How Reliable is Performance-Based Assessment? Comparing Holistic, Analytic, and Comparative Judgment Approaches

Charlotte H. Sievers

Master Thesis - Talent Development and Creativity

S3357171
June 2023
Department of Psychology
University of Groningen
Daily supervisor: MSc. J. C. Arboleda Cardona
Daily supervisor and first examiner: Dr. A.S.M. Niessen
Second examiner: Dr. R.F.A. Cox
………………………..

Abstract

Assessment is crucial in education, and determining its reliability is vital for ensuring consistent and accurate evaluation. Two common assessment approaches are analytic and holistic methods. In educational research, a consensus on which method presents better reliability is still lacking. A promising emerging approach is comparative judgment (CJ) assessment, which relies on relative judgment during decision-making. Given the standardization offered by the analytic method by specifying performance dimensions and the holistic approach relying on intuitive judgment which can often be incoherent, we hypothesized that the analytic method would present the highest inter-rater reliability. CJ assessment, which determines a rank order based on multiple pairwise comparisons, was expected to demonstrate greater reliability than the holistic approach. A convenience sample of $N = 135$ undergraduate students was gathered to assess 30 short-written essays, using a between-subjects design. Raters' perceived complexity of applying the method and general decision-making style (holistic vs analytic) were additionally explored to detect possible barriers to the practical implementation of the most reliable method. The intraclass correlation coefficient estimated inter-rater reliability for the analytical and holistic method, revealing low to moderate reliability. The CJ method's reliability was assessed using scale separation reliability and Pearson's product-moment correlation, returning high inter-rater reliability. Significant differences in perceived complexity and decision-making tendencies were found. Due to this study's methodological limitations, it is challenging to draw definitive implications. Future research should explore the validity of comparing different reliability measures and use a more adapted rubric with fewer criteria.

Keywords: Performance-based assessment, convenience sample, inter-rater reliability, rater perceptions, intraclass correlation, scale separation reliability

How Reliable is Performance-Based Assessment? Comparing Holistic, Analytic, and Comparative Judgment Approaches

Educational institutions are a place for students to acquire skills and abilities in preparation for their future careers. To ensure that learning is taking—and has taken—place, it is necessary to assess students' abilities and learning processes and to give them feedback on their performance (Sadler, 2005; Vercellotti & McCormick, 2021). The predominant performance-based assessment methods in education can be summarized as following either a holistic or an analytic approach (Barkaoui, 2011; Brookhart et al., 2016; Daniels & Harley, 2017; Harsch & Martin, 2013; Jones & Wheadon, 2015; Jönsson et al., 2021; Meadows & Billington, 2005; Sadler, 2009). Although the definitions of these methods vary, holistic assessment generally entails basing the decision on the performance as a whole on the intuitive judgment in the rater's mind, and analytic assessment breaks down performance into dimensions, optionally employing a rating scale, such as a rubric. It may additionally include some sort of decision rule that determines the weight each criterion should carry to arrive at an overall score (e.g., Arboleda et al., 2023). Accordingly, proponents of holistic assessment assume that a student's performance cannot be viewed as separate entities because an overall impression is required to formulate a judgment, expressed in an overall score. In contrast, advocates of analytic assessment consider the assessed skill/ability to have subdimensions, which should be rated separately to enhance objectivity in rating and increase transparency in how the rating is composed (Barkaoui, 2011; Meadows & Billington, 2005; Reddy & Andrade, 2010). A lesser-known assessment approach that is getting more and more attention due to its allegedly easy and intuitive application, is a comparative judgment (CJ) assessment (Verhavert et al., 2019). While this method essentially categorizes as a holistic assessment as it also refers to the performance as a whole in the final score, it distinguishes itself from generic holistic performance-based assessment by requiring the raters to assess two performances of a set of performances at a time, rather than by themselves. Specifically, raters compare a given work with another one to decide which one is better (i.e., making relative judgments), thereby producing a rank order of the assessed performances (Han, 2021).

Investigating the reliability of an assessment is essential for ensuring that the objective under examination is consistently and accurately being assessed to enhance the fairness and independence of the rater (Bannigan & Watson, 2009; Bruton et al., 2000; Reddy & Andrade, 2010; Verhavert et al., 2019). To determine which method leads to consistent assessment results when multiple raters are involved, it is common practice to investigate the method's inter-rater reliability (Arkes et al., 2006; Bruton et al., 2000; Koo & Li, 2016). Inter-rater reliability generally distinguishes between absolute agreement and consistency (Koo & Li, 2016). While absolute agreement indicates that the raters agree on the awarded score for each performance, consistency relates to whether the raters agree on the rank order of assessed performances. For this study, the definition of consistency is used to compare the inter-rater reliability among raters within the respective assessment method, as it offers a more nuanced understanding of the agreement between raters and presents more practical performance-based assessments. Specifically, it allows for examining how raters agree on the rank order even if raters do not agree on the precise rank, due to the inherent subjectivity of rating performance-based assessments (explained in more detail below).

The variety of definitions used for holistic and analytic assessment in research and the mixed findings in terms of which method leads to higher reliability make it challenging for drawing conclusions on which method to endorse (Arboleda et al., 2023). Consequently, whether a rater’s judgment of the performance relies on intuition, prespecified performance dimensions, or a CJ might affect the reliability of the awarded score to a different extent. Mixed results in educational research on which of these assessment methods produce the most valid and reliable results call for increased analysis of the assessment components (e.g., the rating method and integration of information). Particularly, to provide accurate feedback for improvement on the assessed performance and ensure that students are evaluated fairly (Arboleda et al., 2023; Barkaoui, 2011).

Performance-Based Assessment

Thus far, the object of assessment (i.e., the students’ works) is referred to as performances, to demonstrate the mechanisms of these assessment methods. Student performances can range from demonstrating specific skills or abilities to more complex displays of abilities. This is to say, the use of a rubric, provision of a holistic rating, or comparison of two works is applicable to a broad range of assessment situations. Although there might be, a correct answer to the assessed performance is not always discernible due to the inherent complexity of performances (e.g., Vercellotti & McCormick, 2021). The more difficult it is to delineate a correct answer to the task, the greater the potential risk for unreliability (Meadows & Billington, 2005). For instance, when solving an equation, there is only one correct answer, even if the paths on how to arrive there might vary. In tasks where critical thinking or reasoning (i.e., professional skills) are under assessment, the correct answer is not always as clear-cut and has to be considered in the context (Allen et al., 2005). That is to say, there are multiple ways to express the skill under assessment that can be either right or wrong, presenting more of a continuum of performance quality. Performance-based assessment may take many forms, including tests, essays, projects, presentations, or even on-the-job observations, and generally investigates open-ended tasks (Hartell & Buckley, 2021; Jones & Wheadon, 2015; Sadler, 2009; Smith, 2000). In this study, we chose a reasoning task with a less clear-cut correct answer for assessors unfamiliar with the topic to evaluate whether the presented assessment methods can ensure inter-rater reliability of complex tasks.

Prevailing Assessment Methods

Analytic Assessment

According to research, an assessment that involves standardization in terms of rating and integration of information enhances consistency among decision-makers (i.e., raters) for several reasons (Arboleda et al., 2023; Dawes, 1979; Jonsson & Svingby, 2007; Kahneman & Klein, 2009; Sawyer, 1966). Essential for establishing such standardization is the application of specific rating criteria to guide the assessment, division of performance into individual dimensions, and detailed performance levels. A decision rule is beneficial for integrating information by determining how much weight the subscores of the rating criteria receive to formulate a final judgment (e.g., Meadows & Billington, 2005; Vercellotti & McCormick, 2021). This ensures that each rater judges each performance based on the same predefined criteria and that every rater is bound to the same rating procedure, thereby enhancing consistency. People are usually not very consistent in applying the same measures to different performances during assessment and are prone to bias (Dawes, 1979), which is counteracted by standardization to some degree.

In the educational context, rubrics are an application to obtain a standardized rating. A rubric entails multiple dimensions and descriptions reflecting various levels of performance, which help the rater determine the quality of the performance (Vercellotti & McCormick, 2021). Hence, the essential features of rubrics are providing evaluation criteria, quality definitions, and scoring strategies, as weighting schemes can easily be implemented (Reddy & Andrade, 2010).

Thereby, rubrics offer clear guidance to assessors about what to look for during assessment and how to assign ratings based on specific criteria, leading to more consistent and standardized grading decisions (Meadows & Billington, 2005). Such a practice is essential for inter-rater reliability and for providing informative feedback to students.

