Sureeyatanapas *et al.*
Language Testing in Asia (2024) 14:1
https://doi.org/10.1186/s40468-023-00271-z

Language Testing in Asia

RESEARCH Open Access

The analysis of marking reliability through the approach of gauge repeatability and reproducibility (GR&R) study: a case of English-speaking test

Pornphan Sureeyatanapas[1*], Panitas Sureeyatanapas[2], Uthumporn Panitanarak[3], Jittima Kraisriwattana[1], Patchanan Sarootyanapat[1] and Daniel O'Connell[1]

*Correspondence:
pornpsu@kku.ac.th

[1] General Education Division, International College, Khon Kaen University, Khon Kaen, Thailand
[2] Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand
[3] Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand

Abstract

Ensuring consistent and reliable scoring is paramount in education, especially in performance-based assessments. This study delves into the critical issue of marking consistency, focusing on speaking proficiency tests in English language learning, which often face greater reliability challenges. While existing literature has explored various methods for assessing marking reliability, this study is the first of its kind to introduce an alternative statistical tool, namely the gauge repeatability and reproducibility (GR&R) approach, to the educational context. The study encompasses both intra- and inter-rater reliabilities, with additional validation using the intraclass correlation coefficient (ICC). Using a case study approach involving three examiners evaluating 30 recordings of a speaking proficiency test, the GR&R method demonstrates its effectiveness in detecting reliability issues over the ICC approach. Furthermore, this research identifies key factors influencing scoring inconsistencies, including group performance estimation, work presentation order, rubric complexity and clarity, the student's chosen topic, accent familiarity, and recording quality. Importantly, it not only pinpoints these root causes but also suggests practical solutions, thereby enhancing the precision of the measurement system. The GR&R method can offer significant contributions to stakeholders in language proficiency assessment, including educational institutions, test developers and policymakers. It is also applicable to other cases of performance-based assessments. By addressing reliability issues, this study provides insights to enhance the fairness and accuracy of subjective judgements, ultimately benefiting overall performance comparisons and decision making.

Keywords: GR&R, ICC, Inter-rater reliability, Intra-rater reliability, Scoring consistency, Speaking proficiency, Language assessment, Marking reliability, Measurement system analysis, Rubric

Introduction

The trustworthiness of testing scores, or marking reliability, is a topic that has been emphasised and widely discussed in educational literature (Bird & Yucel, 2013). Reliability in this context indicates the extent to which assessments yield stable and consistent results (Golafshani, 2003; Lyness et al. 2021). Maintaining the legitimacy and fairness of assessments and tests in educational systems requires ensuring the reliability of marking. This promotes consistent and valid results, allowing for more accurate interpretations of test scores and fair comparisons between individuals' proficiency (Doosti & Safa, 2021; Khan et al. 2020). The reliability of marking is particularly imperative in cases where students' test scores influence their future employment and education (Marshall et al. 2020).

In the study of English language education, the choices between standardised assessment and performance-based assessment present distinct advantages and disadvantages. Standardised assessments, such as multiple-choice questions, true or false, or matching exercises, often offer a higher degree of marking reliability (Doosti & Safa, 2021; Trevisan, 1991). However, these standardised forms might not be effective or comprehensive enough to accurately reflect students' actual proficiency and progress. Moreover, they fail to simulate real-life communication scenarios. On the other hand, performance-based assessments, such as oral tests, presentations, group performances, and written productions, effectively address these limitations by closely simulating practical language use, and they offer a comprehensive view of students' language abilities (Bland & Gareis, 2018; Brown, 2004; Doosti & Safa, 2021). Nonetheless, the reliability of performance-based testing results may be compromised due to the subjective nature of scoring, influenced by various factors (Doosti & Safa, 2021; Khan et al. 2020; Marshall et al. 2020).

The main source of inconsistent scores is attributed to the raters or examiners (Lyness et al. 2021). Reliability issues often stem from the involvement of multiple examiners within a teaching team, a concept referred to as 'inter-rater reliability' (Bird & Yucel, 2013; Khan et al. 2020; Wang, 2009). The reliability issue becomes more noticeable when all examiners assess the same test. In many universities, especially in non-English-speaking countries, the same English course is provided to students across various disciplines during a single academic year. Consequently, various lecturers participate in grading the same exercises, leading to a comparison of students' language proficiencies based on scores assigned by them (Akeju, 1972). Existing literature reports that low inter-rater reliability can arise due to variations in teaching experience and levels of training among examiners, as well as a lack of effective scoring criteria or standardised marking guidelines (Bird & Yucel, 2013; Doosti & Safa, 2021; Huang et al. 2018). Another aspect of reliability frequently discussed is 'intra-rater reliability,' which pertains to the consistency of scores assigned by a single examiner at different times (Bird & Yucel, 2013; Khan et al. 2020). For intra-rater reliability, factors such as the sequence in which work is evaluated and the time allocated for marking could exert influence (Bird & Yucel, 2013). Additionally, examiners' preconceived biases, lapses in attention, and human errors can also contribute to scoring inconsistencies (Doosti & Safa, 2021).

Drawing from a review of literature on English language learning, reliability concerns have been a substantial topic of discussion, especially in the context of performance-based assessments, notably in the domains of writing and speaking proficiency (Akeju, 1972; Khan et al. 2020; Porter & Jelinek, 2011; Rashid & Mahmood, 2020; Saeed et al. 2019). However, this study narrows its focus to specifically examine the speaking proficiency test, as it appears to face a greater risk of poor reliability compared to the writing test. The primary objective of speaking tests is generally to assess students' ability to use appropriate language in social contexts (Khan et al. 2020). When marking an oral assessment, numerous factors can cause marking bias. These factors include an examiner's familiarity with a student's accent and pronunciation (Carey et al. 2011; Huang et al. 2016), the examiner's linguistic background (Huang, 2013; Winke et al. 2013), and the examiner's understanding of rubrics and scoring criteria (Jeong, 2015; Khan et al. 2020).

Based on a review of the educational literature, most studies have examined marking reliability using various methods, such as analysis of variance, Pearson's correlation, or Kappa coefficients. However, these methods still exhibit several limitations, which will be discussed in the subsequent section. Therefore, this study proposes the application of an alternative method called the study of 'gauge repeatability and reproducibility,' or 'GR&R,' for analysing marking reliability. GR&R is extensively employed to assess the precision of gauges or measurement instruments in the manufacturing and engineering fields. However, the literature review reveals a lack of utilisation of the GR&R study in the realm of education. This might be due to the fact that this technique was originally developed to address imprecision issues of measurement data in the automotive industry (AIAG, 2010). Consequently, it has primarily been integrated into higher education curricula within engineering schools, while it remains absent in education, humanities, and social sciences faculties. Therefore, many educators and language institutions may be unfamiliar with the GR&R study or fail to recognise its direct relevance to educational assessment practices. Another possible reason is the complex computation procedures associated with the GR&R approach, which may not align with the practical perspective of teachers and educators (Başaran et al. 2015). Nevertheless, there are now software tools available that can simplify the analysis of the GR&R study. This study, therefore, aims to promote the adoption of this technique in a broader range of users and contexts.

The objective of this study is to employ the GR&R approach to address concerns regarding reliability issues in language proficiency tests, encompassing both intra- and inter-rater reliabilities. Additionally, it aims to analyse which characteristics of speech are more prone to experiencing low marking reliability or potentially lead to discrepancies between examiners. To validate the findings, the results of the GR&R study will be cross-compared with the intraclass correlation coefficient (ICC), which is a commonly used method for assessing marking reliability in educational literature (Doosti & Safa, 2021; Hallgren, 2012; Saeed et al. 2019; Soemantri et al. 2022).

This study highlights the crucial role of the GR&R method in language proficiency assessment. It is expected to enhance fairness in subjective scoring and strengthen the precision of the assessment system. In comparison to the ICC method, this study seeks to determine whether the GR&R demonstrates superior sensitivity in detecting marking

reliability issues and establishes itself as a more effective evaluation tool in education fields. Additionally, by identifying various causes of reliability challenges in marking processes, the proposed solutions are expected to empower educators and students to achieve more precise and equitable language assessment.

The paper consists of several key sections. After the introduction, there is a review of commonly used methods for assessing marking reliability in education. The subsequent section introduces the concepts of the GR&R study, followed by the 'Methods' section that demonstrates its application in a real-life scenario involving three examiners assessing 30 recording clips. The next section presents the results of this illustrative case, leading to a discussion on reliability issues and practical implications. The conclusions are drawn in the final section.

A review of literature

This section offers a comprehensive review of previous studies that addressed issues related to marking reliability in education. The primary objective is to explore commonly used methods for assessing the level of reliability. The first part of this section presents studies that focused on assessing students' writing skills as a part of English language learning, followed by those examining scoring reliability for speaking proficiency tests. The section also provides a review of methods used in assessing marking reliability in other fields of education. A summary of these methods is presented in Table 1, followed by a critique of their limitations or conditions of use. The final part of this section justifies the adoption of the GR&R approach in this study over the commonly used methods.

