Education and Science

Original Article

2025 Supplement 1 93-110

Adaptive Selection Algorithm and Standard Error Termination Rule in Comparative Judgement: An Application for Assessing Writing Skills *

Sungur Gürel [1], Murat Doğan Şahin [2], İbrahim Uysal [3], Ali İhsan İbileme [4], Tuba Gündüz [5]

Abstract

This study aims to examine the scoring reliability of comparative judgement under different sample sizes and standard error termination rule conditions. For this purpose, a Monte Carlo simulation study with 9 conditions and 82 iterations was conducted with sample sizes of 250, 500 and 1000 and standard error termination rules of 0.40, 0.35 and 0.30. In addition, a application for assessing writing skills was conducted with a sample of 50 students using the standard error termination rule of 0.40 and a maximum number of comparisons of 40. In the simulation study, scoring reliability was determined by true reliability, rank order accuracy and scale separation reliability. In the application, the correlation between scores that are obtained with a holistic rubric and ability estimates that are obtained with adaptive comparative judgement as well as the correlation between scores that are obtained using an analytic rubric and ability estimates that are obtained with adaptive comparative judgement were examined. In addition, scale separation reliability was calculated to obtain ability estimates using adaptive comparative judgement. The simulation results showed a high level of reliability in all conditions. Moreover, reliability was high, independent of the sample size. We conclude that stricter standard error termination rules lead to higher levels of reliability, but this requires performances to be subjected to a higher number of pairwise comparisons. The application results showed high scale separation reliability of .89 and correlations of over 0.70 with the scores obtained by using both holistic and analytic rubrics. Overall, the results of the study suggest that adaptive comparative judgement can be used in both classroom and large-scale assessment applications. In addition, adaptive comparative judgement is considered advantageous because it is easier to administer, does not require a difference in the testing process, and places the abilities on a continuous scale.

Keywords

Comparative judgement
Holistic assessment
Scale separation reliability
Pairwise comparison

Article Info

Received: 10.06.2024
Accepted: 01.07.2025
Published Online: 03.03.2025

DOI: 10.15390/EB.2025.14123

* A part of this study was presented at the International Symposium on Measurement, Selection and Placement held between 4-6 October 2024 as an oral presentation.

[1] Siirt University, Faculty of Education, Department of Educational Sciences, Türkiye, s.gurel@siirt.edu.tr
[2] Anadolu University, Faculty of Education, Department of Educational Sciences, Türkiye, muratdogansahin@gmail.com
[3] Bolu Abant İzzet Baysal University, Faculty of Education, Department of Educational Sciences, Türkiye, ibrahimuysal@ibu.edu.tr
[4] Eskişehir Technical University, Continuing Education Application and Research Center, Türkiye, aiibileme@gmail.com
[5] Muğla Sıtkı Koçman University, Faculty of Education, Department of Educational Sciences, Türkiye, tubagunduz@mu.edu.tr

Introduction

Rubrics are used to ensure scoring reliability in the assessment of the performance outcomes that are produced by individuals. In order to prevent biases that may arise from a single rater, the most commonly used method is to score the performance independently by at least two raters with rubrics consisting of three to five categories, and a high level of agreement between the two scoring is expected. The main purpose of using rubrics with a maximum of five categories is to keep the agreement between the raters as high as possible. In fact, as the number of categories increases, the level of disagreement may also increase (Goossens & De Maeyer, 2018). This situation brings with it the structuring of the performance tasks in accordance with the rubric with a limited number of categories, in other words, preferring higher reliability over higher validity (van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2019).

In particular, rubrics that are used in the assessment of higher-level skills have more than one component; therefore, the use of analytical rubrics requires each component to be scored separately. This increases the time allocated for assessment. Another important issue is that in the assessment of performance outcomes that focus on skills such as creativity, the outcome that is the basis of the assessment has a different identity than the sum of the components of the rubric; in other words, the asssessed skill is different from the sum of the components of the analytic rubric (Jones & Davies, 2023). In order to avoid this problem, if a holistic rubric is used instead of the analytical rubric, it is not possible to assess with enough precision to reveal the differences in individuals' performances. This dilemma causes a significant limitation in situations where differences among individuals need to be revealed in terms of assessed performance outcome.

Another limitation of traditional systems based on two raters is the need for expert raters. What is meant by expertise here is not only competence and experience in the field, but also standardization in the scoring process. In cases where the number of expert raters is limited, there is a need for a scoring approach where decisions can be made based on an approach that requires relatively less expertise.

Although it seems possible to assess performance with reference to an absolute level of competence with the help of a rubric, it is known that raters are influenced by their previous scoring and this is reflected in their scoring performance (Bloxham, 2009; Crisp, 2013). This is in line with experimental psychologist Laming's (2003) statement, “There is no absolute judgement. All judgements are comparisons of one thing with another.” Laming states that although an absolute assessment is desired due to the nature of the trait to be assessed and the purpose of the assessment process, this is not possible due to the nature of the rating behaviour. It seems possible to minimize all these limitations related to the use of traditional methods based on rubrics in the holistic assessment of performance with the use of comparative judgements (CJ).

The CJ is essentially based on Thurstone's principle of comparative judgements, where the rater is presented with the performance of two different individuals and is expected to decide only which performance is holistically better than the other. In this way, each performance is subjected to multiple pairwise comparisons. Ability estimation is performed by the Bradley-Terry-Luce (BTL) model (Bradley & Terry 1952; Luce 1959), which is very similar to the Rasch model. The basic premise of BTL is that it is easier for raters to make relative judgements than absolute judgements (Bramley, 2005). In addition, two raters may not agree on the scoring of two different performances, but they are likely to agree on which one is “better” (Bramley, 2007; Steedle & Ferrara, 2016). In fact, it is considered sufficient to provide limited training to the raters in the assessment using the pairwise comparison method (Heldsinger & Humphry, 2013).

With its advantages, CJ has found its place primarily at the K-12 level with its use in teachers' performance assessment practices in the classroom (Steedle & Ferrara, 2016). In particular, the increase in research on the reliability of teachers' ratings and the observation that teachers' ratings are mostly variable in these studies (Humphry & Heldsinger, 2019) has led to an increase in the number of applications of the CJ that are much easier than rubric-based assessment and thus provide highly reliable results. In this regard, nomoremarking.com is particularly prominent with its studies based on comparative judgements of teachers in approximately 2000 schools in the United Kingdom for five years to assess writing skills longitudinally at the primary education level (Christodoulou, 2024). In other words, in recent years, in both formative and summative assessment practices in the classroom and in large-scale applications, the CJ has found itself more and more in practice.

In assessment practices in Turkey, it can be said that the use of multiple-choice items in national exams for placement has increased the use of multiple-choice items at all grades. However, the Ministry of National Education (MoNE), with its new curricula within the framework of the Turkish Century Education Model, has published a regulation that directs teachers to use open-ended items instead of multiple-choice items by emphasizing productive language skills (MoNE Measurement and Evaluation Regulation, 2023). As a result, an increase in the use of open-ended items is expected at the K-12 level. With this regulation, in addition to this in-class use, some of the written exams at the secondary education level were structured as common written exams across the country. In the continuation of this process, it can be expected that open-ended items will also be used in the high-stakes national placement exams organized by MoNE. Since it is unlikely that the number of expert raters will reach a sufficient level in exams to be administered to large masses of hundreds of thousands, there may be significant concerns about scoring reliability. The fact that the CJ enables high reliability with non-expert raters in both classroom applications and large-scale high-stakes exams offers an important opportunity to eliminate possible problems that may arise in these administrations in Turkey.