However, the literature also reports some drawbacks to the application of analytic assessment (Eckes, 2008; Meadows & Billington, 2005; Panadero & Jonsson, 2020). For instance, developing and validating rubric descriptors can be challenging, as they require time for in-depth consideration of the to-be-included criteria and performance levels, as well as reviewing and evaluating their effectiveness before and after implementation (Han, 2021; Vercellotti & McCormick, 2021).

Holistic Assessment

The core belief of this assessment method is that a generic skill/ability cannot simply be broken down into specific performance qualifications but that there are more aspects to consider that can only be realized by the rater, and these aspects may vary between different performances of similar skills (Barkaoui, 2011; Sadler, 2009; White, 1984). In other words, holistic raters assume the whole is more than its parts. Thereby, it allows for subjective consideration of the performance, based on the assessors' impressions (Meadows & Billington, 2005; White, 1984). Accordingly, predefining criteria and how these criteria should be added may reduce the inclusion of varying distinctive traits (i.e., criteria) the rater might identify within a set of performances (Sadler, 2009). This is not to say that holistic assessment always omits the use of predetermined criteria. The lack of a comprehensive definition of holistic assessment in educational research allows for a range of descriptors of what holistic assessment entails (Arboleda et al., 2023; Barkaoui, 2011; Meadows & Billington, 2005). Even if the assessment includes criteria, holistic assessment distinguishes itself by awarding a single rating for the whole performance, while taking predefined criteria into account. However, this still allows raters to include idiosyncratic criteria in the final rating or to value one criterion to a different extent than another within a set of performances (Barkaoui, 2011).

Comparative Judgment Assessment

A more recent and less commonly used assessment method that has been the focus of some previous educational research is called comparative judgment (CJ) assessment (Hartell & Buckley, 2021; Verhavert et al., 2019; Walland, 2022). It requires a group of assessors to individually judge which one out of two performances is better in relation to a set of (implicit) criteria for a certain number of pairs. The rank order of the performances is estimated by computing the 'wins' each performance has compared to the other performances that are being assessed, and represented by theta values. The Bradley-Terry-Luce (BTL) model is most commonly used for this analysis (Verhavert et al., 2019). The underlying mechanisms of this method root in Thurstone's Law of Comparative Judgment (1927) which states that humans are generally more reliable and accurate in making judgments when they compare one object with a similar object (i.e., relative judgment; in Thurstone, 1994). Conversely, judging an object in isolation (i.e., absolute judgment) is the norm in analytic and holistic assessment, as the quality of a performance is assessed by itself, at most, with some assistance in making a judgment provided by rating criteria. However, the comparison with Thurstone's realizations is not as seamless as often presented in the CJ literature (Kelly et al., 2022). Indeed, Thurstone investigated physical stimuli that could be quickly evaluated, like short statements, which is less commonly the case in performance-based assessment. He furthermore noted that his Law is less easily applied for diverse stimuli, which, again, may very well be the case in performance-based assessment due to its complexity. Thus, the framework for CJ to argue that relative judgment of performance is essentially superior to absolute judgment requires more validation.

As the quality of a performance depends on the assessed set of performances, this method is inherently norm-referenced. This reliance on peer performance might reduce reliability as the reference groups' performance is not necessarily stable, and an absolute standard for comparison is lacking (Lok et al., 2016; Sadler, 2005). However, multiple comparisons per representation are made to estimate the quality of the performance, it presents high reliability nonetheless (Verhavert et al., 2019; Walland, 2022). In a meta-analysis of CJ studies, Verhavert et al. (2019) concluded that a large number of comparisons enhances the level of reliability. Accordingly, an average of 13 comparisons per representation is associated with .7 reliability, and 26-37 comparisons per representation are associated with .9 reliability. However, the meta-analysis did not confirm that the number of assessors has a significant effect on the reliability of the CJ method, only the number of representations per assessor.

Reliability in Educational Studies

A few studies comparing the reliability of the presented assessment methods are available, providing insights into the variability of application with regard to provided criteria, rater training, level of experience among raters, and rating context (Arkes et al., 2006; Daniels & Harley, 2017; Han, 2021; Harsch & Martin, 2013; Jones & Wheadon, 2015; Marshall et al., 2020). Studies comparing CJ with holistic assessment in secondary education present higher inter-rater reliability for CJ both with content experts (Marshall et al., 2020) and peers (Jones & Wheadon, 2015) as raters. These findings indicate that in the absence of criteria (Jones & Wheadon, 2015), raters arrived at a more consistent rank order when comparing performance to another one rather than assessing performance in isolation. Compared to analytic assessment, however, CJ produced inferior inter-rater reliability (Han, 2021). The researchers suggest that the higher inter-rater reliability for analytic assessment is due to more data points per performance being gathered compared to CJ. Specifically, more ratings per performance were produced through the rubric used in the analytic method, relative to the comparisons per performance in the CJ assessment of this study. However, in their study, the CJ assessment did not include criteria, unlike the analytic assessment, which might have been an additional explanation for the lower reliability (Han, 2021). A comparably greater extent of research is directed at comparing holistic and analytic assessment methods, highlighting the ambiguity in assessment reliability research. While some studies fail to demonstrate significant differences between the assessment methods regarding reliability (e.g., Daniels & Harley, 2017), others demonstrate superior reliability in the analytic condition (Arkes et al., 2006; Jönsson et al., 2021).

Perceptions of Raters

The raters' decision-making process plays a role in the rating process (Baker, 2012). Studies' findings of raters' inability to distinguish between performance levels, differences in the relative values placed by different assessors on varying criteria, and difficulties in justifying their assessment results illustrate the inherent subjectivity of performance-based assessment (Brookhart et al., 2016; McMillan, 2003). Findings like this suggest that raters' perceptions of assessment—in terms of how a decision about a performance's quality is formed and how to apply this when rating— are thus important for the practical implementation of the assessment method.

According to Lodato (2008), human judgment is best characterized by intuition and analysis, meaning that one has a preference to combine pieces of information in the mind (holistically) or rationally and consistently (analytically) by using some sort of decision aid.

Generally, holistic approaches are subject to lower reliability as the judgments often differ between raters, as is often reported in other human performance domains (Dawes, 1979; Kahneman & Klein, 2009). Applying this to the educational field, one could assume that while rating a performance, a rater with a tendency to holistic judgment may have already made up their mind on how to grade a specific performance or will direct their focus on aspects that are personally deemed important. In contrast, the rater that tends to analytic judgment may pay attention to the same aspects between performances and take more time to form a decision. Exploring raters' tendencies to either analytic or holistic judgment might thus provide insights into how well the respective method can be implemented in practice.

Additionally, considering raters' ease of applying the assessment method might help to determine the practical use of the respective method. Specifically, given the findings in existing research on assessment methods, raters may have difficulties in discerning the assessment criteria (e.g., Han, 2021). Examining whether raters find it difficult to understand the assessment method itself, especially if they are not experts in the assessed topic may advance the understanding of how likely it is that the most reliable method can be implemented in practice.

Aim and Objectives

To determine the most reliable assessment method for performance-based assessment in higher education, we compared the inter-rater reliability of analytic, holistic, and CJ performance-based assessments. To our knowledge, no study has directly compared all three assessment approaches to determine which assessment approach provides the highest consistency between assessors. Based on the existing literature, it is hypothesized that:

H1. The analytic assessment approach will provide the highest inter-rater reliability and thus will show higher inter-rater reliability than the holistic and CJ approach.

H2. The CJ approach will return higher inter-rater reliability than the holistic, but lower inter-rater reliability than the analytical approach.

Furthermore, the raters' perception of the assessment methods' complexity and how these perceptions differ between conditions, as well as what type of assessment raters generally prefer across conditions was explored. This is relevant because such perceptions likely influence implementation in practice.

Method

Participants

For this study, a sample of first-year students (Dutch and international) of the University of Groningen psychology program was recruited through the SONA systems platform to serve as assessors. Through this platform, students are required to collect points for participating in studies from the university. To investigate the inter-rater reliability of the analytic and holistic method, the intraclass correlation coefficient (Koo & Li, 2016) is used. Due to the nature of the data obtained from CJ, the ICC is not suitable. Accordingly, the Scale Separation Reliability (SSR) together with Pearson's product-moment correlation coefficient (Verhavert et al., 2018; Crompvoets et al., 2022) are the most commonly used and best-fitting reliability indexes. These coefficients are discussed in more detail below. Given the complexity of estimating the required sample size for intraclass correlations, the ambition in this study was to gather as many participants as possible, with ideally 30 participants per condition.