First of all, the review explores methods that previous researchers in English language teaching have employed to assess the reliability of scoring or grading students' writing tasks. For instance, Akeju (1972) conducted a study to evaluate inter-rater reliability in marking English essay papers written by 96 teenage students in Ghana. Seven examiners graded these papers, and the reliability assessment involved calculations of Pearson's correlation coefficients, analysis of variance (ANOVA), and Hartley's test for the equality of variances. Wang (2009) analysed the inter-rater reliability of eight examiners in scoring compositions. ANOVA was employed to test for any significant differences among the scores provided by the examiners. Nimehchisalem et al. (2021) introduced a genre-specific scale for evaluating argumentative essays written by students. They utilised a dataset of 110 samples from Malaysian university students for their study. The experiments involved the participation of five experienced raters. Inter- and intra-rater reliability were evaluated by computing Pearson's correlation coefficients. Li (2022) investigated the perceived effectiveness of a teacher-developed scoring rubric by peer raters in a Chinese EFL writing course at the college level. The study also aimed to determine how well the rubric could differentiate students' writing skills. Intra-rater reliability was assessed using the Infit Mean Square (Infit MnSq) and Outfit Mean Square (Outfit MnSq), while inter-rater reliability was evaluated through the analysis of the point-measure correlation.

When focusing on the English-speaking test, it becomes apparent that the methods for assessing marking reliability are similar to those reviewed previously for writing assessment. This similarity may arise because both skills depend on the subjective judgement of the raters, and most language institutions employ performance-based

Table 1 Methods for evaluating the marking reliability in education

Authors	Pearson's correlation coefficient	Fleiss' kappa or Cohen's kappa	ICC	ANOVA	Percentage of agreement	Standard deviation	Spearman's rank correlation coefficient	Bland–Altman plot	SSR	Split-halves technique	Paired sample *t*-test	Infit MnSq and Outfit MnSq
Akeju (1972)	/			/								
Sullivan and Hall (1997)	/				/							
Wang (2009)				/								
Hallgren (2012)		/	/									
Mukundan and Nimehchisalem (2012)	/											
Bird and Yucel (2013)						/						
Zhao (2013)	/											
Davis (2016)	/	/		/	/							
Saeed et al. (2019)			/									
Aprianoto and Haerazi (2019)		/										
Khan et al. (2020)	/							/				
Marshall et al. (2020)	/								/	/		
Rashid and Mahmood (2020)							/					
Doosti and Safa (2021)			/									
Lyness et al. (2021)		/			/							
Nimehchisalem et al. (2021)	/											
Soemantri et al. (2022)			/									
Li (2022)	/											/
Detey et al. (2023)							/					
Stuart and Barnett (2023)	/	/										
Naqvi et al. (2023)						/					/	
Total	10	5	4	3	3	2	2	1	1	1	1	1

assessment forms rather than standardised ones. Zhao (2013) conducted a study to investigate the validity and reliability of the Diagnostic College English Speaking Test (DCEST), a face-to-face interview test utilised for evaluating students' proficiency in various aspects of English speaking, within the context of EFL learning in China. Again, the study employed Pearson's correlation analysis to assess both inter- and intra-rater reliability. Davis (2016) investigated the impact of several factors (rater training, experience in scoring, and the use of exemplar responses and scoring rubrics) on the consistency and accuracy of raters' judgements. The study involved twenty experienced English teachers as participants and utilised the TOEFL iBT speaking test for experiments. Various analysis methods, including ANOVA, Pearson's correlation, and Fleiss' Kappa coefficients, were employed in this research. Saeed et al. (2019) developed an English-speaking proficiency test along with assessment rubrics. The test evaluated competency not only in grammar but also in the appropriate use of language within specific social contexts. The participants were undergraduate students from a university in Malaysia. To examine the effectiveness of the rubrics, the inter-rater reliability between two examiners was analysed using the ICC as an indicator. Aprianoto and Haerazi (2019) presented the development processes of an intercultural-based English speaking model utilised for teaching speaking skills at the higher education level. The inter-rater reliability of the instrument was assessed using the Kappa coefficient. Khan et al. (2020) evaluated the inter-rater reliability of the speaking test by collecting data from 61 university students in Saudi Arabia. Pearson's correlation coefficient and Bland–Altman plot were used to examine the agreement among six raters. Doosti and Safa (2021) investigated the impact of rater training on improving inter-rater reliability in an English-speaking test. In their study, four raters scored the performance of 31 Iranian EFL learners on the IELTS speaking test in two rounds (before and after training). The ICC was used to evaluate the inter-rater reliability.

In other educational domains apart from English language teaching, similar methods can also be found for determining the marking reliability. A number of studies were conducted in native English-speaking countries such as New Zealand, Australia, and the UK. Sullivan and Hall (1997) examined the effectiveness of students' self-assessment by inviting undergraduate students in New Zealand to grade and mark their own writing paper—a review of literature. They then compared the self-assessment results with the lecturer's grading of the same work. The level of agreement between the students and the lecturer was assessed using the percentages of 'hits' and 'misses' between their grading results and Pearson's correlation coefficients between their marking scores. Bird and Yucel (2013) proposed a programme aimed at enhancing the marking reliability of students' writing tasks—laboratory reports. To assess the inter-rater reliability, eight laboratory demonstrators from various universities in Australia were tasked with grading forty students' scientific reports. The evaluation of inter-rater reliability was based on the analysis of standard deviation. Marshall et al. (2020) investigated whether using comparative judgement in assessing a writing task could lead to more reliable outcomes when multiple examiners are involved. The writing tasks were collected from high school students in New Zealand. To test the reliability, the researchers employed the scale separation reliability (SSR) and split-halves

techniques. Stuart and Barnett (2023) proposed a new tool for assessing the quality of writing tasks in higher education in the UK, called the Writing Quality Scale (WQS). To demonstrate the applicability of this tool, scripts from 120 students were evaluated by two trained examiners. The inter-rater and test-retest reliabilities were analysed using Cohen's kappa and Pearson's correlation.

Hallgren (2012) provided an overview of methodological issues frequently encountered when assessing inter-rater reliability. At the end, Kappa statistics and the ICC were thoroughly described as commonly used measures of the reliability. Mukundan and Nimehchisalem (2012) assessed their newly developed evaluation checklist for English language teaching textbooks. They engaged two English teaching experts to apply the checklist to evaluate the same textbook, and the inter-rater reliability was examined through Pearson's correlation analysis. Additionally, the proposed checklist was tested for its correlation with the results obtained using Skierso's checklist, a widely utilised tool for textbook evaluation. Lyness et al. (2021) examined the inter-rater reliability of trained examiners who assessed the scores of 19 candidates on the teacher performance assessment (TPA). The study utilised both Cohen's kappa coefficient and the percentage of agreement for analysis. Detey et al. (2023) evaluated phonetic fluency in a reading task among Japanese learners of French. Four examiners assessed the reading proficiency of twelve Japanese learners. The inter-rater agreement was analysed using Spearman's rank correlations for the given ratings. Naqvi et al. (2023) developed an online placement test for use in higher education in Oman. The test encompasses all four main English proficiency skills: reading, writing, listening, and speaking. The tests were completed by a sample of students and then marked by two examiners. The test's reliability was evaluated using the standard deviation of the students' scores and the paired sample *t*-test.

Potential factors that influence the reliability of marking scores were revealed by several studies. For instance, Rashid and Mahmood (2020) examined the factors affecting the inter-rater reliability of marking exam papers in high-stake testing for secondary school students in Pakistan. They employed a questionnaire-based survey to collect data from 98 raters. The analysis of reliability was done using the Spearman correlation coefficient. The study found that the training of examiners significantly influenced their reliability in marking the exams. This finding is consistent with that of Doosti and Safa (2021), mentioned earlier, which also found that the training effectively enhanced the reliability of the scores given by the examiners. Soemantri et al. (2022) analysed the inter-rater reliability to evaluate the effectiveness of two different rubrics used for assessing reflective writing in an undergraduate medical course at a university in Indonesia. The assessment was conducted by two examiners, and the ICC was used as a measure of inter-rater reliability. The results indicated that a more detailed rubric led to a higher reliability score compared to a less detailed one.

Table 1 summarises the methods used to evaluate the marking reliability in the reviewed studies. SPSS software was used to facilitate the analysis in most of these studies. From the summary, the top five common methods for evaluation are found to be Pearson's correlation, Kappa, ICC coefficients, ANOVA, and the percentage of agreement. However, each method suits specific testing conditions and has some limitations. For instance, Pearson's correlation coefficient is only a measure of the linear relationship between continuous variables and does not indicate whether those variables are

equivalent or consistent (Akeju, 1972). ANOVA only reveals if there is any significant difference between the means of different data groups, regardless of whether each individual test is rated equally by different examiners or in different marking rounds. Additionally, the types of measurement scales and test assumptions play a significant role in ANOVA. The dependent variable must be at a continuous level of measurement (interval or ratio scales), while the independent variables must be categorical (nominal or ordinal scales). ANOVA is a parametric test with certain assumptions that justify the results. It assumes that the data are normally distributed, and the variances among data groups are approximately equal (Statistics Solutions, 2013). However, these assumptions may not always be practical in all cases. Next, the percentage of agreement seems to be the simplest indicator used to judge marking agreement. However, its key issue is that the level of agreement could be overestimated due to the ignorance of chance agreement (Hallgren, 2012; Lyness et al. 2021). Kappa coefficients are among the most widely used measures of inter-rater reliability. They remove the chance agreement by estimating the extent to which the raters could agree by chance (DeVellis, 2005; Stokes, 2011). Cohen's Kappa indicates the agreement between two raters, while Fleiss' Kappa is used when more than two raters are involved. Nevertheless, the Kappa coefficients are only suitable when the assessment scores are on a categorical scale. One of the concerns is that Kappa considers only exact agreement while treating near agreement similarly to extreme disagreement. Although this is typically sensible when dealing with nominal categories, a 'near miss' is preferable to a 'far miss' for some types of data (e.g., ordinal, interval, and ratio scales), such as data from marking students' English proficiencies. Furthermore, the more categories there are for a given dataset, the smaller the Kappa is likely to be. This concern is similar to the simple percentage of agreement in terms of the fact that reducing the number of categories (by combining small groups into a single category) can boost the 'hit rate' (DeVellis, 2005). The ICC is another index commonly used to evaluate the reliability of marking exams or scoring educational tests. This is due to its flexibility, as it is capable of measuring both inter- and intra-rater reliability. Additionally, its formulation can be applied to both continuous and discrete variables (Mehta et al. 2018).