Development of Comparative Judgement

The main components of the CJ implementation can be analyzed in four parts: selecting object pairs to be compared, ability estimation, reliability, and termination rule. Although the first step is selecting the object pairs to be compared, the introduction of these components starts with ability estimation, since chronologically the work on ability estimation is at the forefront in the development of the model.

Ability estimation

Thurstone's (1927) law of comparative judgement provided the first method for estimating distances between objects on a latent scale. Subsequently, Bradley and Terry (1952) and Luce (1959) showed how logistic functions can be applied to analyze comparative judgement data, and Andrich (1978) showed that Thurstone's model overlaps with the Rasch logistic model:

$$P_k(a_j) = p(X_{jk} = 1|a_j, a_k)\frac{\exp(a_j - a_k)}{1 + \exp(a_j - a_k)}$$

The expression $P_k(a_j)$ or $p(X_{jk} = 1|a_j, a_k)$ indicates the probability that object j is preferred to object k. In this case $X_{jk} = 1$means that object j is preferred to object k. Here a_j and a_k represent the ability estimates of objects j and k in logit units, respectively.

The CJ practices are based on a relative assessment on the basis that an absolute assessment is not possible due to the nature of human beings. This may bring to researchers' minds the issue of how this method can be used in situations where there are absolute assessment expectations. Here, anchor objects are used to place the ratings on an absolute scale (Heldsinger & Humphry, 2013; Using anchors to link judging sessions, 2016). Accordingly, objects that have been rated as absolute by an expert group

are also included in the CJ process. In this way, it is possible both to place the ratings on an absolute scale and to assess individuals by equating them on the same scale in applications performed for the same purpose at different times (Benton, 2021).

Selecting the Object Pairs to be Compared

It is seen that studies on CJ have increased in recent years (Benton, 2021; Bramley & Vitello, 2019; Crompvoets, Béguin, & Sijtsma, 2020; Crompvoets, Béguin, & Sijtsma, 2022; Holmes, Meadows, Stockfor, & He, 2018; Humphry & Heldsinger; 2019; Lesterhuis, Bouwer, Van Daal, Donche, & De Maeyer, 2022; van Daal et al., 2019; Verhavert, Bouwer, Donche, & De Maeyer, 2019; Verhavert, Furlong, & Bouwer, 2022). Among these, the study by Pollit (2012) is particularly important. This study introduced the idea of selecting object pairs with a method called the Swiss System instead of randomly generating pairs of objects, and with this selecting process, the method was named scoring with adaptive comparative judgement (ACJ). However, this selection method has been criticized for allegedly leading to unrealistically high estimates of scale separation reliability, which is a measure of reliability in CJ applications (Bramley, 2005; Bramley & Vitello, 2019). Considering that high scale separation reliability should be achieved with a low number of pairwise comparisons for assessment precision in CJ, it can be said that it is a necessity to ensure adaptiveness in selecting objects. This necessitated the emergence of new adaptive methods other than Pollit's proposal.

The process in the ACJ can be likened to a typical computerized adaptive test (CAT) administration (Crompvoets et al., 2020). This is particularly due to the adaptive nature of selecting object pairs to be compared in ACJ. Similar to the item selection algorithm in the CAT, the selection of object pairs to be compared in ACJ is done using the Fisher information function (Pollit, 2012). However, in the method proposed by Crompvoets et al. (2020), after each tentative ability estimation, the object with the highest standard error is determined from the objects in the pool and its counterpart is selected through a probability density function. With $\theta_i \sim N[\theta_i, \mathrm{SE}(\theta_i)]$ (where SE is the estimated standard error and θ_i is the estimated tentative ability level), the probabilities are obtained by dividing the density value of each possible object j by the total density value of each possible object j to be selected. When selecting according to the probability density function, the selected object will be similar to the initial object in terms of ability level but with a higher standard error. This adaptive selection method is utilized in this study.

Reliability

Reliability in CJ is calculated by scale separation reliability (SSR). SSR provides information about the level of agreement of the raters regarding the levels of performances (Verhavert et al., 2019). SSR, which is an indicator of reliability, was formulated by Andrich and Douglas (1977) as follows (as cited in Gustafsson, 1977);

$$\mathrm{SSR} = \frac{\sigma_a^2 - MSE}{\sigma_a^2}$$

in the above formula, σ_a^2 represents the variance of the ability estimates and MSE represents the arithmetic mean of the squares of the standard errors as follows:

$$MSE = \frac{\sum_j^n se_{aj}^2}{n}$$

where *se* represents the standard error.

Termination Rule

Another important component in CJ is the termination rule. When the studies are analyzed, it is seen that a termination rule based on a fixed number of pairwise comparisons or an average number of pairwise comparisons is mostly used (Crompvoets et al., 2020; Lesterhuis et al., 2022; Pollit, 2012; Sims, Cox, Eckstein, Hartshorn, Wilcox, & Hart, 2020; Thwaites, Kollias, & Paquot, 2024). However, subjecting object pairs at different ability levels to an equal number of pairwise comparisons leads to

different levels of error in the ability estimates; in this case, the standard error of the estimates increases as we move to the two ends of the ability distribution (Crompvoets et al., 2020; Uysal, Gürel, Şahin, İbileme, & Yıldırım Görgülü, 2024). In order to find a solution to this issue, Verhavert et al. (2022) used SSR as a termination rule, according to which the pairwise comparison process is terminated when a predetermined reliability value is reached. However, even a high SSR does not guarantee a given standard error value for all ability estimations. Therefore, it would be more accurate to ensure that the standard errors of the ability estimates of all objects are at a predetermined level. In other words, similar to the CAT applications, the termination rule should be based on the standard error of the ability estimates.

Current Study

While some of the studies in the literature use semi-random selection for selecting objects, it is seen that studies using adaptive selection have come to the fore in recent years (Crompvoets et al., 2020; Crompvoets et al., 2022; Holmes et al., 2018; Lesterhuis et al., 2022; Pollit, 2012; Sims et al., 2020; Thwaites et al., 2024; van Daal et al., 2019; Verhavert et al., 2019; Verhavert et al., 2022). In these studies, SSR values between 0.70 and 0.95 were obtained with different numbers of pairwise comparisons based on the domain to which the objects belong. However, the termination rule is generally based on the average number of pairwise comparisons per object; as mentioned above, there have also been new studies based on SSR-based termination. However, these termination approaches can result in low standard errors for some objects and high standard errors for others. It would be possible to set a limit value for the standard error of each ability estimation by using a standard error-based termination rule, just like in CAT applications. However, there is no study in the literature that uses a standard error-based termination rule in ACJ studies. In this study, the performance of an algorithm based on the adaptive selection of objects to be compared and the standard error termination rule is demonstrated for the first time. For this purpose, the first part of the study examines the changes in the average number of pairwise comparisons, true reliability, rank order accuracy, and SSR values when different standard error values are used as a termination rule by using data sets generated for different sample sizes. The results of this Monte Carlo simulation study, which compares the performance of the termination rule based on standard errors of 0.30, 0.35 and 0.40 in three different sample sizes that can be considered as medium (250) and large (500 and 1000) considering the CJ studies, will guide researchers in real-world applications. The choice of 0.30, 0.35 and 0.40 as standard error values is due to the fact that these values are the most preferred values within the framework of the standard error termination rule in CAT applications. Of course, it can be predicted that reliability will increase as the standard error value decreases; however, it should not be ignored that there will be a significant increase in the number of pairwise comparisons as a result of this. Therefore, the main purpose of this research is to find the most appropriate standard error value and to reach a value where the number of pairwise comparisons in the application is reasonable, in other words, to focus on assessment precision. It should be kept in mind that while very good values for standard error and SSR can be obtained, dramatic increases in the average number of pairwise comparisons would create a significant usability problem in real-world applications. For this reason, in the second part of the study, the real-world application, a standard error value of 0.40 was chosen for the termination rule. However, a maximum number of pairwise comparisons was determined as in the CJ applications. Accordingly, even if the standard error of 0.40 was not achieved in 40 pairwise comparisons, the pairwise comparison process for the relevant object was terminated. The maximum value of 40 pairwise comparisons was determined in alliance with the simulation results.