A total of *N* = 135 students ultimately participated in the study with a mean age of 20.28 (S*D* = 2.05); 67% identified as female and 0.7% as other. Having gained rating experience prior to participation was reported by 90 of the participants. Of those participants, 43% had relied on their expertise or intuition and 52% had used rubrics for previous assessments. One student reported experience with comparing two works for assessment. Informed consent from all participants and ethical approval from the Ethical Committee of the University of Groningen (code: PSY-2223-S-0236) were received prior to the experiment. For part-taking in our study participants were rewarded with course credits estimated based on the 60 min duration of the study.

Procedure

Participants received detailed information (see Appendix A) before giving informed consent. Through the SONA system, participants were able to sign up for a timeslot to perform the experiment in the University's computer lab. This study used a between-subjects design to investigate and compare the inter-rater reliability of the holistic, analytic, and CJ assessment methods. Hence, per time slot, one of the assessment methods was presented to the participants. Before commencing the grading, participants were asked to state their age and previous grading experience and subsequently received a small training for the respective assessment method.

The training differed per condition in accordance with the respective assessment method, while the assessed assignments used for the training were identical in all conditions. In the holistic condition, 48 participants across 12 sessions were introduced to the definition of holistic rating and a list of criteria (Appendix C). To get acquainted with the assessment method, each participant received two assignments to practice rating, after which they had the opportunity to approach the researcher to resolve possible unclarities. The analytic condition with 44 participants was essentially set up identically, except that they were presented with a rubric (Appendix D) that contained more detailed criteria dimensions and performance levels. Lastly, in the CJ condition, 43 participants received the same list of criteria as the holistic condition and were directed to the website NoMoreMarking (NMM), for which the comparison mechanisms were explained. Their training involved two assignments that they were asked to compare. Subsequently, participants were asked to rate or compare the 30 short-written assignments, which were identical in each condition but presented to the participants in random order. Additionally, the participants were introduced to the questions that had to be answered in the assignments and received a sample of a well-written assignment that accurately responded to these questions. The NMM website automatically chose 30 random comparisons based on the provided assignments and ensured that all assignments were assessed approximately the same amount of times. After the judgment task, participants in all three conditions filled in the 11 items of a scale that measured decision-making tendencies (holistic and analytic), and the perceived complexity scale that consisted of three items.

Only data with active consent were used for analysis before any identifiers were removed. All data is stored for at least 10 years on the secure server of the university to which only the researchers employed at the University of Groningen have access.

Measures and Materials

The Assignments

The 30 assignments used for this study were essays obtained from the course Test Theory, in which second-year Bachelor's students of the psychology program of the University of Groningen had to formulate conclusions about different types of validity of selection instruments used for the Psychology program admission of their university. The word limit for these assignments was 100-250 words. All 30 essays were randomly selected from a pool of approximately 300 essays, pseudonymized, and finally assessed. The number of essays was partially based on the suggestion by Koo and Li (2016) to involve at least 30 different to-be-rated essays. Crompvoets et al. (2022) and Verhavert et al. (2019) recommended 26-41 comparisons per assignment to achieve high reliability. To conform to this we aimed at 30 comparisons per essay, resulting in (30x30) *n* = 900 pairwise comparisons.

Rating Scale and Scoring

The rating criteria (i.e., task fulfillment, organization, and academic writing style) were inspired by a study investigating the reliability of analytic and holistic methods by Harsch and Martin (2013), who based their criteria on the Common European Framework of Reference (CEF; Council of Europe, 2001), and the criteria of a rubric from a skill-course of the University of Groningen. Given the need for students to express their conclusions with adequate grammar and to understand the order of argumentation, these criteria were deemed fit for this study. These criteria were identical across conditions. However, while in the holistic and CJ conditions participants were provided with short descriptions of each criterion (see Appendix B), the analytic condition included a rubric (see Appendix C) in which the descriptions of each criterion entailed detailed indicators of each performance level (very good, good, sufficient, and insufficient). Respectively, the criteria helped to inform the judgment on the quality, either of the overall performance, in comparison with another performance, or in detail on the specific aspects of the performance.

Decision-Making Tendencies

Furthermore, the selection decision-making style (SDMS; Lodato, 2008) scale was adapted to explore participants' decision-making tendencies. This scale entails 11 statements on participants' stance towards holistic—including CJ—and analytic judgment. The original scale reported a Cronbach's alpha of α = .82 for the intuitive (i.e., holistic) subscale and α = .77 for the analytic subscale, respectively (Lodato, 2008). Similar internal consistency was achieved in this study, with α = .82 for the analytic and α = .75 for the intuitive subscale, thereby indicating acceptable to good internal consistency. The participants indicated their level of agreement with each statement on a 5-point Likert scale, where 1 = strongly disagree and 5 = strongly agree. The scale included items such as "*I think it is important to rely on your "gut" when rating a student's work*" (agreement indicative of a holistic decision-making style) and "*Using scores on appropriate pre-defined criteria is a good way to rate students' work*" (agreement indicative of an analytic decision-making style; see Appendix D for more detail).

Perceived Task Complexity

To explore the participants' perceived complexity in applying the assessment method, the perceived complexity scale which had an acceptable internal consistency (α = .70; Langer et al., 2022) was adopted for this study. It consisted of three items (e.g., "*The assessment method I used to grade students and how it works is easy to understand*."), that were each rated on a 5-point Likert scale as well (1 = strongly disagree; 5 = strongly agree; Appendix E). This study obtained a slightly lower internal consistency for this scale (α = .66), indicating questionable reliability.

Data Analysis

The data collected from the Qualtrics questionnaire and the NNM website were analyzed using the R studio software (R Core Team, 2017). Prior to analysis, the reverse-coded items of the decision-making scale and perceived complexity scale were transformed. To assess the inter-rater reliability of the analytic and holistic assessment methods, the intraclass correlation coefficients (ICC; Koo & Li, 2016) were estimated using a 2-way random-effects model of which the mean of a single rater served as an assessment basis. Specifically, the *ICC* was used to investigate the consistency of raters. For this, the means of the awarded scores were computed across the eight criteria per essay for the analytic condition to arrive at an overall score per essay, per rater. The *ICC* scores were computed for each condition based on the mean scores awarded to each essay by each rater and indicated using 95% confidence intervals, where 0.5 = poor, 0.5 - 0.75 = moderate, 0.75 - 0.9 = good, and greater than 0.90 = excellent reliability (Koo & Li, 2016).

To analyze the CJ data, the raw data of the 30 pairwise comparisons made by each rater were converted into a scaled rank order by use of the BTL model, which calculates pairwise comparison parameters using maximum likelihood estimation (Bradley & Terry, 1952; Luce, 2005). This was conducted automatically by the NMM system, similar to the study by Han (2021), Marshall et al. (2020), and Jones and Wheadon (2015), who also investigated the inter-rater reliability of the CJ method. Subsequently, the SSR and Pearson's product-moment correlation coefficients were computed to provide an estimate of the inter-rater reliability of the method. Although the SSR is usually used as a measure of internal consistency, it is acceptable for estimating inter-rater reliability as well (Verhavert et al., 2018). This is because it informs on "the reproducibility of the results across groups of assessors who are equivalent in background characteristics" (Verhavert et al., 2019, p. 543). The Pearson product-moment correlation is a measure of inter-rater reliability that is computed using the split-halves technique and returns a value between -1 and 1, with a value close to 1 representing a perfect positive linear relationship. The SSR defines a reliability *SSR* > .7 as acceptable, *SSR* > .8 as good, and *SSR* > .9 as very good (Notion, n.d.). No p-value of the SSR and Pearson's product-moment correlation will be reported as the primary interest is the strength and direction of the linear relationship rather than its statistical significance.