The literature review clearly demonstrates that there is no evidence of utilising the GR&R study to evaluate reliability issues within educational contexts. While the methods commonly used to assess the reliability of exam marking have faced criticism, as described earlier, introducing the GR&R method from manufacturing to the realm of education presents a challenging yet worthwhile endeavour. The idea is that, while GR&R is traditionally employed to assess the consistency of measurements in manufacturing, it can similarly assess the consistency of marking and grading in education. It is a statistical method designed to ensure the consistency and stability of any measuring system, which also includes the assessment of students' language proficiency. Through a GR&R study, one can determine the variability of data within the overall assessment system and classify this variability into different sources, a feature not offered by other methods (Low et al. 2009). When sources of marking errors can be categorised, such as inter- or intra-rater inconsistency or issues within the speaking recording clips themselves, this provides valuable guidelines for further improvement, ultimately leading to a higher quality of assessment (Sennaroglu & Yurtsever,

2018). For this study, the ICC method is used to validate the findings determined by the proposed GR&R method. The next section describes the basic concept of the GR&R study.

Basic concept of gauge repeatability and reproducibility (GR&R) study

GR&R is a statistical method within the framework of measurement system analysis (MSA). The GR&R study focuses on analysing the consistency and stability of a measurement system, which hold significant importance in quality and process improvement within the manufacturing domain (Low et al. 2009; Montgomery, 2013). In the context of this study, the terms 'measurement system' or 'gauge,' as used in the original GR&R methodology, can encompass all examiners responsible for evaluating students' speaking proficiency.

The basic concept of MSA is illustrated in Equation (1), where y represents the score assigned by the examiner, x signifies the actual score of each recording clip or the score a specific student should ideally receive, which is typically unknown in practice, and ε represents the marking error. When assessing multiple clips, the variance of the overall observed scores, denoted as σ^2_{Total}, can be modelled through Equation (2), with σ^2_C representing the variance of the true scores of all clips and σ^2_{Gauge} denoting the variance of the marking errors (Montgomery, 2013).

$$y = x + \varepsilon \tag{1}$$

$$\sigma^2_{Total} = \sigma^2_C + \sigma^2_{Gauge} \tag{2}$$

The GR&R study is capable of isolating the components of the total observed variability and determining the extent of the variability that is actually caused by the examiners (Montgomery, 2013). According to this concept, the precision capability of the examiners relies on two components of the marking error: repeatability and reproducibility. Repeatability indicates whether each examiner assigns the same score when marking the same clip several times under identical conditions, while reproducibility focuses on the differences in scores given by different examiners when evaluating a single clip. Based on this concept, σ^2_{Gauge} can be decomposed into variances originating from the repeatability and reproducibility of the examiners, represented as $\sigma^2_{Repeatability}$ and $\sigma^2_{Reproducibility}$, as described in Equation (3) (Montgomery, 2013).

$$\sigma^2_{Gauge} = \sigma^2_{Repeatability} + \sigma^2_{Reproducibility} \tag{3}$$

The 'average and range' and the ANOVA methods are the two primary approaches generally used to calculate GR&R results in most studies (Sennaroglu & Yurtsever, 2018). The primary theoretical advantage of using the ANOVA method over the average and range method is its ability to measure variance influenced by the interaction between the examiners and the recording clips. This interaction cannot be identified by the average and range (AIAG, 2010). With the average and range method, only variabilities arising from different clips, examiner repeatability, and their reproducibility can be estimated.

However, these factors are generally considered sufficient for assessing marking reliability in the context of education. The average and range method is chosen for this study since its calculation can be done manually or quickly performed via a computer programme such as Minitab, while the calculation procedure of the ANOVA method is much more complicated and only recommended when a computer-aided calculation is available (AIAG, 2010). Given that such a requirement may not be practical or feasible in most educational institutions, the average and range method proves to be a more accessible and suitable choice.

The important indices to indicate gauge capability include the percentage of precision-to-tolerance (%*P/T*), the percentage of precision-to-total variation (%*P/TV*), and the number of distinct categories (*ndc*). However, %*P/T* is not considered in this study since it represents the percentage of gauge variation to product tolerance, which always refers to a range of product specification limits. While it is a critical indicator in manufacturing processes, it may not be applicable to marking educational exams or language proficiency tests. The percentage of *P/TV*, also known as %GR&R in some academic sources, represents the proportion of gauge variability due to repeatability and reproducibility issues to the total observed variability. It indicates the magnitude of the measurement error compared to the total variation (Cepova et al. 2018). According to the Automotive Industry Action Group (AIAG), if %*P/TV* is less than 10%, it indicates a small gauge variability, and the group of examiners is considered to be acceptable. If the percentage is greater than 30%, this group of examiners is deemed unacceptable, and corrective action should be taken. If it falls within the range of 10–30%, its acceptability depends on the importance and seriousness of the application (Cepova et al. 2018; Pan, 2006). The *ndc* indicates how many distinct categories of all recording clips the marking process can distinguish (Cepova et al. 2018). According to a guideline provided by the Minitab software (in the 'Help' function), the assessment system is considered insufficient if it cannot discriminate between at least five different levels of all clips ($ndc < 5$). Also, according to Cepova et al. (2018), the group of examiners cannot provide reliable information about the difference of students' speaking proficiency when *ndc* is less than five. The experiment design and calculation procedures for the two indices, %*P/TV* and *ndc*, are described below, following AIAG (2010) and Ploypanichcharoen (2010).

The calculation procedures are explained here using a case involving three examiners who marked 30 students based on their speaking test recordings. Each examiner marked each recording twice, with the second round of marking taking place 2 weeks after the first. AIAG (2010) and Ploypanichcharoen (2010) recommend that the GR&R study should include at least ten sample parts (recording clips in this study), and the marking scores should vary, covering the entire range of feasible data. Furthermore, there should be a minimum of two examiners, and each examiner should assess each clip 2–3 times without having access to each other's results. The markings should be conducted in a random order of clips. The data sheet is presented in Table 2.

Step 1: Estimate the examiner repeatability, denoted by $\sigma_{Repeatability}$, using Equation (4).

$$\sigma_{Repeatability} = \frac{\bar{\bar{R}}}{d_2} \tag{4}$$

Table 2 The data sheet for a GR&R study

Recording clips	Examiner 1			Examiner 2			Examiner 3			Clip Average
	Trial 1	Trial 2	*R*	Trial 1	Trial 2	*R*	Trial 1	Trial 2	*R*	
1										$\overline{X}_{C1}$
2										$\overline{X}_{C2}$
3										$\overline{X}_{C3}$
…										…
…										…
…										…
28										$\overline{X}_{C28}$
29										$\overline{X}_{C29}$
30										$\overline{X}_{C30}$
Column average	$\overline{X}_1$		$\overline{R}_1$	$\overline{X}_2$		$\overline{R}_2$	$\overline{X}_3$		$\overline{R}_3$	

Note that $\overline{\overline{R}}$ represents the average of all ranges (R) when considering repeated markings under identical conditions. In this specific case, there are 90 Rs resulting from repeatability, which can be found in the data presented in the fourth, seventh and tenth columns of Table 2. Therefore, $\overline{\overline{R}} = \left(\overline{R}_1 + \overline{R}_2 + \overline{R}_3\right)/3$. The value of d_2 depends on the number of replications (m), $d_2 = 1.128$ when $m = 2$. The appropriate values of d_2 for other cases where $m > 2$ can be found in AIAG (2010), Wheeler (2006), or Ploypanichcharoen (2010).

Step 2: Estimate the reproducibility of the group of examiners, referred to as $\sigma_{Reproducibility}$, by using Equation (5).

$$\sigma_{Reproducibility} = \sqrt{\left(\frac{R_{\overline{X}}}{d_2^*}\right)^2 - \left(\frac{{\sigma_{Repeatability}}^2}{n \times m}\right)} \tag{5}$$

$R_{\overline{X}}$ represents the examiner range or the range between average scores given by each examiner. According to Table 2, $R_{\overline{X}}$ is calculated as the difference between the maximum and minimum scores among $\overline{X}_1, \overline{X}_2$, and $\overline{X}_3$, or $R_{\overline{X}} = \max\left\{\overline{X}_1, \overline{X}_2, \overline{X}_3\right\} - \min\left\{\overline{X}_1, \overline{X}_2, \overline{X}_3\right\}$. For this step, the value of d_2^* depends on the number of examiners involved in the study. Note that when the number of subgroups (k) is less than 20 (there is only one subgroup in this case), d_2^* is used instead of d_2. When $k = 1$, d_2^* is equal to 1.91155 for three examiners. For other numbers of examiners, please refer to AIAG (2010), Wheeler (2006), or Ploypanichcharoen (2010) for the corresponding values of d_2^*. Lastly, n represents the number of sample clips.