Method

The research consists of two parts: simulation study and application. The simulation study and the application are described below, respectively.

Simulation Study

The first part, a Monte Carlo simulation study, used a fully crossed design. Three sample sizes (250, 500 and 1000), three standard error values for termination (0.40, 0.35 and 0.30), 2 factors and 9 conditions were studied. Since the current research examines CJ in large-scale applications, 500 and 1000 conditions were defined for the sample size as well. In addition, the 250 condition was considered to represent a medium sample size. Steedle and Ferrara (2016) used a sample of 200 objects in their study, while Pollitt (2012) used a sample of 1000 objects. The study is designed to examine the variation between these sample sizes.

Objects' abilities were generated using a normal distribution with a mean of 0 and a standard deviation of 1. As the standard deviation values of the generated abilities change, the standard error estimates vary. In order to generate reasonable standard errors, the standard deviation was taken as 1 as suggested by Crompvoets, Béguin, & Sijtsma (2021). The approach of Crompvoets et al. (2020) is taken as a reference for the coding of the adaptive selection algorithm and ability estimates. The coding of the standard error termination rule is original. The data were generated and analyzed using R software (R Core Team, 2023) on a Linux operating system. The analysis was performed on the Turkish National Science Infrastructure (TRUBA) of Tübitak and 56-core computers were used. Tasks were distributed to the cores in the computer by parallelization. The *doParallel* package (Daniel, Microsoft Corporation, Weston, & Tenenbaum, 2022) was used at this stage. *Slurm* was used to send tasks to the TRUBA system. Due to the large size of the data matrix, the TRUBA system was resorted to because of the time constraints when running the TRUBA system. For this reason, 82 iterations could be completed for each condition in the study.

When evaluating the findings obtained as a result of the simulation study, the average number of pairwise comparisons indicating the number of times each object was compared, the SSR, rank order accuracy and true reliability values described above were used. Rank order accuracy is calculated via the Spearman rank difference correlation between the generated and predicted ability ranks, while true reliability is calculated via the square of the Pearson product-moment correlation coefficient between the generated and predicted abilities.

Application

The application was carried out under an observational design. In this context, students' abilities were estimated through ACJ. In the application, a software that is developed by the researchers within the scope of a project called "Development of a System and Software for Scoring Open-Ended Items with Adaptive Comparative Judgement " supported by ÖSYM was used. Within the scope of the research, the data set in the The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus (Crossley et al., 2023) was used. The ELLIPSE Corpus dataset is a dataset collected as part of an automatic rating study and made available to researchers for scientific purposes through a CC BY-NC-SA 4.0 DEED Attribution-NonCommercial-ShareAlike 4.0 International license. The dataset includes the essays written by 8th-grade students in the USA, whose second language is English, on "the effects of technology on human life" and the scores obtained by the assessment of these essays by expert raters. The essays were previously scored by two raters both holistically in terms of general English language proficiency and analytically in terms of cohesion, syntax, vocabulary, phraseology, grammar and conventions. Both the scores obtained holistically and the scores obtained analytically with the sum of the 6 sub-dimensions were transferred to this study. Out of the 250 writing samples given in the relevant chapter, 50 were selected in such a way that the standardized analytic total score

followed the standard normal distribution as closely as possible. The standardized analytical total scores ranged from -1.88 to 2.63 with a mean of 0.10 and a standard deviation of 0.96. The rationale for this choice is that ability levels in the simulation study were generated by a standard normal distribution. The holistic scores for the English language proficiency of the 50 responses ranged between 2 and 5. One student scored 2, seven students scored 2.5, thirteen students scored 3, eighteen students scored 3.5, eight students scored 4, two students scored 4.5 and one student scored 5. The total number of words in 50 writing samples ranged between 152 and 1532, with a median of 449 words and an average of 535 words. Since it is aimed to assess the samples holistically in terms of language skills, it is suitable for rating with CJ. Full details about the relevant dataset can be found at https://github.com/scrosseye/ELLIPSE-Corpus.

The ACJ was performed by four of the authors of this study. The rater authors had English proficiency at the C1 level and did not receive any scoring instructions other than selecting the better writing performance in general. The main purpose here was to test the expectation of achieving high reliability even with non-field expert raters, which is one of the main advantages of the CJ. In order to evaluate the performance of the ACJ after the application, the correlation between the ability estimates obtained with the ACJ and the English language proficiency holistic scores and standardized analytic total scores scored by the experts was reported along with the SSR.

Results

Simulation Study Results

In this study, which aims to determine the consistency of ability estimates when adaptive selection is preferred in CJ, at different sample sizes and when different standard error termination rules are applied, the relationship between ability estimates and standard errors is first presented. To determine the reliability of the ability estimates, the true reliability, rank order accuracy and SSR obtained under different simulation conditions are reported. Finally, to evaluate the usefulness of the ACJ, we report the sample size and the average number of pairwise comparisons to satisfy each termination rule.

