 0.4 are excellent. For Rit-values lower than 0.2, the item discriminates poorly (Mehta & Mokhasi, 2014; Rao et al., 2016).

Rit-values might take a negative value. For negative Rit-values, low ability candidates answer an item correctly while high ability candidates answer the item incorrectly. If this happens, the item needs to be critically examined, since a negative Rit-value might indicate a defective item, a trick question, or a mistake in the answer key (Rao et al., 2016).