Step 3: Calculate the marking system variation, also known as Gauge R&R (GR&R) or σ_{Gauge}, using Equation (6). This equation is adapted from the concept introduced in Equation (3).

$$GR\&R = \sigma_{Gauge} = \sqrt{{\sigma_{Repeatability}}^2 + {\sigma_{Reproducibility}}^2} \tag{6}$$

Step 4: Estimate the clip-to-clip variation, or σ_C, by quantifying the variability of the 'true' scores for all recording clips, free from the impact of marking errors. This can be achieved using Equation (7).

$$\sigma_C = \frac{R_C}{d_2^*} \tag{7}$$

R_C represents the range encompassing average scores across all sample clips. Given that the true score of each individual clip remains elusive, the average score of all markings for a particular clip is assumed to be its true score. The average score of clip i is denoted as $\overline{X}_{Ci}$, where i ranges from 1 to 30 in this context (referring to the final column of Table 2). Therefore, $R_C = \max\left\{\overline{X}_{C1}, \overline{X}_{C2}, \ldots, \overline{X}_{C30}\right\} - \min\left\{\overline{X}_{C1}, \overline{X}_{C2}, \ldots, \overline{X}_{C30}\right\}$. In this step, the value of d_2^* depends on the number of clips. When $k = 1$, $d_2^* = 4.147$ for a set of 30 sample clips. For alternative cases, the specific values of d_2^* are provided by AIAG (2010), Wheeler (2006), and Ploypanichcharoen (2010).

Step 5: Compute the total variation (σ_{Total}) using Equation (8), which is adapted from the concept illustrated in Equation (2).

$$\sigma_{Total} = \sqrt{{\sigma_{Gauge}}^2 + {\sigma_C}^2} \tag{8}$$

Step 6: Calculate $\%P/TV$ and ndc by applying Equations (9) and (10), respectively.

$$\%P/TV = \left(\frac{\sigma_{Gauge}}{\sigma_{Total}}\right) \times 100 \tag{9}$$

$$ndc = \frac{1.414 \times \sigma_C}{\sigma_{Gauge}} \tag{10}$$

The calculation can be quickly performed using Minitab software. Additionally, the software offers various types of graphical tools, including the average $(\overline{X})$ chart, range (R) chart, individual plots by clip, and box-plots by examiner. These tools are valuable for conducting in-depth analyses to comprehend the underlying causes of inconsistencies. The following section describes the data collection process for this experiment.

Methods—experimental process and data collection

As mentioned in the introduction, the primary objective of this study is to demonstrate the application of the GR&R study in assessing marking reliability within the context of English-speaking assessments. The data used for this analysis, employing the GR&R and ICC methods, were drawn from a sample group of 30 students. In this research, the data were not primarily collected; instead, they were obtained from audio clips submitted by students in the previous academic year (2022). These audio recordings were part of the evaluation process for a course taught by one of the researchers, who already had access to this data. In this assessment, students were tasked with imagining an online video call with their favourite idol or famous person and had 1 min to explain how they knew this individual and why they admired them. A sample group of 30 students was purposefully selected to represent a uniform distribution of scores. This selection was made in accordance with the recommendations of the GR&R and ICC testing methods. The aim was to minimise the influence of the distribution of testing scores, as suggested by Mehta et al. (2018). To ensure a uniform distribution of scores, the students' test scores were categorised into four levels: less than 10, 10-12, 12-14, and higher than

14. Afterward, 30 samples of video clips were intentionally chosen to have a comparable frequency of scores within each level.

This study ensured student anonymity, with no connection between student names and the selected audio clips. The recordings provided by the students did not contain any identifying information, preventing the research participants from identifying the speakers. All the filenames of the clips were changed to numerical values ranging from (1) to (30).

Three lecturers who teach English language courses at a public university in Thailand were invited to participate in the experiment and were asked for their consent as volunteers. The researchers conducted online meetings through the Zoom application with each lecturer (hereafter called 'examiner') to evaluate the English language speaking proficiency of the 30 students. The researchers began by explaining the marking criteria, which include five assessment criteria: 'Content', 'Grammar & Vocabulary', 'Pronunciation', 'Fluency', and 'Presentation style'. This was done to ensure that each examiner had a clear understanding of these criteria. A session of recording preliminary review and rater training was not conducted before the experiment, despite being aware of their benefits. This decision was made in line with the primary objective, which is to align with the established traditional practices of the organisation where the three examiners are affiliated. Additionally, this decision aligns with the common practice among classroom teachers who base their marking assessments solely on provided rubrics (Jeong, 2015; Knoch et al. 2007). This study aims to propose policy changes or strategies if the experimental results indicate low marking reliability issue.

The researchers played individual audio clips of the speaking assessments, and the examiners could request to replay each clip multiple times without any limitations. They could also take breaks before listening to the next clip without any time restrictions. After reviewing each recording clip, the examiners were asked to record the assessment scores in the provided data recording form. They then returned the completed forms to the researcher via email.

After a 2-week interval since the first evaluation, the researchers conducted another Zoom meeting with the same examiners. During this meeting, the researchers requested the examiners to evaluate and provide scores for the same set of 30 recordings. However, the order of the clips had been randomised and was different from the first round. After evaluating the second round of assessments, the examiners sent the scores for all 30 clips to the researchers via email. These two Zoom meetings did not involve any recording of audio, images, or videos whatsoever.

The evaluation scores received from the three examiners were analysed for intra-rater and inter-rater reliability using the GR&R and ICC methods. After analysing the scores using both tools, the researchers conducted a group discussion with the three examiners to address clips that exhibited repeatability and reproducibility problems, as well as those that received consistent scores from all three examiners. The purpose of this discussion was to identify the features of the video clips that influenced score variations and consistency, and to develop guidelines for improving the standard of assessing students' speaking abilities in the future. The research findings are reported here only as a general overview without disclosing any information that can identify or be linked to the examiners to ensure their anonymity.

Results of the GR&R study

This section aims to demonstrate the practical application of the GR&R study in analysing the reliability of marking language proficiency tests using a real-life case. The collected data is presented in Table 3, with calculations displayed in the 4th, 7th, 10th, 11th columns, and the final row of the table.

The GR&R results, obtained using Minitab software, are depicted in Fig. 1, and the manual calculations are provided in the Appendix. These results illustrate how the total variability in the observed scores is attributed to various sources, including examiner repeatability and their reproducibility (combined as 'Total Gauge R&R' in Fig. 1), as well as clip-to-clip variability, which pertains to the variability in the scores due to different clips. It is worth noting that the 'Part-to-Part' variability in the results provided by Minitab software, as shown in Fig. 1, is equivalent to the clip-to-clip variation, or σ_C, explained in the previous section.

Table 3 The data from the case of three examiners marking 30 clips twice

Recording clips	Examiner 1			Examiner 2			Examiner 3			Clip Average
	Trial 1	Trial 2	R	Trial 1	Trial 2	R	Trial 1	Trial 2	R	
1	16	17	1	16	16	0	9	11	2	14.17
2	11	12	1	14	12	2	8	8	0	10.83
3	16	16	0	14	15	1	12	12	0	14.17
4	13	15	2	14	14	0	11	12	1	13.17
5	15	15	0	18	14	4	12	14	2	14.67
6	13	14	1	14	11	3	7	10	3	11.50
7	16	18	2	16	18	2	9	13	4	15.00
8	12	11	1	13	13	0	6	10	4	10.83
9	16	15	1	14	10	4	10	13	3	13.00
10	16	15	1	10	11	1	8	9	1	11.50
11	15	11	4	12	10	2	8	9	1	10.83
12	11	12	1	8	10	2	7	7	0	9.17
13	15	15	0	6	11	5	5	7	2	9.83
14	13	12	1	10	11	1	6	5	1	9.50
15	15	15	0	15	13	2	15	16	1	14.83
16	14	15	1	12	12	0	9	9	0	11.83
17	16	14	2	6	13	7	14	14	0	12.83
18	13	12	1	11	12	1	6	6	0	10.00
19	13	12	1	6	11	5	9	13	4	10.67
20	17	17	0	16	17	1	15	16	1	16.33
21	13	11	2	10	9	1	5	6	1	9.00
22	16	16	0	12	12	0	12	11	1	13.17
23	16	18	2	14	16	2	15	15	0	15.67
24	13	12	1	8	8	0	6	6	0	8.83
25	15	16	1	14	14	0	9	14	5	13.67
26	12	11	1	11	11	0	5	9	4	9.83
27	17	17	0	18	16	2	14	17	3	16.50
28	15	13	2	14	14	0	11	14	3	13.50
29	12	10	2	8	12	4	10	12	2	10.67
30	15	14	1	8	15	7	10	15	5	12.83
Column average	14.183		1.10	12.383		1.97	10.267		1.80	

In the context of this study, repeatability signifies the variation arising from the same examiner marking the same clip twice, while reproducibility reflects the variation resulting from different examiners marking the same item. Ideally, repeatability and reproducibility should contribute minimally to the overall variability, with differences between clips (Part-to-Part) accounting for the majority of the variability. The distribution of variability should be reflected in the '%Contribution' displayed in Fig. 1, indicating the relative contribution of different sources of variation to the total variation. However, in this particular case, a contrasting situation is observed. Figure 1 reveals a significant percentage of the total gauge R&R (63.93%), which is nearly twice as high as the clip variability (36.07%). The primary source of data variability is found to be the reproducibility issue, as evidenced by its contribution percentage (42.73%), which is twice as high as that of the repeatability issue (21.19%).