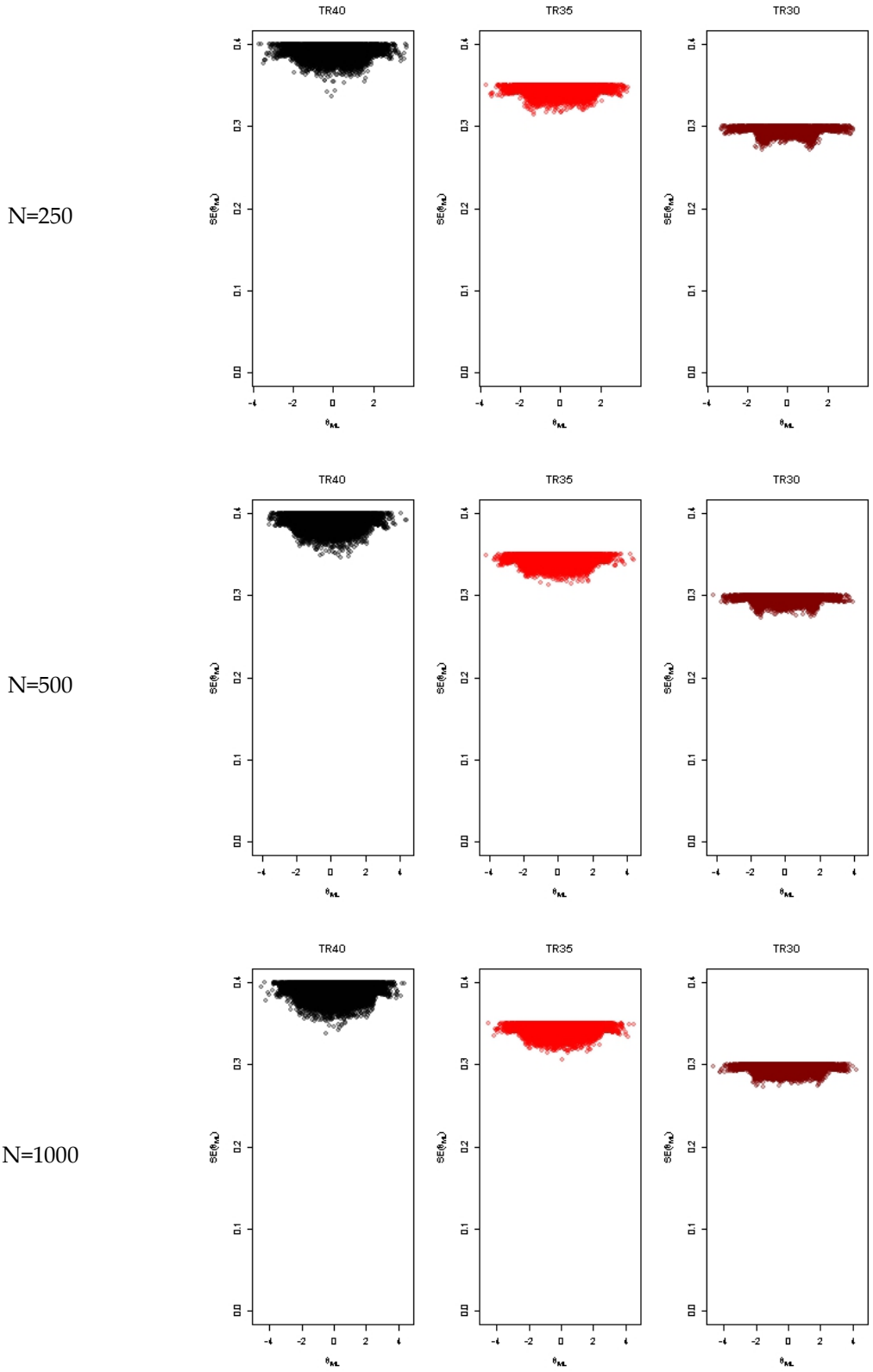

Figure 1. The relationship among ability estimates and standard error in different sample sizes

Figure 1 reveals the relationship between the ability estimates obtained with different standard error termination rules at different sample sizes and the standard errors of the ability estimates. In Figure 1, TR40 refers to the results obtained with a standard error termination rule of 0.40, TR35 refers to the results obtained with a standard error termination rule of 0.35, and TR30 refers to the results

obtained with a standard error termination rule of 0.30. In all the conditions analyzed, the standard error value, which is the basis for the termination rule, was achieved. In the centre of the ability distribution, much better standard error estimates were obtained than with the termination rules. In addition, the termination rule of 0.40 produced standard error estimates over a wider range, while the termination rule of 0.30 produced standard error estimates over a narrower range.

Table 1. Averages for true reliability, rank order accuracy and SSR by diferent simulation conditions

	True Reliability			Rank Order Accuracy			SSR		
N	**TR40**	**TR35**	**TR30**	**TR40**	**TR35**	**TR30**	**TR40**	**TR35**	**TR30**
250	0.859	0.888	0.916	0.918	0.936	0.951	0.841	0.881	0.913
500	0.858	0.889	0.917	0.920	0.938	0.954	0.843	0.883	0.916
1000	0.860	0.889	0.916	0.922	0.938	0.953	0.844	0.883	0.916

Table 1 presents the averages of true reliability, rank correlation and SSR under different simulation conditions over iterations. When Table 1 is examined, it is determined that the averages of the estimated statistics do not differ much by the sample size. All of the reliability statistics obtained indicate a good degree of reliability (Pollit, 2012). It can be concluded that the true reliability, rank order accuracy and SSR statistics improve as we move from the relatively flexible TR40 termination rule to the relatively strict TR30 termination rule. Figure 2, Figure 3 and Figure 4 show the distributions of the consistency statistics obtained in 82 iterations under each simulation condition.

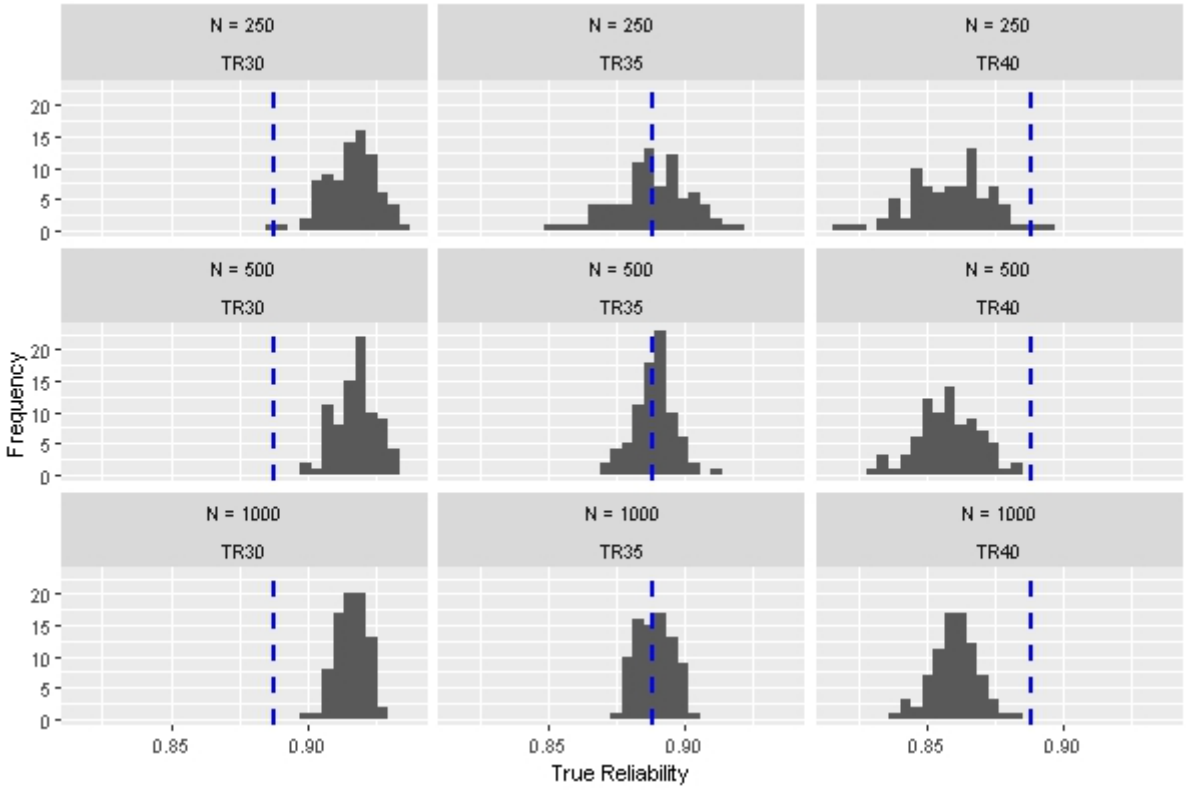

Figure 2. The distribution of true reliability statistics under different simulation conditions

When the true reliability statistics obtained under different simulation conditions, presented in Figure 2, are examined, it is found that the shape of the distribution differs in each condition. Although acceptable or good levels of true reliability statistics were obtained in each condition, it was found that the estimated true reliability statistics spread over a narrower area with increasing sample size. Higher levels of true reliability were also obtained with stricter standard error termination rules.