Results

Data Inspection and Preliminary Analysis

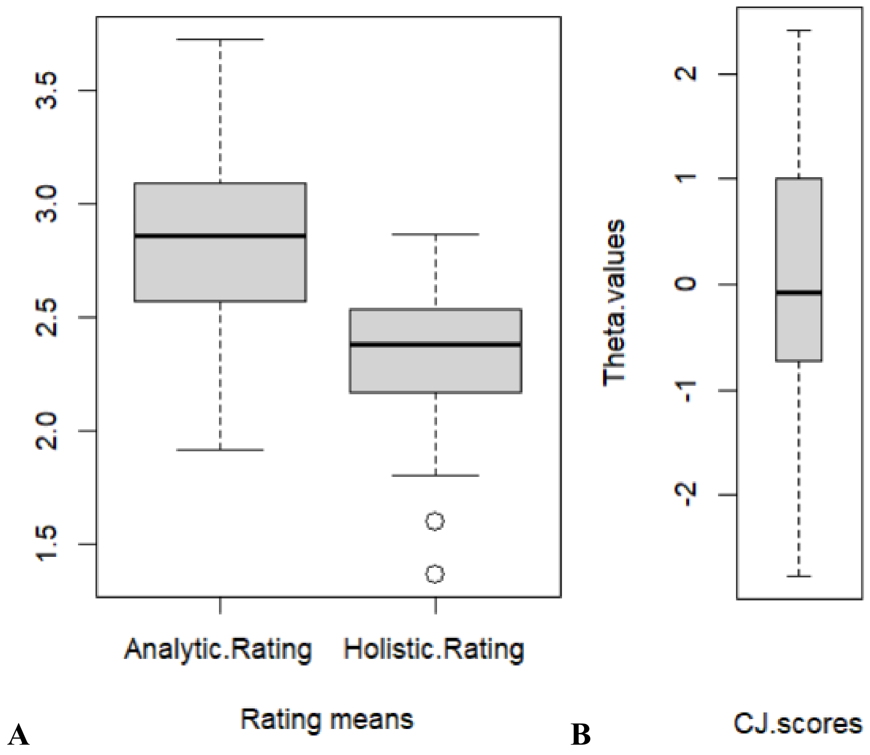
Prior to the hypothesis testing, the ratings of the analytic and holistic condition were transformed to numeric values so that the ratings of “very good”, “good”, “sufficient”, and “insufficient” correspond to 4, 3, 2, 1, respectively. Similarly, the indications of the 5-point-Likert scale of the decision-making and perceived complexity scales were transformed to 1 (strongly disagree), 2 (somewhat disagree), 3 (neither agree nor disagree), 4 (somewhat agree), and 5 (strongly agree), after reverse coding item 11 of the decision-making scale and items 1 and 2 of the PC scale. This was done to perform a meaningful analysis of ordinal data so that means could be attained for each grading condition and each scale. Subsequently, the data were screened for outliers by analyzing the duration participants took for partaking and whether extreme responses were awarded repeatedly by participants to identify inattentive respondents. No outliers were detected and the Shapiro-Wilk test returned no significant p-value for the holistic ($W = .96$, $p > .1$) or analytic ($W = .99$, $p > .9$) assessment, indicating that the mean score distribution can be considered approximately normal. For the CJ assessment, outliers were checked by scanning the infit statistics of each judge, which represents the average of all residual squared values per the decision made. According to the NMM website, judges with an infit above 1.3 should be excluded from the moderation task to ensure consistency (Notion, n.d.). In this sample, eight judges matched this description. However, repeated analysis without these judges’ data revealed virtually identical results. Therefore, results on the full sample are reported. The descriptive statistics of the mean ratings per condition can be found in Table 1 and their distribution is illustrated in Figure 1. In contrast to the analytic and holistic condition, which resulted in scores between 1 and 4, the scores obtained from the CJ condition are theta values, ranging between -2.77 and 2.42.

Table 1
Means and Standard Deviations of Assessment Conditions' Rating

	Analytic	Holistic	CJ
M (SD)	2.83 (*.40*)	2.32 (*.31*)	0 (*1.19*)
Range	1.91 - 3.72	1.37 - 2.87	-2.77 - 2.42
n	44	48	43

Note. Standard deviations are presented in parentheses. The range indicates minimum to maximum rating, averaged across all raters per condition. Values represent scores, where 1 = insufficient, 2 = sufficient, 3 = good, and 4 = very good.

Figure 1
Boxplots of Rating Means per Condition



Note. Quartiles are presented. Panel A: Mean ratings per analytic and holistic condition. The y-axis represents the rating scale from 1 (insufficient) to 4 (very good). Panel B: Theta values of the CJ scores.

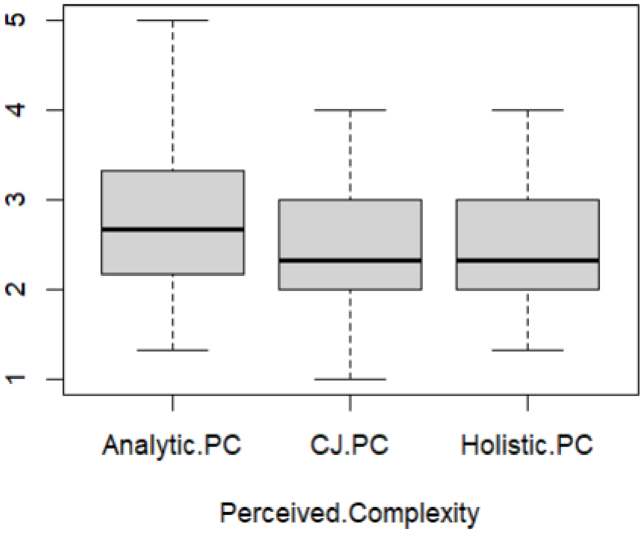
The three conditions were quite similar in their composition of experienced raters. Specifically, in the analytic condition, 77% of the participants had prior experience in assessment, for which 47% applied a rubric (rather than providing an overall rating, based on their intuition. In the holistic condition, 66.7% of the participants indicated prior experience in assessment, of which 38% based the assessment on their intuition (compared to using a rubric). Similarly, 65% of the participants in the CJ condition reported having experience with assessment, one of which applied CJ before, and 46% used a rubric. The time spent on the assessment varied significantly ($F(2,132) = 22.92$, $p < .001$). Post hoc comparisons using Tukey's Honestly Significant Difference (HSD) test at the $\alpha = 0.01$ significance level show significant differences between the analytic ($M = 51.04$, $SD = 15.72$) and holistic condition ($M = 32.12$, $SD = 11.73$), with $t(79.23) = 6.50$, $p < 0.05$, $d = 1.37$. Significant differences were also found between the analytic and CJ condition ($M = 40.59$, $SD = 12.54$), with $t(81.75) = 3.44$, $p < 0.05$, $d = .74$, and the CJ and holistic condition, $t(86.28) = 3.31$, $p < 0.05$, $d = .70$.

Perceptions

Similarly, a significant difference in perceived complexity was revealed by applying an ANOVA ($F(2, 132) = 3373$, $p < .05$). Tukey's HSD test was conducted at a significance level of $\alpha = 0.05$ to examine these differences. Significant differences between the analytic ($M = 2.78$, $SD = .84$) and holistic condition ($M = 2.39$, $SD = .69$) were revealed, with $t(83.21) = 2.41$, $p < 0.05$, $d = .51$. However, no significant differences were observed between the CJ and analytic, $t(84.29) = 1.97$, $p > 0.05$, nor the holistic condition $t(85.66) = .35$, $p > 0.05$. Figure 2 displays the comparison of perceived complexity in each group.

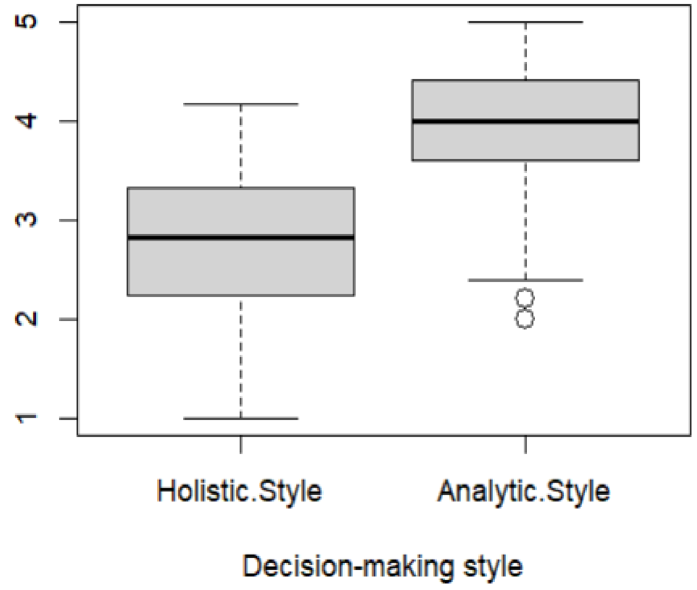
The overall distribution of the holistic and analytic decision-making style is presented in Figure 3. The tendency for analytic decision-making across all conditions (*M* = 3.90*, SD* = .68) was significantly higher than the tendency for holistic decision-making (*M* = 2.77*, SD* = .74), *t*(134) = 13.25, *p* < .01, *d* = 1.60. Per condition, the tendencies for either decision-making style were quite balanced, with the holistic subscale revealing means of *M* = 2.78 (*SD* = .75) in the analytic condition, *M* = 2.68 (*SD* = .78) in the holistic condition, and *M* = 2.85 (*SD* = .67) in the CJ condition. The analytic subscale displayed means of *M* = 3.83 (*SD* = .68) in the analytic condition, *M* = 3.99 (*SD* = .68) in the holistic condition, and *M* = 3.88 (*SD* = .78) in the CJ condition.