The column labelled '%Study Var' or '%*SV*' in Fig. 1 also contains significant information. Please note that the %*SV* of the total gauge R&R, as provided by Minitab, represents the same value as %*P*/*TV* explained previously. As mentioned earlier, this value should not exceed 30% of the study variation, and it is considered ideal to keep it below 10%. However, in this study, %*P*/*TV* is recorded at 79.95%, indicating a severe deviation from the desired range. Such a high value suggests that the current assessment system, involving the three examiners, the rubric, and the testing environment, is deemed unacceptable, and it indicates a likelihood of producing unreliable marking results.

Gage R&R Study - XBar/R Method

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	6.23878	63.93
Repeatability	2.06825	21.19
Reproducibility	4.17053	42.73
Part-To-Part	3.52060	36.07
Total Variation	9.75938	100.00

Source	StdDev (SD)	Study Var (6 * SD)	%Study Var (%SV)
Total Gage R&R	2.49776	14.9865	79.95
Repeatability	1.43814	8.6288	46.04
Reproducibility	2.04219	12.2531	65.37
Part-To-Part	1.87633	11.2580	60.06
Total Variation	3.12400	18.7440	100.00

Number of Distinct Categories = 1

Fig. 1 GR&R results derived by Minitab software

Figure 1 also provides crucial information regarding the number of distinct categories, denoted as '*ndc*'. As mentioned earlier, a measurement system should ideally have an *ndc* value of at least five to demonstrate its capability to differentiate between different clips or samples. However, in this study, the *ndc* is recorded at only 1. This indicates that the current system is unable to distinguish the varying speaking performances of the 30 students. In other words, this implies that the marking system lacks the necessary sensitivity to identify differences in the speaking performances of the students. This limitation raises concerns about the system's ability to provide accurate and reliable assessments, as it fails to recognise variations that may exist among the students' performances.

The analysis using Minitab software offers an advantage over manual calculations in terms of providing several types of graphs to summarise the results, as demonstrated in Figs. 2 to 5.

Figure 2 displays the $\overline{X}$-chart and R-chart generated from the performances of the three examiners. The R-chart is a control chart illustrating the consistency within each examiner by plotting ranges (maximum data − minimum data). *UCL* and *LCL* denote the upper and lower control limits for the overall range of the marking system. In an ideal scenario, the range should be zero. When a point exceeds *UCL*, it signifies that the examiner's scores are inconsistent, and the level of inconsistency goes beyond common causes of variation. Observing the figure, the first examiner generally has the lowest ranges, indicating the best repeatability among all the examiners. The second examiner has a few out-of-control points, suggesting the highest degree of inconsistency compared to the others. The $\overline{X}$-chart compares the fluctuation of average scores for the 30 clips assessed by each examiner. Ideally, the clips selected for the GR&R study should cover a wide range of scores. Consequently, this chart should exhibit higher variation between assessment scores, surpassing the specified control limits and resulting

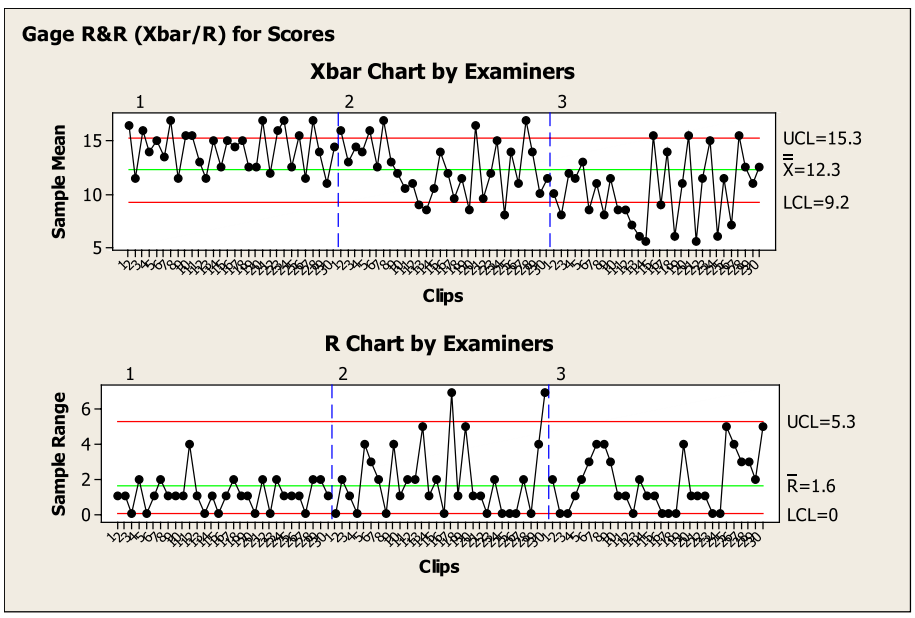

Fig. 2 $\overline{X}$-chart and R-chart by the three examiners

in several out-of-control points. Moreover, the fluctuation of scores should be similar among the three examiners. However, based on the average results, Fig. 3 indicates that the three examiners assigned varying scores for each clip, with the first examiner consistently awarding the highest scores.

Figure 3 illustrates the distribution of scores for each clip. The dots represent the scores, while the circle-cross symbols indicate the means. It is noticeable that Clips 3, 4, 15, 20, 23, 27, 28, and 29 have narrow ranges (between 2 and 4), whereas Clips 7, 13, and 17 show the widest range (between 9 and 10).

Figure 4 presents a boxplot of scores assigned by each examiner. In general, Examiner 1 assigns the highest scores, followed by Examiners 2 and 3. Furthermore, Examiner 1's scores exhibit the narrowest distribution. Figure 5 displays the plot of average scores by clips and examiners. In summary, Figs. 4 and 5 collectively provide a clear illustration of the reproducibility issue among the three examiners.

To evaluate the effectiveness and validity of employing the GR&R study for marking reliability analysis, its results are compared with those obtained through the ICC method, a commonly utilised approach in the field of education. The ICC method considers a coefficient value of 0.70 as acceptable, above 0.80 as good, and above 0.90 as excellent (Saeed et al. 2019). Tables 4 and 5 present the results of the ICC analysis for both intra- and inter-rater reliability, along with their respective 95% confidence intervals (*CI*). These findings align with some aspects of the GR&R study. Specifically, when examining the obtained ICC indices, the inter-rater reliability emerges as a significant concern within this assessment system, as it generally yields lower ICC values compared to the intra-rater analysis. Within the realm of intra-rater reliability, the first examiner demonstrates the highest ICC, indicating the best repeatability among the three examiners. While Examiner 3's ICC is only slightly lower than that of the first examiner, its 95% *CI* is significantly

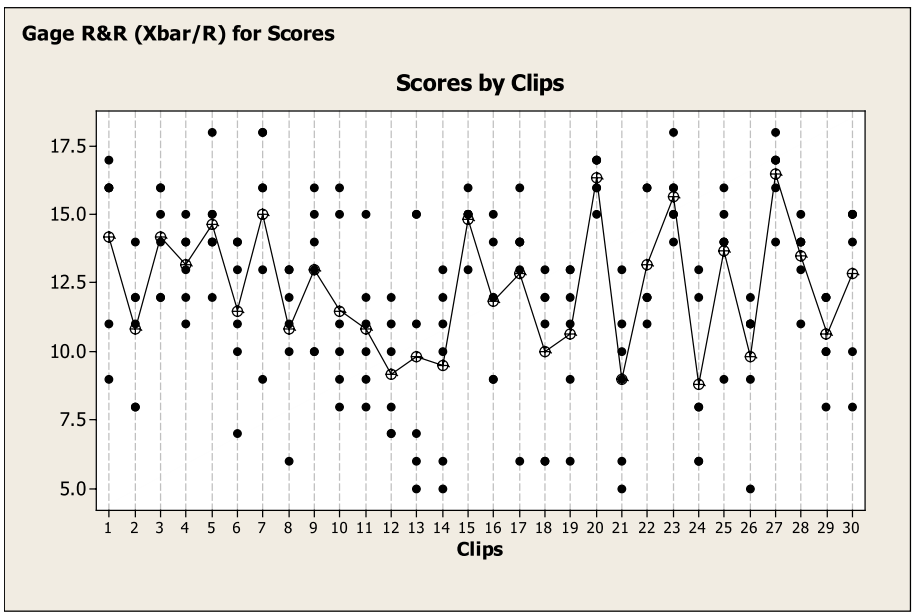

Fig. 3 The individual value plot of scores by clips

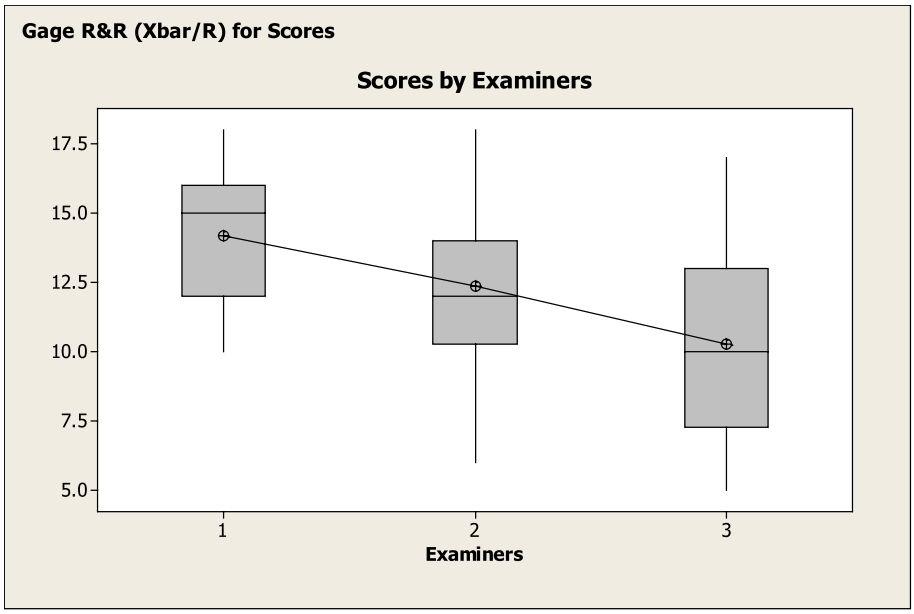

Fig. 4 The boxplot of scores by examiners

broader, suggesting higher variability in marking repeatability. Conversely, the second examiner exhibits the lowest ICC, along with a wide range in the 95% interval.