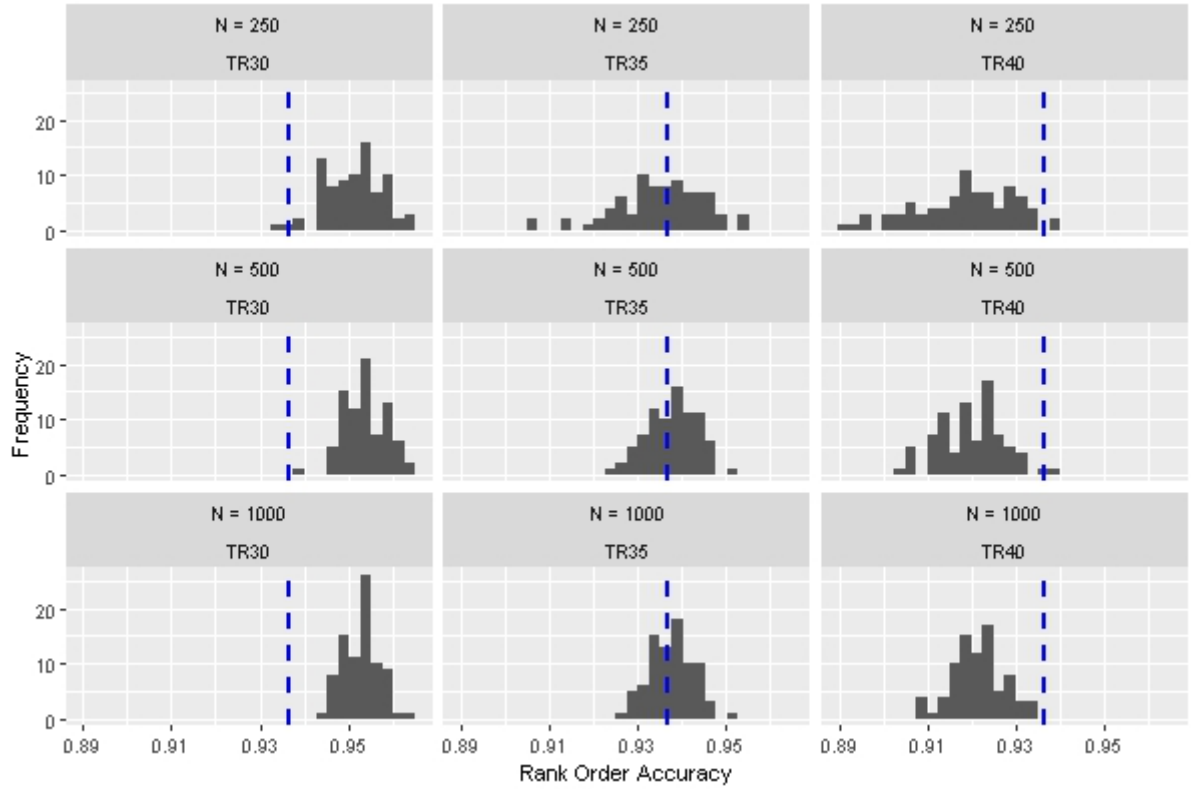

Figure 3. The distribution of rank order accuracy statistics under different simulation conditions

When the rank order accuracy statistics obtained in different simulation conditions presented in Figure 3 are examined, it is found that the shape of the distribution differs in each condition. Although acceptable or good levels of rank order accuracy statistics were obtained in all conditions, it was found that the estimated rank order accuracy statistics spread over a narrower area with increasing sample size. Higher levels of rank order accuracy were also obtained with a stricter standard error termination rule. Comparing the actual reliability and the rank order accuracy, the rank order accuracy was found slightly higher.

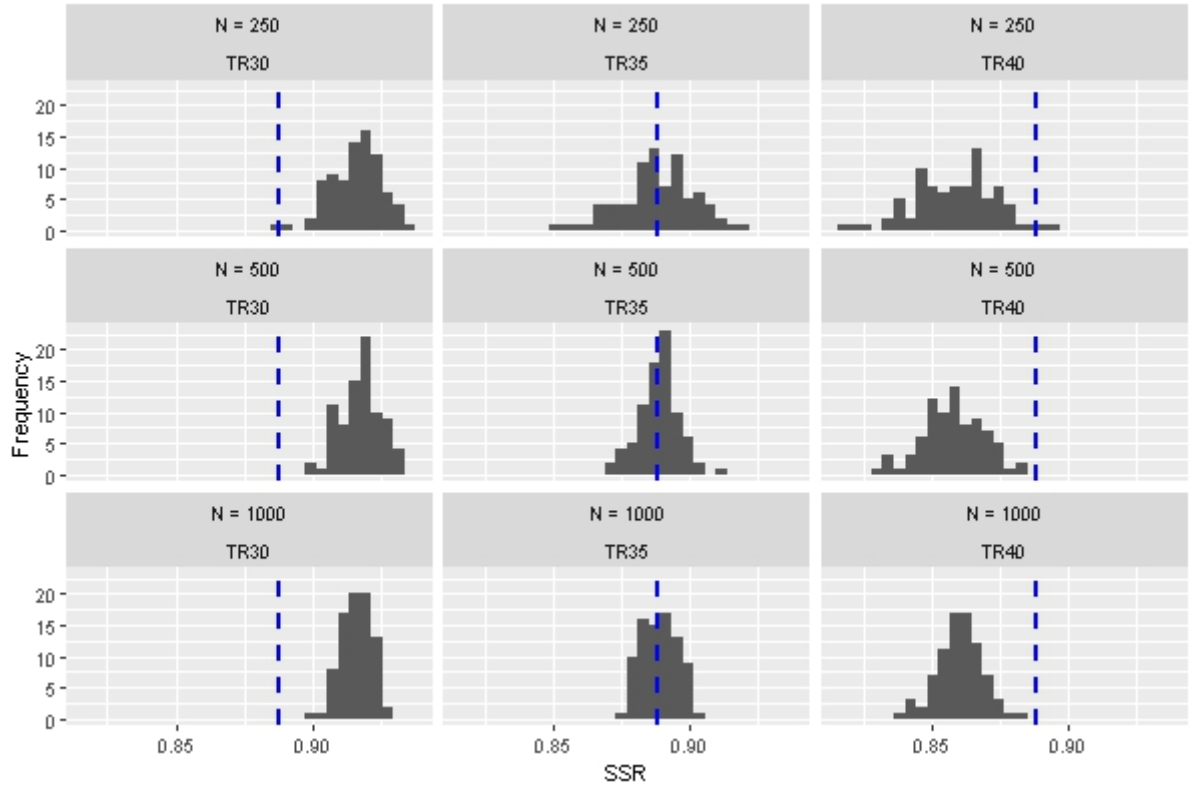

Figure 4. The distribution of SSR statistics under different simulation conditions

When the SSR statistics obtained in different simulation conditions presented in Figure 4 are analyzed, it is found that the shape of the distribution differs in each condition. Although acceptable or good levels of SSR statistics were obtained in each condition, it was found that the estimated SSR statistics spread over a narrower area with increasing sample size. Higher levels of SSR were also obtained with a stricter standard error termination rule. Table 2 presents the findings on the average number of pairwise comparisons required for objects when different levels of standard error values are used as the termination rule.