Figure 2
Boxplots of Perceived Complexity per Condition



Note. Quartiles are represented. The y-axis represents the Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Figure 3
Boxplots of Decision-making Styles of the Participants

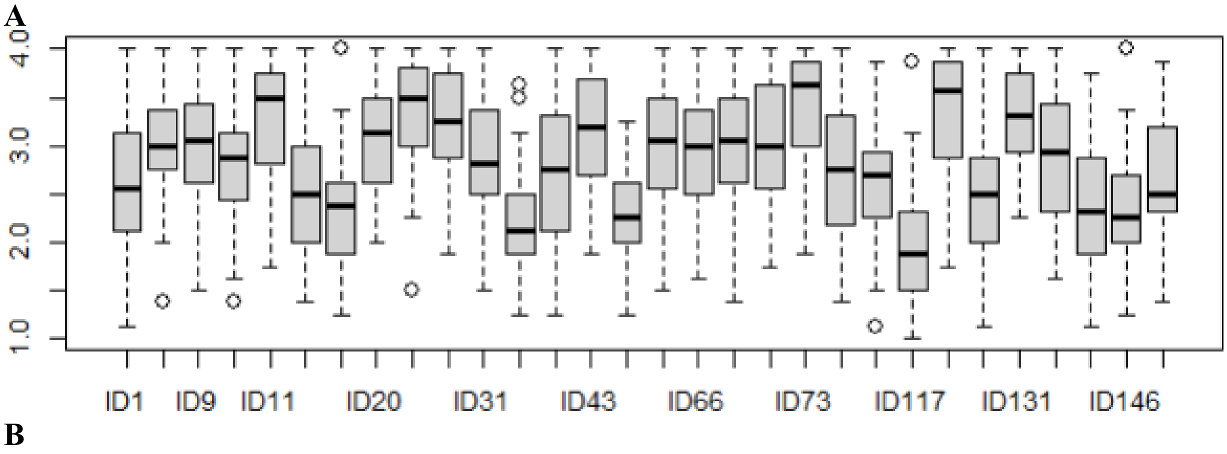


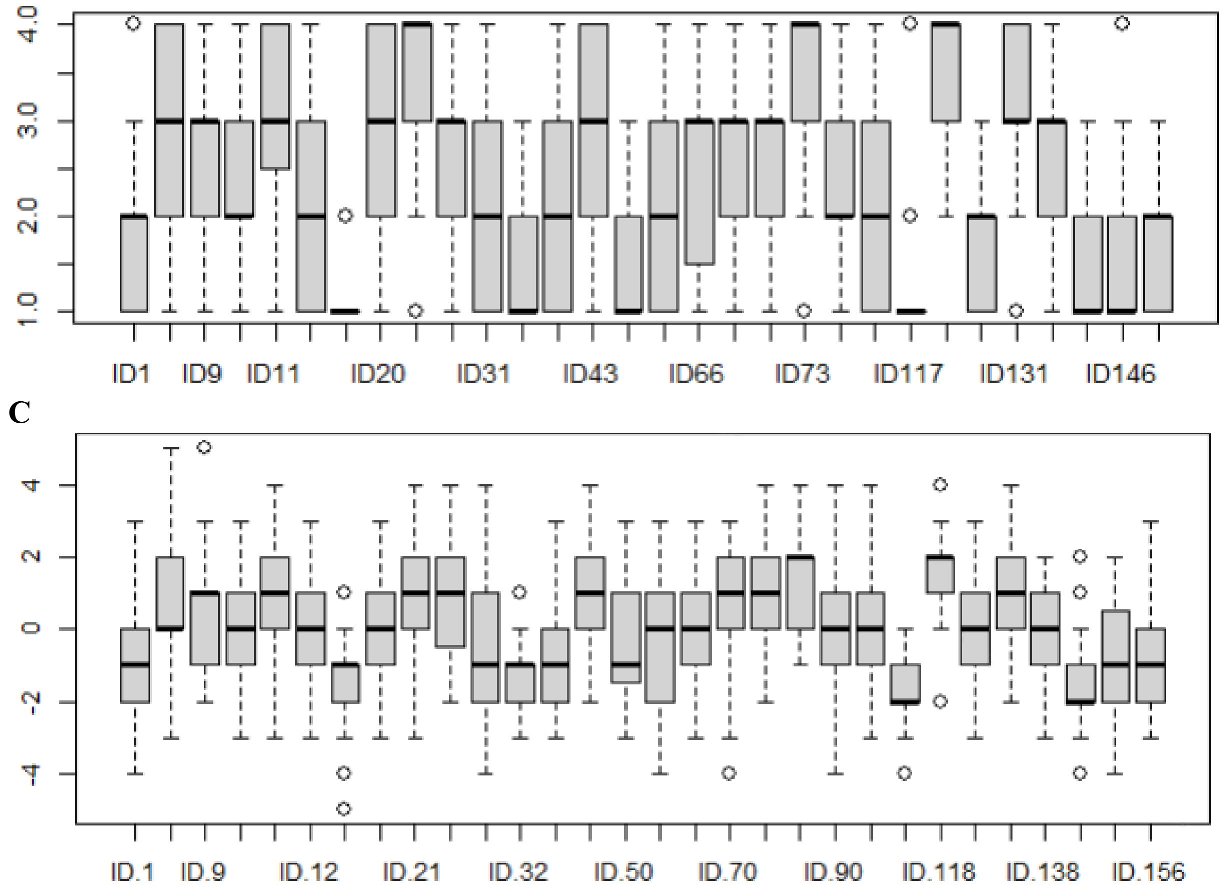
Note. Boxplots are represented. The y-axis represents the Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Inter-Rater Reliability Comparison

An illustration of the distribution of the essay scores per applied method can be found in Figure 4. Wider within variance (size of the grey area per essay) is indicative of lower consistency between raters, as this signifies the distribution of scores awarded for one essay by all raters. The *ICC* estimates and their 95% confidence intervals were computed by the software. For the analytic assessment, the expected inter-rater reliability was ICC_A = .39 (*F* (29, 1247) = 29, 95% CI [.28, .54], $p < .01$), indicating poor to moderate reliability. Conversely, the holistic assessment presented inter-rater reliability of ICC_H = .45 (*F* (29, 1363) = 41, 95% CI [.34, .60], $p < .01$), indicating poor to moderate reliability as well. The inter-rater reliability of the CJ assessment revealed the highest reliability, with *SSR* = .95 and Pearson's *r* = .90, indicating very good reliability. Regarding the first hypothesis that the analytic method would provide better inter-rater reliability than the other methods, these results cannot confirm the hypothesized superiority of inter-rater reliability when using the analytic method. The CJ method provided the highest inter-rater reliability coefficient, thereby partially confirming the second hypothesis. Partial confirmation because while these results confirm the assumption that the CJ method would produce superior inter-rater reliability than the holistic method, they also indicated superior inter-rater reliability than the analytic method.

Figure 4
Boxplot with Quartiles of the Analytic, Holistic, and CJ Rating Distribution





Note. ID = Essay identifier. Panel A: Distribution of essay ratings in the analytic condition. Panel B: Distribution of essay ratings in the holistic condition. Panel C: Distribution of essay ratings in the CJ condition. For A and B the y-axis represents the rating values from 1 (insufficient) to 4 (very good). For C, the y-axis represents the number of losses (-) and wins in comparison to the other essays. Ratings for each essay encompass the average decision of each rater.