Despite these consistent findings, it seems that ICC may not have sufficient sensitivity to detect reliability issues, as most of the indices in Tables 4 and 5 surpass the acceptable threshold (0.70). Solely relying on the ICC analysis might lead to the general conclusion that the assessment system is currently acceptable and does not require immediate

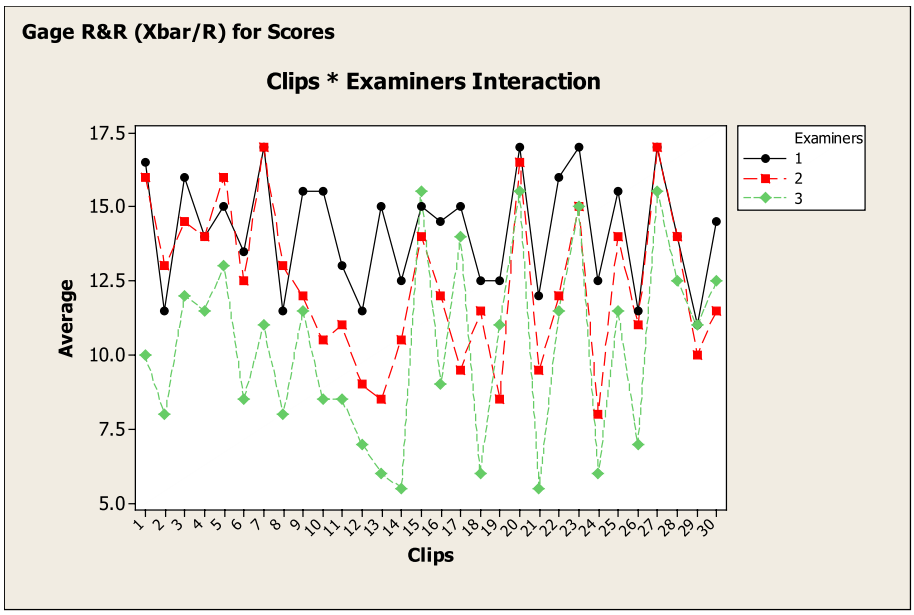

Fig. 5 The plot of average scores by clips and examiners

Table 4 Intra-rater reliability with the intraclass correlation coefficient

	ICC	95 % CI for ICC
Examiner 1	0.869	(0.728, 0.937)
Examiner 2	0.719	(0.417, 0.865)
Examiner 3	0.864	(0.345, 0.954)

Table 5 Inter-rater reliability with the intraclass correlation coefficient

	ICC	95 % CI for ICC
Trial 1	0.502	(0.040, 0.755)
Trial 2	0.75	(0.409, 0.888)

action. However, this conclusion contradicts the earlier findings from the GR&R study. This highlights the advantages of adopting the GR&R approach in the field of education.

Discussion on the reliability problems

As illustrated in Fig. 3, recordings numbered 3, 4, 15, 20, 23, 27, 28, and 29 exhibit a narrow range of scores, implying high reliability in the evaluations provided by the three examiners. During the post-interview phase, when these clips were re-observed, the examiners offered feedback, noting that all these clips featured students speaking English clearly and with a strong, audible voice. This clarity enabled the examiners to discern the accuracy of grammar usage. Additionally, each of these clips showed students using body language during their presentations, indicating effective communication of the subject matter. This alignment between verbal and nonverbal expression underlined the students' comprehensive understanding of the topics they were discussing.

Figure 6 presents a scatterplot illustrating the average scores for each clip and the corresponding ranges, which are derived from the six scores provided by the three examiners. The plot reveals a noticeable negative correlation trend, indicating that students proficient in the speaking skill tend to encounter fewer reliability issues during the evaluation of their recorded performances. This alignment between the trend depicted in Fig. 6 and the feedback from all three examiners underscores the importance for students to prioritise thorough preparation, topic comprehension, and accurate pronunciation to ensure higher assessment reliability.

When examining the range data of each examiner in the fourth, seventh, and tenth columns of Table 3 to assess repeatability, it became evident that twelve clips (Clips 5, 7, 8, 9, 11, 13, 17, 19, 25, 26, 29, and 30) exhibited issues, as their ranges (from at least one examiner) exceeded 3. An underlying factor affecting repeatability lies in the incorrect estimation of the overall group's language proficiency. The examiners' interviews revealed that in the first round of scoring, their focus was more on the students' fluency, and they were unaware of the language proficiency levels across all 30 clips. Some examiners admitted to overestimating the group's proficiency during this phase, potentially influenced by the clips they had assessed earlier. The examiners generally compared whether the students spoke more fluently and accurately compared to the clips they had listened to previously. The standard of evaluations seemed to change in the second

Fig. 6 The scatterplot of the average scores and the ranges

round; the focus shifted towards content, grammar structures, and vocabulary as the examiners had already recognised this group of students' proficiency level in their use of the English language in the first round.

Furthermore, all three examiners shared the perspective that considering multiple criteria simultaneously can result in variability in scoring consistency. The complexity of the rubric might prompt them to emphasise certain aspects in the initial round and subsequently pivot to address different criteria in the subsequent round. Another issue affecting the repeatability of examiners involves the limited accessibility and universality of chosen topics. This term emphasises the idea that the reliability of marking can be affected by the topics chosen by students. If a topic is too specific and only understandable by a particular group of people, it may hinder the examiner's ability to accurately assess the student's communication skills. For instance, a student spoke about a favourite video gamer, narrating stories related to various games that the examiners were unfamiliar with. This lack of familiarity led the examiners to assign a lower score in the initial assessment round. However, upon re-evaluating the student in the second round, the examiners gained a clearer understanding of the content being discussed. This shift in comprehension prompted the examiners to revise the student's scores upward. This circumstance is consistent with the research by Jensen and Hansen (1995), which indicated that prior knowledge significantly influenced participants' listening comprehension, particularly in technical topics compared to non-technical ones. This illustrates how the specificity of the topic can impact the repeatability of examiners and contribute to inconsistencies. In another clip where a student discussed a Thai folk song singer, Examiner 3 encountered similar difficulty in understanding. In contrast, the other two examiners, who were more familiar with the Thai context, assigned nearly similar scores, showing alignment in their evaluations.

When considering reproducibility, an analysis of Fig. 5 and the raw data revealed that the recordings numbered 1, 7, 10, 13, 14, 16, 17, 18, 21, and 24 exhibit inconsistencies among all three examiners. During the group discussion, audio clips that posed reproducibility issues were reviewed by the examiners for re-evaluation.

The issues of unknown levels of language proficiency of the entire class, the complexity of the rubric, and limited accessibility and universality of the chosen topic continue to surface in the context of reproducibility. Each examiner held different perspectives on the overall group performance and emphasised various criteria in each evaluation round. This rubric's complexity can lead to human errors as examiners may focus on different aspects or inadvertently overlook certain errors made by the students. For example, in the case of Clip 13, Examiner 1 awarded a higher score due to recognising the student's accurate pronunciation. However, Examiner 2 noticed that the student used vocabulary beyond her proficiency level, resulting in unnatural language usage. Meanwhile, Examiner 3 observed that the student employed vocabulary that was too advanced for her proficiency level, leading to a presentation that appeared rigid due to memorisation and extended pauses caused by memory lapses. As a result, scores were deducted for issues related to fluency and presentation. This issue aligns with Jeong (2015), who found that the inconsistencies among raters stem from their personal impressions. Some raters tend to assign higher scores for a particular criterion compared to other criteria. Finally, when a student selects a topic that is not universally understood or is limited to a specific group, each examiner, based on their varying prior knowledge, may interpret it differently. This highlights the challenges in achieving reproducibility in assessments.

Another factor contributing to inconsistencies among the three examiners is their familiarity with accents. Examiner 3, who has only recently begun teaching in Thailand, mentioned that he was unaccustomed to the students' accents, which led to lower scores due to challenges in comprehending their speech.

Another factor contributing to the occurrence of reproducibility problems is the unclear rubric. For instance, in the 'Grammar & Vocabulary' category, only parts of speech, sentence structures, and tenses are specified. Issues arise when students make errors in other aspects of grammar. Even though the researcher had explained the descriptors of each scale and criterion before the marking process started, uncertainty still surfaced. All three examiners shared the same uncertainty about whether points should be deducted for errors not listed in the rubric. Another example of an issue arising from an unclear rubric (in the 'Content' category) is when the students go off-topic. From the interviews, some students misunderstood the topic while speaking. The task required students to imagine having an online conversation with a celebrity they admire, emphasising that points would be awarded for students who provided reasons why they admired the celebrity with strong supporting details or explanations. However, several students interpreted it as narrating stories about their favourite celebrity. They discussed why they liked them, leading to the use of 'he/she' instead of the pronoun 'you' to refer to the celebrity they were supposed to converse with. Consequently, examiners were uncertain about the appropriate point deduction in such cases.