Table 2. The required average number of pairwise comparisons to meet the related termination rule

	Standard Error Termination Rule		
N	**TR40**	**TR35**	**TR30**
250	30.56	39.16	52.55
500	30.46	38.84	51.68
1000	30.42	38.74	51.38

Table 2 shows the average number of pairwise comparisons each object should be included in to satisfy the relevant standard error termination rule under different simulation conditions. Given the information presented in Table 2, the average number of pairwise comparisons is not affected by the sample size. However, to ensure a stricter standard error termination rule, many more pairwise comparisons are required.

Application Results

In the second part of the study, a total of 754 pairwise comparisons were conducted in order to obtain ACJ ability estimates for 50 writing performances. On average, each pairwise comparison was performed in 2 minutes and 41 seconds, totaling 33 hours, 48 minutes and 28 seconds of scoring.

The ACJ ability estimates range from -2.80 to 2.18 with a mean of 0 and a standard deviation of 1.19. 46 out of 50 writing performances met the standard error termination rule of 0.40, and only 4 performances were excluded from further pairwise comparisons based on the maximum number of comparisons of 40 rule. The standard error estimates for these 4 performances are 0.403, 0.409, 0.417, and 0.472 respectively. All of these performances are at high achievement levels and their ability estimates range from 1.53 to 2.18. The relationship between ACJ ability estimates and numbers of pairwise comparisons is presented in Figure 5.

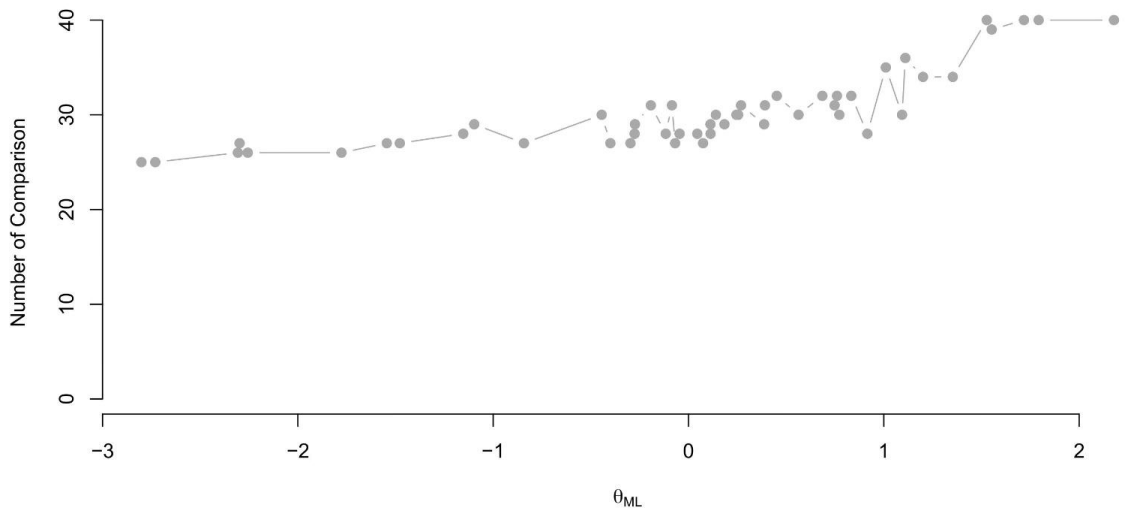

Figure 5. The relationship between ACJ ability estimates and numbers of pairwise comparisons

Given the findings presented in Figure 5, it is possible to say that objects with higher levels of abilities are involved in a higher number of pairwise comparisons. Moreover, the fact that each object is involved in an average of 30.16 pairwise comparisons is in line with the finding presented in Table 2 that each object needs to be involved in an average number of pairwise comparisons in order to meet the standard error termination rule of 0.40 regardless of the sample size.

The SSR statistic obtained for the ACJ ability estimates is 0.89. This finding shows that the reliability is quite high. This result can be considered as an indication that even non-expert raters can rate with high reliability. In addition, correlations of 0.71 and 0.73 were estimated between ACJ ability estimates and standardized analytical total scores and English language proficiency holistic scores, respectively. The relationship between the ACJ ability estimates and the other two scores is visualized in Figure 6 in the Appendix. In this context, it can be concluded that the ability estimates obtained with the ACJ are compatible with the scores that are obtained using the analytic and holistic rubrics.

Discussion, Conclusion and Suggestions

The aim of this study was to compare the reliability of ACJ at different levels of standard error termination rules in different sample sizes. The results of the Monte Carlo simulation show that when 0.30, 0.35 and 0.40 standard error termination rules are applied, lower standard errors are found at the midpoints of ability; moreover, the 0.40 standard error termination rule estimates a wider range of standard errors than the 0.35 and 0.30 standard error termination rules. Sample size makes only small differences in the results obtained in terms of true reliability, rank order accuracy and SSR. However, contrary to the current research, Crompvoets et al. (2020) stated that sample size has an effect on the rank order accuracy and true reliability value. At this point, it is important to note that Crompvoets et al. (2020) determined the sample size as 20, 25, 30 and 100. The current study, on the other hand, includes samples of at least 250 people. However, considering the iterations, it is realized that as the sample size increases, the true reliability, rank order accuracy and SSR values are found in a narrower range.

In general, the rank order accuracy was higher than the true reliability and SSR values. Sample size did not make a difference in terms of the average number of pairwise comparisons, and a larger number of pairwise comparisons were required to satisfy a strict standard error termination rule. In the study, actual reliability, rank reliability, SSR and average number of comparisons differed with respect to the standard error termination rule. When the results are considered in general, the rank order accuracy is approximately 0.92, 0.94 and 0.95 for the standard error termination rule of 0.40, 0.35 and 0.30, respectively. This result shows that it is appropriate to use all the standard error stopping rules considered in the simulation in cases where ranking is at the forefront, and the standard error termination rule can be determined with respect to the acceptable standard error relative to the importance of the decisions to be made. However, it should be kept in mind that when applying the standard error termination rule of 0.40, 0.35 and 0.30, approximately 30, 39 and 52 average pairwise comparisons are required, respectively. When the true reliability values were considered, approximately 0.86, 0.89 and 0.92 values were obtained for 0.40, 0.35 and 0.30 standard error termination rules, respectively. The SSR, which is prominent in comparative judgement scoring research, was found to be 0.84, 0.88 and 0.92 for the standard error termination rule of 0.40, 0.35 and 0.30, respectively. Crompvoets et al. (2020) state that a true reliability and SSR value of 0.80 can be achieved with 20-22 pairwise comparisons. The results obtained are consistent with this evidence. Verhavert et al. (2019) stated that approximately 26-37 pairwise comparisons should be made to reach a reliability of around 0.90. In the current study, reliability values close to 0.90 were reached in an average of approximately 39 comparisons. Pollitt (2012), on the other hand, reached a reliability above 0.90 after 9 rounds of comparison. However, it is believed that the result was caused by the standard deviation value of 3.85 for ability estimates in Pollitt's (2012) study.