Discussion

The objective of this study was to compare the inter-rater reliability of the most common assessment methods of analytic and holistic rating, and the rather novel approach of CJ assessment to determine which aspects of assessment may enhance the inter-rater reliability in the context of higher education. The hypothesis that the analytic method would produce the greatest inter-rater reliability compared with the holistic and CJ methods was not supported. The analytic and holistic methods both resulted in poor to moderate inter-rater reliability, while the CJ method revealed very good reliability among raters. Contrary to what was hypothesized, this study found the holistic method to have somewhat higher consistency between raters than the analytic method. The second hypothesis stated that the CJ method would return higher inter-rater reliability than the holistic but lower reliability than the analytic method. For this, only the first part was confirmed by these results.

A possible explanation for the generally low reliability of the analytic method might be the duration of the study. Assessing 30 essays on each descriptor of the criteria is a rather time-consuming task that might have led to rater fatigue or carelessness. Raters were possibly not as attentive toward the essays or did not consider their choice of rating in the course of the assessment in as much detail as when they started rating, which would lead to lower consistency in the assessment. Moreover, when the assessment method is perceived as complex, it might have been difficult for the raters to apply the criteria consistently, which would lead to low inter-rater reliability. The average (i.e., mean) indication of participants in the analytic condition to perceive the method as relatively more complex than simple suggests that this could indeed be a cause for the low inter-rater reliability. Although the holistic and CJ method also tended to agree—more than disagree—that the method was complex, it appears that the inter-rater reliability of the CJ method was the least affected by this perceived complexity. This was the case even though it was perceived as more complex than the holistic method, albeit the difference in means was not significant. Other possible explanations of low inter-rater reliability are a possible ambiguity in the criteria or inadequate training for the analytic and holistic condition. These topics will be more thoroughly explained in the limitations section.

The slightly higher ICC estimate of the holistic compared to the analytic method might be due to highlighted discrepancies between raters in the analytic condition, due to the analytic raters having provided ratings on each of the criteria descriptors (see also Barkaoui, 2011). Thus, more information was provided per rating. If the essays were, for instance, quite obvious in their quality, holistic raters had it easier to express their agreement by choosing a rating between one and four. In contrast, in the analytic condition—even if the raters agreed on the overall quality of the essay—they provide more variety with subscores on the eight criteria descriptors. Continuing on this example, the obvious differences would be even more readily apparent in the CJ method, as two performances were directly compared to one another. This might further explain the high consistency of raters using the CJ method.

Lastly, as Verhavert et al. (2018) note, the SSR and Pearson's product-moment correlation might overestimate the true inter-rater consistency. This is due to the underlying method for which the set of ratings is split in half and these halves are then correlated. Since both correlated halves share the same set of raters, they are not entirely independent. This might explain the large gap in reliability between the CJ assessment to the other two assessment methods.

Furthermore, exploratory research on participants' decision-making style in terms of holistic and analytic, as well as perceived complexity was performed to examine whether these personal factors affect the implementation of the assessment methods. In line with previous research in this field (e.g., Han, 2021), the analytic method was the most time-intensive and perceived as the most complex. The decision-making tendency of the sample overall was significantly more analytic than holistic, which was quite equally represented in each condition as well. This implies that participants were overall more in favor of standardized decision-making and less likely to rely on their intuition when making decisions, regardless of the method used. Accordingly, the raters overall did not perceive the assessment method as requiring their intuition or gut feeling and agreed that the pre-defined criteria were useful for assessment.

Limitations and Future Recommendations

A noticeable limitation of this study is the use of a convenience sample as it may heighten the risk of confounding variables due to possible limited variability in relevant characteristics in the sample. Thus, by only using first-year psychology students, the findings are not necessarily generalizable to the general rating population in higher education. This is due to the lower levels of experience and expertise with assessment overall and the specific topic at hand. However, a convenience sample of first-year psychology students was deemed sufficient as the main objective was to analyze the inter-rater reliability of the three different assessment methods. Future research on the comparison of assessment methods is advised to sample assessors employed in the field of assessment at higher education institutions for more practical insights.

Furthermore, the reliability coefficients used to compare the inter-rater reliability of the methods are not the same. It ought to be noted that the measures vary in the data structure required for computation, computation itself, and, hence, in the meaning conveyed. Given the different nature of the data, there was not one measure of inter-rater reliability that could be applied to all three conditions. Specifically, the ratings for both the analytic and holistic assessment relied on a scale from 1-4, while the CJ ratings were represented in a binary fashion. Future research should focus on the validity of comparing different measures of reliability, or on finding a measure that can be applied to each method and can be interpreted in the same way, to allow a better comparison. Moreover, the estimation of the SSR as well as Pearson's product-moment correlation lacked a confidence interval, which is essential for taking into account the uncertainty with which the true value of the reliability has been determined (Kelley & Pronprasertmanit, 2016). It is also to be noted that the means rather than the medians were estimated for the analytic ratings, despite the rating scale being ordinary. This should be treated with caution as the relative nature of ordinal data does not necessarily represent equal intervals between each rank. As a result, the differences between the ranks may not be accurately captured by the assigned numerical values that are required for estimating the mean. Since the focus of this study was more on the rating consistency between raters rather than an accurate assessment of the performances, however, estimating the mean values were found to be acceptable for this study.

The rubric used in this study, including the criteria, was loosely adapted from the study by Harsch and Martin (2013) and existing rubrics for other university courses. Although the descriptors of the criteria and anchors that indicate the performance level for each criterion were matched to the requirements of the essay, the ambiguity of their meaning cannot be fully ruled out. That is because the researcher lacked the expertise and experience of constructing rubrics and formulating criteria descriptors that fully and explicitly define the performance under assessment. Thus, it is possible that a certain vagueness in the rubric has led to reduced consistency between raters, simply by leaving room for interpretation. The estimation of the amount of criteria is challenging, for an increased number enhances specificity but may also be overwhelming for raters. Future research may avoid the potential threat of unclarity/unfit quantity of criteria by applying pre-use reviews of experts to investigate its validity, reliability, and practicality (Vercellotti & McCormick, 2021). The training in the respective method was additionally rather short and might not have given the participants sufficient time to familiarize themselves with the method and to discuss unclarities. It might be beneficial to check in with the raters after the training to ensure that no misunderstandings or misinterpretations of the criteria are open before the assessment in future studies.

Given the focus on the inter-rater reliability of the assessment methods, this study cannot answer whether the essay actually assessed the professional skills it was intended to assess, nor does it allow insights into whether the assessed performance aligns with the learning outcomes that were intended to be tested with the assignment. The lack of comparison with other established measures further prevents insights into the assessments' criterion-related and concurrent validity. Being a conditional component of validity, though, reliability is a very relevant aspect to determine the internal validity of the assessment methods (Arkes et al., 2006; Plonsky & Derrick, 2016). To explore the validity of the assessment to a greater extent, future research is thus recommended to determine the most appropriate specified criteria to most accurately assess what is meant to be tested with a given performance and suggested to include some comparison to other performance measures that provide insights into students' performance quality. This is relevant but also challenging, given that the quality of performances can be substantially affected by aspects of the task or of the context that do not necessarily represent the students' ability of the task alone (e.g., Schoonen, 2005).

Lastly, it is important to consider the questionable internal consistency of the perceived complexity scale used in this study. Although the acceptable value for Cronbach's alpha varies between studies and depends on the research done, the general level of acceptable internal consistency is considered to range from .70-.95 (Tavakol & Dennick., 2011). As the perceived complexity scale only consisted of three items, omitting the item that would have enhanced the consistency could have led to reduced or obscured meaning. Since the original scale presented good internal consistency, the rephrasing of the items may have introduced some variability or ambiguity. The results of the perceived complexity are thus to be considered with caution.