Another concern is that some students submitted audio clips with very low volume, excessive background noise, or while wearing face masks. This significantly impacted the audio quality and made it almost impossible for examiners to discern. The challenge here

is that while volume is one of the specified criteria, it also affects scoring in other areas such as content, grammar structures, and vocabulary. Understanding what is being said remains difficult. Similar to the previously mentioned issue, the rubric does not offer specific guidelines for this aspect. It becomes unclear how scores should be deducted when the audio quality is so poor that comprehension is compromised.

This section underscores the complexity of the evaluation process, summarising the factors that influence examiners' repeatability and reproducibility in Table 6. Addressing these factors can provide insights into implementing strategies to enhance assessment processes, as discussed in the next section.

Practical implications

Table 7 outlines potential strategies to mitigate the causes of inconsistency in marking English-speaking tests. Firstly, according to the interviews, examiners mentioned that in the first round of scoring, they did not yet have a comprehensive view of the proficiency levels of all the students. Consequently, they provided scores based on previous

Table 6 Factors influencing the examiners' repeatability and reproducibility

Factors	Repeatability	Reproducibility
Unknown levels of language proficiency of the entire class	/	/
Influence of previously assessed clips on scoring	/	
Complexity of the rubric	/	/
Limited accessibility and universality of the chosen topic	/	/
Examiners' lack of familiarity with a student's accent		/
Lack of clarity in the rubric		/
Poor recording quality	/	/

Table 7 Implementing strategies to minimise the inconsistency factors

Factors	Strategies			
	Conducting a preliminary review	Conducting a marking calibration	Creating well-defined criteria in a rubric for examiners	Providing a clear instruction along with a scoring rubric to students
Unknown levels of language proficiency of the entire class	/	/		
Influence of previously assessed clips on scoring	/	/	/	
Complexity of the rubric		/	/	
Limited accessibility and universality of the chosen topic				/
Examiners' lack of familiarity with a student's accent	/			
Lack of clarity in the rubric		/	/	
Poor recording quality				/

clips they had assessed. Furthermore, according to Examiner 3, there were challenges in understanding the accents and pronunciation of certain students. Therefore, it is advisable for examiners to review sample responses several times before commencing the actual marking process. This practice can significantly assist examiners in comprehending the overall performance of the entire group. It empowers examiners to make scoring decisions without relying solely on the individual speaking abilities of the students under assessment or referring to prior evaluations. Moreover, the preliminary review enables examiners to adapt perceptually to foreign accents, fostering consistent evaluations across all assessment rounds by exposure to such accented speech (Huang, 2013; Huang et al. 2018).

However, a limitation of the preliminary review approach is its impracticality when dealing with a large number of students to be evaluated due to time constraints. Therefore, providing calibration meetings or rater training sessions could offer a solution to streamline the process, reducing the time spent on preliminary reviews. The marking calibration, using sample videos, allows a group of examiners to collectively experience a diverse range of response patterns and varying levels of proficiency. This also ensures that examiners receive comprehensive information on the assessment criteria and protocols (Lumley & McNamara, 1995). Moreover, the training sessions enable examiners to practice scoring a set of students' responses together, engage in discussions about their interpretations of the criteria and a given rubric, determine detailed guidance on handling challenging cases, and then reach a consensus (Davis, 2016; Doosti & Safa, 2021; Jeong, 2015). These goals can be achieved by incorporating exemplars that illustrate how language proficiency at each level corresponds to distinct speaking abilities (Jonsson & Svingby, 2007). The timing of organising training or calibration meetings is also crucial. It should be scheduled close to the time when the examinations, both oral and written, need to be assessed. If there is a significant time gap between them, it can lead to the occurrence of inconsistency issues (Weigle, 1998).

In terms of the third strategy, 'Creating well-defined criteria in a rubric for examiners,' it is possible to maintain rater reliability without the need for extensive rater training by employing uncomplicated rubrics with only three or four criteria and three levels (Koizumi et al. 2022). Consequently, this approach can alleviate inconsistent scoring resulting from factors such as the rubric's complexity, examiner bias which may arise when examiners prioritise certain evaluation aspects over others, unclear rubrics, and even the influence of marking orders. Marking calibration and refining criteria thus emerge as effective methods to address these challenges.

Another strategy is to provide a clear instruction of the assignment along with a scoring rubric to students and make sure that they all understand. This could mitigate the issues of off-topic responses and the quality of recordings. There is evidence that examiners frequently penalise test-takers for delivering responses that are off-topic or classify them as ineligible for scoring. Nonetheless, some examiners often exhibit a willingness to grant leniency to test-takers who presented responses deviating from the intended topic by compensating their score with their proficiency (Burton, 2020). To address this, the rubric criteria for content relevance should offer additional clarification on which responses are considered eligible for scoring to promote fairness and consistency in assessment among examiners. Furthermore, language teachers should emphasise

the importance of relevance as a key aspect of the scoring criteria. When students are aware that off-topic responses will be penalised, they will be more likely to attentively listen to the task prompt and generate responses that are directly relevant and responsive. Another strategy that could improve instruction and lessen students' confusion is presenting marking criteria in advance. Presenting the criteria and providing additional illustrative examples to students facilitates their comprehension of the important dimensions associated with each criterion during the assessment (Jonsson & Svingby, 2007). Lastly, they should ensure good sound quality by finding quiet locations to prevent background noise, as this can have a significant impact on scoring.

Conclusions

This research addresses concerns related to reliability issues in speaking proficiency tests through the application of the GR&R approach. By conducting a comprehensive analysis encompassing both intra- and inter-rater reliabilities, this study shed light on the factors that contribute to discrepancies among examiners in assessing speech characteristics. The results of this investigation were subsequently cross-compared with the commonly used ICC method, demonstrating the efficacy of GR&R in detecting marking reliability issues in a more sensitive manner. This highlights the significance of GR&R as a better alarm measurement tool for assessing reliability.

The experiment not only confirms the effectiveness of the GR&R approach but also demonstrates the application of Minitab software in facilitating in-depth analysis. The graphical tools generated through the software reinforce the analysis of discrepancies among examiners and delineate the distribution of marking data within each examiner's evaluation.

Furthermore, this research delves into various causes of reliability problems in the marking process by utilising a case study approach. By examining clips with varying levels of consistency and engaging in discussions with examiners, key factors influencing scoring inconsistencies can be identified. These factors include the unknown of overall group performance, the sequence of work presentation, a complex and unclear rubric, the student's chosen topic, the examiners' unfamiliarity with a student's accent, and poor recording quality. Importantly, this study not only identifies these common root causes but also proposes practical strategies to improve the precision of the measurement system.

The findings of this study hold substantial importance for language proficiency assessment stakeholders, including educational institutions, test developers, and policy makers. It offers insights to improve the fairness and accuracy of speaking proficiency tests by addressing reliability issues.

This research not only presents a valuable contribution to the field of language proficiency testing but also emphasises the importance of employing advanced statistical tools such as the GR&R approach to enhance the quality of assessments. The practical solutions proposed in this study offer a checklist for improving the reliability of speaking proficiency tests, benefiting both educators and students alike in their pursuit of accurate and fair language assessment.

In this study, the sample size of 30 individuals aligns with the theoretical recommendation, which suggests that a GR&R analysis should involve a sample size of at least ten individuals. However, it is important to note that increasing the sample size in future studies may yield different reliability results and allow other causes of marking inconsistency to emerge. Additionally, this study used only a single speaking test item to assess proficiency. If different test items with varying contexts were included, it could potentially lead to different analytical results.

Appendix

Manual calculation for the GR&R analysis

Step 1: From the data shown in Table 3, calculate $\sigma_{Repeatability}$ using Equation (4), $d_2 = 1.128$.

$$\sigma_{Repeatability} = \frac{\overline{\overline{R}}}{d_2} = \frac{(1.10 + 1.97 + 1.80)/3}{1.128} = 1.43$$

Step 2: Estimate $\sigma_{Reproducibility}$ using Equation (5), $R_{\overline{X}} = 14.183 - 10.267 = 3.91667$, $d_2^* = 1.91155$

$$\sigma_{Reproducibility} = \sqrt{\left(\frac{R_{\overline{X}}}{d_2^*}\right)^2 - \left(\frac{{\sigma_{Repeatability}}^2}{n \times m}\right)} = \sqrt{\left(\frac{3.91667}{1.91155}\right)^2 - \left(\frac{1.438^2}{30 \times 2}\right)} = 2.04$$

Step 3: Calculate the GR&R using Equation (6).

$$GR\&R = \sigma_{Gauge} = \sqrt{{\sigma_{Repeatability}}^2 + {\sigma_{Reproducibility}}^2} = \sqrt{1.43^2 + 2.04^2} = 2.49$$

Step 4: Estimate σ_C using Equation (7). From Table 3, $R_C = 16.50 - 8.83 = 7.67$, $d_2^* = 4.147$

$$\sigma_C = \frac{R_C}{d_2^*} = \frac{7.67}{4.147} = 1.85$$

Step 5: Calculate σ_{Total} using Equation (8).

$$\sigma_{Total} = \sqrt{{\sigma_{Gauge}}^2 + {\sigma_C}^2} = \sqrt{2.49^2 + 1.85^2} = 3.10$$

Step 6: Calculate $\%P/TV$ and ndc using Equations (9) and (10), respectively.

$$\%P/TV = \left(\frac{\sigma_{Gauge}}{\sigma_{Total}}\right) \times 100 = \left(\frac{2.49}{3.10}\right) \times 100 = 80.32$$

$$ndc = \frac{1.414 \times \sigma_C}{\sigma_{Gauge}} = \frac{1.414 \times 1.85}{2.49} = 1.05$$

Please note that σ_C, σ_{Total}, $\%P/TV$, and ndc might not be exactly equal to the results obtained from Minitab software, as illustrated in Fig. 1. This discrepancy arises from using different decimal digits in the calculations.