The study was conducted under the condition of a maximum standard error of 0.40 for all objects. With this standard error termination value, it was aimed that the amount of error for each object's ability estimate would be low. Thus, the decisions made individually about the objects would be more qualified. As a matter of fact, when the literature is examined, it is seen that the average standard error of ability estimations reaches up to 1.5 (Verhavert et al., 2022). As a result of the current study, it was concluded that the adaptive selection algorithm performed adequately in comparative judgement. Depending on the importance of the decisions to be made, it was determined that one of the standard error termination rules in the current research could be preferred. Moreover, the results of the study show that increasing the sample size does not differentiate the results of the research. Therefore, in real-life administrations, when sufficient raters are reached, comparative judgement can be performed in large samples through the adaptive selection algorithm.

In the second part of the study, after the simulation results were obtained and it was determined that the sample size did not differentiate the results, the application was carried out with a sample of 50 students. In the application carried out under similar conditions to the simulation study, a standard error termination rule of 0.40 and a maximum of 40 pairwise comparisons for each student were taken into consideration. As a result of the application, findings similar to the simulation results were obtained. As a matter of fact, 30 pairwise comparisons were made on average per object in the application and the result is very close to the average number of pairwise comparisons determined for the 0.40 standard error termination rule in the simulation study. While the SSR value is 0.89 in the application, it is 0.84 in the simulation study conducted with the standard error termination rule of 0.40. The fact that the standard deviation of the ability estimates was 1.19 in the application may have led to a higher SSR value in the application.

A striking factor in the results of the application is that the number of pairwise comparisons increases and the standard errors are higher at higher ability levels. This may indicate that raters may have difficulty in deciding which performance is better for objects at a higher level of ability. In addition, the correlations between the holistic scores and ability estimates through ACJ and standardized analytic scores and the abilities estimated through ACJ were found to be above 0.70. It is reported in the literature that the correlations between rubric scores and ACJ ability levels can range between 0.38 and 0.92 (Steedle & Ferrara, 2016). The fact that the ACJ in the application was carried out by four researchers who are not experts in the field of English language and who have C1 level English skills and a correlation above 0.70 was reached is a promising result for the use of ACJ. Bartholomew, Nadelson, Goodridge, and Reeve (2018) conducted CJ with non-teaching individuals in their study and found comparable results to this study. The standard errors and reliability estimates obtained as a result of the application indicate that ACJ can be used in classroom assessment at the K12 level. In addition, using ACJ does not make a difference in students' usual test practices (Pollitt, 2012).

The ACJ allows the scoring process to be realized with a much easier decision. In addition, it is an important advantage that the ability estimates are very similar to item response theory models. Moreover, unlike rubrics, ability estimates are obtained on a continuous scale, which makes it possible to reveal the differences between individuals more precisely in terms of the measured trait. In the application within the scope of the research, a reliability value of approximately 0.90 can be obtained with an average of 30 pairwise comparisons. In order to reduce the number of pairwise comparisons in real-life administrations, relatively lower reliability levels can be targeted. Especially if a reliability level of 0.70 is targeted in classroom applications, it seems possible to achieve this target with much fewer pairwise comparisons.

Considering the average pairwise comparison time obtained for 754 comparisons, the final scoring of a performance requires approximately 40 minutes. It should not be ignored that this time is achieved with raters who are not English Language experts and in a situation where the average word count is high. In addition, given the recent decisions of the Ministry of National Education, it seems likely that open-ended questions will be used in high-stakes exams in the short to medium term. If these exams are scored with holistic rubrics consisting of five or six categories, the problem of discrimination may arise in these exams, which are essentially for ranking purposes; therefore, performances will need to be scored separately in several dimensions with an analytical rubric as in this study. Considering that in rubric-based scoring methods, each performance has to be scored by at least two raters and an experienced rater is involved in case of disagreement, it can be said that rating time with ACJ is quite reasonable.

MoNE and ÖSYM have been working on four-skills language proficiency exams for some time. These exams are intended to measure not only writing skills but also speaking skills. The ACJ can be used to score not only writing skills but also speaking skills, which is another productive skill. The fact that high reliability values can be obtained by focusing only on which performance is better, makes it possible to use the ACJ in different dimensions of such exams that are suitable for holistic assessment. However, it should be kept in mind that this study is the first application research on the use of the ACJ method at the K-12 level in Turkey. For this reason, it can be stated that more research should be conducted on the usability of ACJ at the K-12 level.

Based on the results, a high level of reliability was found in all conditions evaluated in the study. It is generally accepted that when CJ is utilized, raters can be asked to consider only validity when making their decisions. Indeed, the results were found to be highly reliable (Jones & Davies, 2023; Pollitt, 2012).

Based on the results of the current study, researchers may be recommended to develop an alternative selection algorithm to the adaptive selection algorithm in future studies. In addition, a reference set-based comparative judgement study can be carried out, which can also mediate the standardization process. In the reference set-based approach, adaptive selection algorithm, Bayesian adaptive selection algorithm or new selection algorithms to be developed by researchers can be utilized. The research is limited by the standard error termination rule and sample size in terms of simulation factors. In future research, the number of simulation factors can be increased. In this direction, different standard deviation values can be used when generating the ability distribution with a normal distribution. The effects of using the standard error termination rule and the adaptive selection algorithm can be examined in the case of more variation in objects' ability levels. The number of raters was not examined in this research. In future studies, the number of raters can be differentiated, and the effects on the reliability of ACJ can be examined by making changes in the level of agreement among raters.

Acknowledgements

We would like to thank ÖSYM Research and Development Department for supporting the project titled "Development of a System and Software for Scoring Open-Ended Items with Adaptive Comparative Judgement" which was used in the preparation of this study.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 451-462. doi:10.1177/014662167800200319

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H., & Reeve, E. M. (2018). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educational Assessment, 23*(2), 85-101. doi:10.1080/10627197.2018.1444986

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education*, 6, 775203. doi:10.3389/feduc.2021.775203

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209-220. doi:10.1080/02602930801955978

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. doi:10.1093/biomet/39.3-4.324

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement, 6*(2), 202-223.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246-294). Qualifications and Curriculum Authority.

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. doi:10.1080/0969594X.2017.1418734

Christodoulou, D. (2024). *Using comparative judgement to improve writing [webinar]. The Education Hub.* Retrieved from https://theeducationhub.org.nz/using-comparative-judgement-to-improve-writing/#:~:text=Comparative%20judgement%20is%20a%20process,double%20marking%2C%20but%20much%20quicker

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*(1), 127-144. doi:10.1080/0969594X.2012.741059

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics, 45*(3), 316-338. doi:10.3102/1076998619890589

Crompvoets, E. A. V., Beguin, A. A., & Sijtsma, K. (2021). *Pairwise comparison using a Bayesian selection algorithm: Efficient holistic measurement*. PsyArXiv. doi:10.31234/osf.io/32nhp

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*, 788202. doi:10.3389/feduc.2021.788202

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., . . . Boser, U. (2023). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research, 9*(2), 248-269. doi:10.1075/ijlcr.22026.cro

Daniel, F., Microsoft Corporation, Weston, S., & Tenenbaum, D. (2022). doParallel: Foreach parallel adaptor for the 'parallel' package (Version 1.0.17) [Computer software]. Retrieved from https://CRAN.R-project.org/package=doParallel

Goossens, M. ve De Maeyer, S. (2018). How to obtain efficient high reliabilities in assessing texts: Rubrics vs comparative judgement. Technology enhanced assessment. TEA 2017. Communications in Computer and Information Science, Springer, Cham. doi:10.1007/978-3-319-97807-9_2

Gustafsson, J.-E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program*. Göteborg: Göteborg University.