Implications

Despite the limitations, the study also provides some notions for research in educational assessments. For instance, this study indicates that the type of assessment method influences the inter-rater reliability of the assessments. The CJ method having the highest reliability among raters, further suggests that students with little content knowledge of the assessed material are most likely to rate the performances consistently when performing pairwise comparisons. The poor to moderate inter-rater reliability of the analytic method may indicate that standardization in a rating procedure does not necessarily result in high inter-rater consistency. With the CJ method relying on comparisons of assessments in the same cohort, it can be considered norm-referenced. The holistic and analytic assessment method, on the other hand, depended on criteria to identify the quality of the performances, they are criteria-referenced. Although the CJ method involved the provision of the same criteria as the other methods, the criteria might not necessarily have been required by the raters to formulate a quality judgment in the CJ condition to the same extent. By providing criteria for the CJ method as well, however, this study contributes to the advice that these approaches are best combined (Lok et al., 2016). Specifically, as also noted in this study, the challenge with criteria-referenced assessment is the design and interpretation of the specific criteria used to assess the performance. Norm-referenced assessment alone, however, relies on comparison with other students, which might undermine the individual achievement of the students themselves and presents as “essentially self-correcting” (Sadler, 2005, p. 187), because it is not compared to an objective standard. As Lok et al. (2016) argue, combining both assessment approaches enhances internal consistency through a resulting feedback loop that compares the performance to both specified criteria and the performance level of the cohort. The herein-presented findings thus offer support for the benefit of combining these grading approaches.

Conclusion

Overall, comparing the inter-rater reliability of analytic, holistic, and CJ performance-based assessment methods yielded valuable insights. While the hypothesized benefit of increased standardization in assessment through the analytic method was not detected, the inter-rater reliability of the holistic approach was similarly weak. More research is needed to investigate the reason behind the low consistency between raters when applying a rubric and to improve the methodological quality. However, this study also strengthens the assumption that relative judgments might in fact positively affect rater consistency. More research is needed to make explicit interpretations of these findings.

References

Allen, J., Ramaekers, G., & Van der Velden, R. (2005). Measuring competencies of higher education graduates. *New Directions for Institutional Research, 126*(126), 49-59. https://doi.org/10.1002/ir.147.

Arboleda, J. C., Meijer, R. R., Tillema, M., Niessen, S. (2023). *Protocol of the article "Unraveling Discrepancies in Analytic and Holistic Assessment Approaches in Performance-based Assessment: A Systematic Review".* [Unpublished manuscript]. Department of Psychometrics, University of Groningen.

Arkes, H. R., Shaffer, V. A., & Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioral Decision Making, 19*(5), 429-439. https://doi.org/10.1002/bdm.503.

Baker, B. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225–248. https://doi.org/10.1080/15434303.2011.637262.

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing, 18*(23), 3237-3242. https://doi.org/10.1111/j.1365-2702.2009.02939.x.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*, 324-345. https://doi.org/10.1093/biomet/39.3-4.324.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848. https://doi.org/10.3102/0034654316672069.

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: what is it, and how is it measured? *Physiotherapy, 8*6(2), 94-99. https://doi.org/10.1016/S0031-9406(05)61211-4.

Council of Europe. (2001). *A common European framework of reference for language learning and teaching*. Cambridge: Cambridge University Press. https://www.coe.int/t/dg4/linguistic/source/Framework_EN.pdf.

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2022). On the bias and stability of the results of comparative judgment. *Frontiers in Education, 6*, 1-10. https://doi.org/10.3389/feduc.2021.788202.

Daniels, V. J., & Harley, D. (2017). The effect on reliability and sensitivity to level of training of combining analytic and holistic rating scales for assessing communication skills in an internal medicine resident OSCE. *Patient Education and Counseling, 100*(7), 1382-1386. https://doi.org/10.1016/j.pec.2017.02.014.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185. https://doi.org/10.1177/0265532207086780.

Han, C. (2021). Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta, 66*(2), 337-361. https://doi.org/10.7202/1083182ar.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281-307. https://doi.org/10.1080/0969594X.2012.742422.

Hartell, E., Buckley, J. (2021). Comparative Judgment: An Overview. In A. Marcus-Quinn & T. Hourigan (Eds.), *Handbook for Online Learning Contexts: Digital, Mobile and Open*. (pp. 289–307). Springer, Cham. https://doi.org/10.1007/978-3-030-67349-9_20.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93-101. https://doi.org/10.1016/j.stueduc.2015.09.004.

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice, 28*(3), 212-227. https://doi.org/10.1080/0969594X.2021.1884041.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002.

Kahneman, D., & Klein, G. (2009). Conditions for Intuitive Expertise A Failure to Disagree. The *American Psychologist, 64*(6), 515-26. https://doi.org/10.1037/a0016755.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69–92. https://doi.org/10.1037/a0040086.

Kelly, K. T., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: a call for clarity. *Assessment in Education: Principles, Policy & Practice*, *29*(6), 674-688. https://doi.org/10.1080/0969594X.2022.2147901.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). "Look! It's a computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 581*, 1-28. https://doi.org/10.1145/3491102.3517527.

Lodato, M. A. (2008). *Going with your gut: An investigation of why managers prefer intuitive employee selection* [Doctoral dissertation, Bowling Green State University]. Retrieved from: http://rave.ohiolink.edu/etdc/view?acc_num=bgsu1206311034.

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450-465. https://doi.org/10.1080/02602938.2015.1022136.

Luce, R. D. (2005). The basic theory. In *Individual choice behavior: A theoretical analysis.* (pp. 1–37). Dover Publications. https://doi.org/10.1037/14396-001.

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies, 55*, 49-71. https://doi.org/10.1007/s40841-020-00163-3.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34-43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x.

Meadows, M. & Billington, L. (2005). A review of the literature on scoring reliability. *National Assessment Agency.* Retrieved from: https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf.

Notion. (n.d.). Choosing and signing up your judges to a task. *No More Marking*. Retrieved February 14, 2023, from https://nmm.notion.site/Choosing-and-signing-up-your-judges-to-a-task-562fbbfa564040c3aab11f73ec30f001.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, 100329. https://doi.org/10.1016/j.edurev.2020.100329.

R Core Team (2017) R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859.

Sadler, R. D. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 30*(2), 175–194. https://doi.org/10.1080/0260293042000264262.

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.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*(3), 178–200. https://doi.org/10.1037/h0023624.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30. https://doi.org/10.1191/0265532205lt295oa.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. *Studies in Immigrant English Language Assessment*, *1*, 159-189. http://www.ameprc.mq.edu.au/__data/assets/pdf_file/0019/241435/Research_Series_11.pdf#page=87.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116%2Fijme.4dfb.8dfd.

Thurstone, L.. (1994). A Law of Comparative Judgment. *Psychology Review. 34,* 273-86. https://doi.org/10.1037/0033-295X.101.2.266.

Vercellotti, M. L., & McCormick, D. E. (2021). Constructing Analytic Rubrics for Assessing Open-Ended Tasks in the Language Classroom. *TESL-EJ, 24*(4), n4. https://eric.ed.gov/?id=EJ1288720.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice. 26*(5), 541-562. https://doi.org/10.1080/0969594X.2019.1602027.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment? *Applied Psychological Measurement, 42*(6), 428-445. https://doi.org/10.1177/0146621617748321.

Walland, E. (2022). Judges' Views on Pairwise Comparative Judgement and Rank Ordering as Alternatives to Analytical Essay Marking. *Research Matters, 33*, 48-67. https://eric.ed.gov/?id=EJ1343622.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400-409. https://doi.org/10.2307/357792.

Appendix A: Information About the Study

"HOW RELIABLE IS PERFORMANCE-BASED ASSESSMENT? COMPARING HOLISTIC, ANALYTIC, AND COMPARATIVE JUDGMENT APPROACHES"

EC code: PSY-2223-S-0236

➢ **Why do I receive this information?**

As a first-year Psychology bachelor's student who is registered in the SONA system, you are invited to participate in this study. I, Charlotte Sievers, and my research team from the Department of Psychometrics and Statistics (J.C. Arboleda, dr. A.S.M. Niessen, and prof. Dr. R. Meijer) are doing research on the quality of different performance-based assessment methods. We are looking for first-year students to test our research concept.

➢ **Do I have to participate in this research?**

Participation in the research is voluntary. However, your consent is needed. Therefore, please read this information carefully. Ask all the questions you might have, for example, because you do not understand something. Only afterward do you decide if you want to participate. If you decide not to participate, you do not need to explain why, and there will be no negative consequences for you. You have this right at all times, including after you have consented to participate in the research.

➢ **Why this research?**

Higher education institutions are a place for students to develop professional skills (e.g., critical thinking) but how to ideally conduct performance-based assessments of these skills is still unclear. In our study, we are looking into several different approaches to performance-based assessment.

- **What do we ask of you during the research?**

Firstly, we ask for your consent to participate in this study, as well as to provide some information about yourself (age). After a short training, you will proceed to grade 30 short assignments (max. 250 words each). Then, we will ask you to fill in a short questionnaire on how you perceived the grading. Every step of the study will take place in our lab where you will be provided with a computer; the duration of the whole study will be approximately 60 minutes.

- **What are the consequences of participation?**

Participation in our study will give you some insights into how performance-based assessment is performed in higher education. Other than that, there are no direct or indirect positive or negative consequences of participating in this study. In addition, you will be rewarded 1.5 SONA credits [this will be filled out after the amount of credits is determined by the SONA administrator] for complete participation.

- **How will we treat your data?**

All answers given while filling in the questionnaire will be treated confidentially. This means

that the questionnaires and answers are kept secure on the university drive and that only the

researchers can see the completed Questionnaires. As soon as data collection is complete

your SONA number is replaced with a random identifier, and no data can be directly traced back to you. The anonymized data will be securely stored on the university server for 10

years. You have the right to request insight and removal of your data until 2.6.2023.

- **What else do you need to know?**

You may always ask questions about the research: now, during the research, and after the

end of the research. You can do so by speaking with one of the researchers present right now

or by emailing c.sievers@student.rug.nl or j.c.arboleda.cardona@rug.nl.

Do you have questions/concerns about your rights as a research participant or about the

conduct of the research? You may also contact the Ethics Committee of the Faculty of

Behavioural and Social Sciences of the University of Groningen: ec-bss@rug.nl.

Do you have questions or concerns regarding the handling of your personal data? You may also contact the University of Groningen Data Protection Officer: privacy@rug.nl.

Appendix B: Information for the Holistic and CJ Condition

Essay Requirements:

Write a short conclusion in which you answer the following questions on the basis of the analyses you conducted above:

- ➔ What can you conclude about the predictive validity of the four different selection instruments?
- ➔ What can you conclude about the incremental validity of the conscientious scale and the math test on top of the cognitive test?
- ➔ How do you explain the incremental validity of both instruments?
- ➔ Which instrument or which combination of instruments would you advise the university?
- ➔ To what extent does the use of this instrument/these instruments contribute to the aim of increasing the quality of the admitted students (that is: students with a mean grade of 6 or higher)?

Use between 100 (minimum) and 250 (maximum) words.

Criteria	**Description**
Task Fulfillment	A correct conclusion for the predictive validity of the four different selection instruments is provided.
	Advice on which selection instruments the university should use is provided.
	Incremental validity (i.e., the added benefit that a specific predictor variable has over another predictor) is correctly explained and considered for both instruments.
Organization	The essay consists of at least a topic sentence, a supporting sentence, and a concluding sentence.
	All sentences relate to the main theme of the instruments' validities.
	The essay has a coherent narrative with clearly explained (sub)topics.
Academic Writing Style	The text is objective and concise.
	The language is grammatically correct, appropriately formal, and avoids colloquial language.

Addition only for the Holistic Condition

Rating Scale Description

Very good	Good	Sufficient	Insufficient

Appendix C: Information of the Analytic Condition

Essay Requirements:

Write a short conclusion in which you answer the following questions on the basis of the analyses you conducted above:

- ➔ What can you conclude about the predictive validity of the four different selection instruments?
- ➔ What can you conclude about the incremental validity of the conscientious scale and the math test on top of the cognitive test?
- ➔ How do you explain the incremental validity of both instruments?
- ➔ Which instrument or which combination of instruments would you advise the university?
- ➔ To what extent does the use of this instrument/these instruments contribute to the aim of increasing the quality of the admitted students (that is: students with a mean grade of 6 or higher)?

Use between 100 (minimum) and 250 (maximum) words.

The Rating Scale and the Assessment Criteria

	Very Good	**Good**	**Sufficient**	**Insufficient**
Task Fulfillment				
A correct conclusion for the predictive validity of the four different selection instruments is provided.	The essay clearly describes a correct conclusion about the predictive validity of the four different selection instruments that are persuasively justified.	The essay describes a correct conclusion about the predictive validity of the four different selection instruments that are partially justified.	The conclusion about the predictive validity of some selection instruments is described.	The conclusion about the predictive validity of the four different selection instruments is not described or is difficult to grasp/understand.
Advice on which selection instruments the university should use is provided.	The advice on which combinations of instruments the university should use is exceptionally sound and supported by argumentation.	The advice on which combinations of instruments the university should use is sound and supported by argumentation.	The advice on which combinations of instruments the university should use is provided but unclear.	No advice is provided, or difficult to understand.
Incremental validity (i.e., the added benefit that a specific predictor variable has over another predictor) is correctly explained and considered for both instruments.	The considerations of the selection instruments are correct, critically analyzed, and insightful with regard to their incremental validity.	The essay correctly considers the incremental validity of the instruments.	The essay considers the incremental validity of some instruments.	The considerations of incremental validity of the instruments are not stated or are difficult to understand.
Organization				

The essay consists of at least a topic sentence, a supporting sentence, and a concluding sentence.	The essay's topic is clear, consistently introduced by a topic sentence, supported by related sentences, and a clear concluding sentence.	The essay's topic is clear, in most cases introduced by a topic sentence, and supported by related sentences. A concluding sentence is provided.	The topic of the essay is mostly clear but sometimes too broad/narrow, topic sentences are sometimes lacking, and/or sentences are unrelated.	The topic of the essay is mostly unclear, topic sentences are lacking, and/or sentences are unrelated.
All sentences relate to the main theme of the instruments' validities.	The essay explicitly relates to the posed questions.	The essay relates to the posed questions, and most of the claims are explicitly related.	Most claims in the essay relate explicitly or implicitly to the posed questions.	Claims in the essay appear to be unrelated to the posed questions.
The essay has a coherent narrative with clearly explained (sub)topics.	The essay is ordered logically and the answers to the questions transition fluently. Very good coherence.	The essay is ordered logically and most of the transitions between sentences are fluent. Good coherence.	Most of the essay is ordered logically, the transitions are not always fluent and sometimes abrupt. There is (some) coherence.	The essay is ordered illogically, the transitions are abrupt. There is no coherence.
Academic Writing Style				
The text is objective and concise	The essay describes the topic accurately and impartially. It always contains concise and relevant descriptions of the aspects of the posed questions.	The essay generally describes the topic accurately and impartially. It mostly contains concise and relevant descriptions of the aspects of the posed questions.	The essay describes the topic somewhat accurately and mostly impartially. At times biased or imprecise terminology is used. There is some irrelevant information.	The essay uses biased or imprecise terminology. The text is wordy and long-winded.

The language is grammatically correct, appropriately formal, and avoids colloquial language.	The text uses entirely appropriate language and terminology for an academic paper. The use of language and punctuation avoids ambiguity and any potential for misunderstanding.	The text generally uses appropriate language and terminology for an academic paper. The use of language and punctuation has minor issues but mostly avoids ambiguity and the potential for misunderstanding.	There is some inappropriate use of language, terminology, or punctuation but without impairing the overall understanding of the research question/focus. The use of language and punctuation is sufficient but may exhibit some ambiguities.	The use of language, grammar, or punctuation impairs understanding of the text or leads to a misinterpretation of the research question/focus.

Appendix D: Decision-Making Scale Items

Adapted from the Selection Decision-Making Style (SDMS) scale (Lodato, 2008), rated on a five-point Likert scale (1 = strongly disagree; 5 = strongly agree)

The holistic subscale consists of items:

1. I think it was important to rely on my "gut" when rating the students' work
2. I think it was important to rely on my instincts when rating 'the students' work
3. I think it was important to rely on my intuition when rating the students' work
4. I think that rating a student's work is more of an art than a science
5. You can't always explain why a student's work is the best one – You just know it
6. You can "read between the lines" to detect whether a student did well

The analytic subscale consists of items:

7. Basing scores on pre-defined criteria is a valid way to rate students' work
8. Assessments with specific criteria are an effective way to assess students' work
9. Assessing abilities with standardized rubrics is an effective way to evaluate a student's ability
10. I believe that computing a score based on separately rated criteria is a good way to rank students
11. I do not believe that formal ratings (e.g., rubrics) are useful for assessing students (r)

Appendix E: Perceived Complexity Scale Items

Adapted from the perceived complexity scale (Langer et al., 2022), rated on a five-point Likert scale (1 = strongly disagree; 5 = strongly agree)

1. The assessment method I used to grade students and how it works is easy to understand (r)
2. *Holistic/analytic/CJ assessment* is understandable even for laypeople. (r)
3. *Holistic/analytic/CJ assessment* is complex.