Heldsinger, S. & Humphry, S. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational research, 55*(3), 219-235. doi:10.1080/00131881.2013.825159

Holmes, S. D., Meadows, M., Stockford, I., & He, Q. (2018). Investigating the comparability of examination difficulty using comparative judgement and Rasch modelling. *International Journal of Testing, 18*(4), 366-391. doi:10.1080/15305058.2018.1486316

Humphry, S. M., & Heldsinger, S. (2019). A two-stage method for classroom assessments of essay writing. *Journal of Educational Measurement, 56*(3), 505-520. doi:10.1111/jedm.12223

Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education, 47*(2), 170-181. doi:10.1080/1743727X.2023.2242273

Laming, D. (2003). *Human judgment: The eye of the beholder.* Thomson Learning.

Lesterhuis, M., Bouwer, R., Van Daal, T., Donche, V., & De Maeyer, S. (2022). Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education, 7*, 823895. doi:10.3389/feduc.2022.823895

Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis*. New York: John Wiley & Sons.

MoNE Measurement and Evaluation Regulation. (2023). *Resmi Gazete* (Sayı: 32304). Retrieved from https://www.resmigazete.gov.tr/eskiler/2023/09/20230909-2.htm

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300. doi:10.1080/0969594X.2012.665354

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. https://www.r-project.org/ adresinden erişildi.

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric rating with MFRM versus randomly distributed comparative judgment: A comparison of two approaches to second-language writing assessment. *Educational Measurement: Issues and Practice, 39*(4), 30-40. doi:10.1111/emip.12329

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay. *Applied Measurement in Education, 29*(3), 211-223. doi:10.1080/08957347.2016.1171769

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286. doi:10.1037/h0070288

Thwaites, P., Kollias, C., & Paquot, M. (2024). Is CJ a valid, reliable form of L2 writing assessment when texts are long, homogeneous in proficiency, and feature heterogeneous prompts?, *Assessing Writing*, 60, doi:10.1016/j.asw.2024.100843

Using anchors to link judging sessions. (2016). Retrieved from https://nomoremarkingltd.freshdesk.com/support/solutions/articles/16000029952-using-anchors-to-link-judging-sessions.

Uysal, İ., Gürel, S., Şahin, M. D., İbileme, A. İ., & Yıldırım Görgülü, Y. (2024). *Açık uçlu maddelerin karşılaştırmalı yargıyla puanlanmasında sabit sayıda uyarlamalı ve rassal eşlemeye dayalı bir simülasyon çalışması.* 9th. International Conference on Measurement and Evaluation in Education, Anadolu University, Eskişehir.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice, 26*(1), 59-74. doi:10.1080/0969594X.2016.1253542

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. doi:10.1080/0969594X.2019.1602027

Verhavert, S., Furlong, A., & Bouwer, R. (2022). The accuracy and efficiency of a reference based adaptive selection algorithm for comparative judgment. *Frontiers in Education*, 6, 785919. doi:10.3389/feduc.2021.785919

Appendix 1

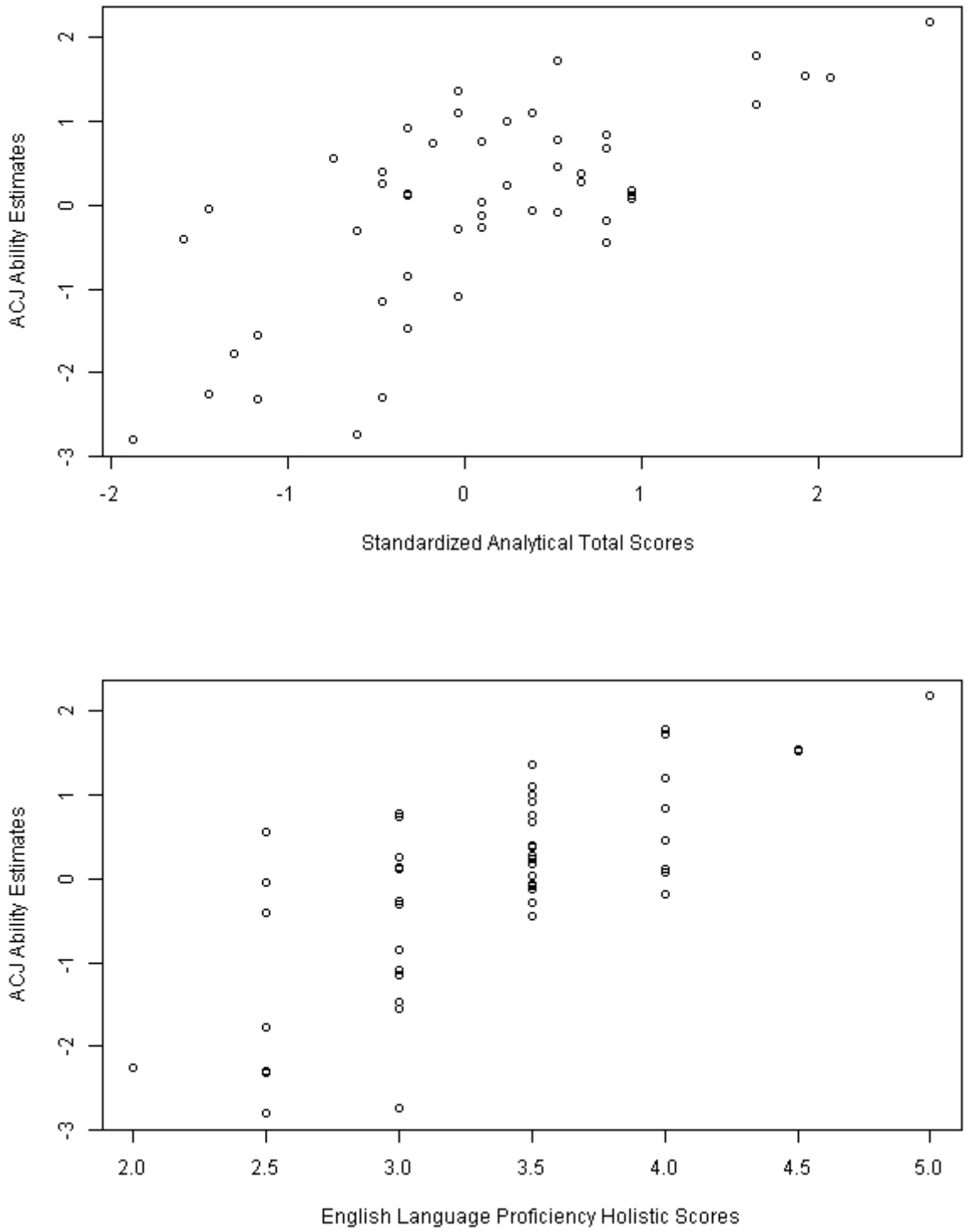

Figure 6. The relationship between ACJ ability estimates and two rubric scores