Abbreviations

GR&R or σ_{Gauge}	Marking system variation caused by gauge repeatability and reproducibility
ICC	Intraclass correlation coefficient
ANOVA	Analysis of variance
SSR	Scale separation reliability
TPA	Teacher performance assessment
MSA	Measurement system analysis
%P/T	Percentage of precision-to-tolerance
%P/TV	Percentage of precision-to-total variation
ndc	The number of distinct categories
AIAG	Automotive Industry Action Group
$\sigma_{Repeatability}$	Variation caused by the examiner repeatability
R	Range
$\overline{R}$	Average of all ranges
m	The number of replications
$\sigma_{Reproducibility}$	Variation caused by the examiner reproducibility
$R_{\overline{X}}$	Range across all examiners
n	The number of sample clips
k	The number of subgroups
σ_C	Clip-to-clip variation
R_C	Range across all sample clips
σ_{Total}	Total variation
$\overline{\overline{X}}$	Average value
UCL	Upper control limit
LCL	Lower control limit
CI	Confidence intervals

Acknowledgements
This research was supported by a research grant provided by International College, Khon Kaen University. In addition, the authors extend their gratitude to the three participants who took part in the experiments of this study, referred to as 'examiners' throughout this manuscript.

Authors' contributions
Pornphan S served as the principal investigator, conceptualising the study, reviewing relevant literature, engaging with participants, and overseeing the experimental data collection and the results discussion. She was a major contributor in drafting most sections of the manuscript. Panitas S contributed by reviewing relevant literature, performing GR&R analysis, and writing sections of the manuscript related to statistical analyses. UP conducted the ICC analysis and read and approved the final manuscript. JK, PS, and DO participated in the results discussion and read and approved the final manuscript.

Funding
This research was financially supported by International College, Khon Kaen University.

Availability of data and materials
All data generated or analysed during this study are included in this published article.

Declarations

Ethics approval and consent to participate
The research has received approval for ethics declarations, and the need for consent has been waived by the Centre for Ethics in Human Research at Khon Kaen University. It falls under the category of 'Exemption Research'.

Competing interests
The authors declare that they have no competing interests.

Received: 16 September 2023 Accepted: 20 December 2023
Published online: 11 January 2024

References

AIAG (2010). *Measurement system analysis (MSA)*, (4th ed.,). Automotive Industry Action Group.
Akeju, S. A. (1972). The reliability of general certificate of education examination English composition papers in West Africa. *Journal of Educational Measurement*, *9*(3), 175–180.
Aprianoto, D., & Haerazi, D. (2019). Development and assessment of an interculture-based instrument model in the teaching of speaking skills. *Universal Journal of Educational Research*, *7*(12), 2796–2805.
Başaran, M., Özalp, G., Kalender, İ., & Alacacı, C. (2015). Mathematical knowledge and skills expected by higher education in engineering and the social sciences: Implications for high school mathematics curriculum. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(2), 405–420.
Bird, F. L., & Yucel, R. (2013). Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme. *Assessment & Evaluation in Higher Education*, *38*(5), 536–553.

Bland, L. M., & Gareis, C. R. (2018). Performance assessments: A review of definitions, quality characteristics, and outcomes associated with their use in k-12 schools. *Teacher Educators' Journal*, *11*, 52–69.
Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.
Burton, J. D. (2020). Raters' measurement of test-task authentic engagement in L2 oral-performance assessment: An exploration of scale development. *System*, *90*, 102233.
Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219.
Cepova, L., Kovacikova, A., Cep, R., Klaput, P., & Mizera, O. (2018). Measurement system analyses - gauge repeatability and reproducibility methods. *Measurement Science Review*, *18*(1), 20–27.
Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135.
Detey, S., Fontan, L., Le Coz, M., & Jmel, S. (2023). Computer-assisted assessment of phonetic fluency in a second language: A longitudinal study of Japanese learners of French. *Speech Communication*, *125*, 69–79.
DeVellis, R. F. (2005). Inter-rater reliability. *Encyclopedia of Social Measurement*, *2*, 317–322.
Doosti, M., & Safa, M. A. (2021). Fairness in oral language assessment: Training raters and considering examinees' expectations. *International Journal of Language Testing*, *11*(2), 64–90.
Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–606.
Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.
Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, *13*(1), 25–41.
Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41*(3), 770–785.
Huang, L. F., Kubelec, S., Keng, N., & Hsu, L. H. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, *8*, 14.
Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing*, *12*(1), 99–119.
Jeong, H. (2015). Rubrics in the classroom: do teachers really follow them? *Language Testing in Asia*, *5*, 6.
Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.
Khan, R. M. I., Radzuan, N. R. M., Shahbaz, M., & Kumar, T. (2020). An investigation of the reliability analysis of speaking test. *Asian EFL Journal*, *27*(31), 356–373.
Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*((1), 26–43.
Koizumi, R., Hatsuzawa, S., Isobe, R., & Matsuoka, K. (2022). Rater reliability in speaking assessment in a Japanese senior high school: Case of classroom group discussion and debate. *JALT Journal*, *44*(2), 281–322.
Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing*, *35*, 2409–2431.
Low, S. M., Lee, S. Y., & Yong, W. K. (2009). Application of GR&R for productivity improvement. In *The 11th Electronic Packaging Technology Conference*. Singapore: Institute of Electrical and Electronics Engineers (IEEE).
Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71.
Lyness, S. A., Peterson, K., & Yates, K. (2021). Low inter-rater reliability of a high stakes performance assessment of teacher candidates. *Education Sciences*, *11*, 648.
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.
Mehta, S., Bastero-Caballero, R. F., Sun, Y., Zhu, R., Murphy, D. K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Statistics in Medicine*, *37*, 2734–2752.
Montgomery, D. C. (2013). *Statistical Quality Control*, (7th ed.,). John Wiley & Sons.
Mukundan, J., & Nimehchisalem, V. (2012). Evaluating the validity and economy of the English language teaching textbook evaluation checklist. *World Applied Sciences Journal*, *20*(3), 458–463.
Naqvi, S., Srivastava, R., Al Damen, T., Al Aufi, A., Al Amri, A., & Al Adawi, S. (2023). Establishing reliability and validity of an online placement test in an Omani higher education institution. *Languages*, *8*(1), 61.
Nimehchisalem, V., Mukundan, J., Rafik-Galea, S., & Samad, A. A. (2021). Assessment of the analytic scale of argumentative writing (ASAW). *Pertanika Journal of Social Science and Humanities*, *29*(S3), 1–25.
Pan, J.-N. (2006). Evaluating the gauge repeatability and reproducibility for different industries. *Quality and Quantity*, *40*(4), 499–518.
Ploypanichcharoen, K. (2010). *Measurement system analysis (MSA)*, (2nd ed.,). TPA Publishing.
Porter, J. M., & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance. *International Journal of Educational Policies*, *5*(2), 74–87.
Rashid, S., & Mahmood, N. (2020). High stake testing: Factors affecting inter-rater reliability in scoring of secondary school examination. *Bulletin of Education and Research*, *42*(2), 163–179.
Saeed, K. M., Ismail, S. A. M. M., & Eng, L. S. (2019). Malaysian speaking proficiency assessment effectiveness for undergraduates suffering from minimal descriptors. *International Journal of Instruction*, *12*(1), 1059–1076.
Sennaroglu, B., & Yurtsever, O. (2018). Evaluating measurement system by gauge repeatability and reproducibility. In *The 2nd European International Conference on Industrial Engineering and Operations Management*. Paris, France: The IEOM Society International.
Soemantri, D., Mustika, R., & Greviana, N. (2022). Inter-rater reliability of reflective-writing assessment in an undergraduate professionalism course in medical education. *Education in Medicine Journal*, *14*(1), 87–97.

Statistics Solutions. (2013). *ANOVA (Analysis of Variance)*. Retrieved 21 February 2023 from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/anova/
Stokes, E. K. (2011). *Rehabilitation Outcome Measures*. Elsevier: Churchill Livingstone.
Stuart, N. J., & Barnett, A. L. (2023). The writing quality scale (WQS): A new tool to identify writing difficulties in students. *British Journal of Special Education*, 1–10. https://doi.org/10.1111/1467-8578.12464.
Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment & Evaluation in Higher Education*, *22*(3), 289–305.
Trevisan, M. S. (1991). Reliability of performance assessments: Let's make sure we account for the errors. In *The Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors*. Chicago, Illinois: Education Resources Information Center, the Institute of Education Sciences, the United States Department of Education.
Wang, P. (2009). The inter-rater reliability in scoring composition. *English Language Teaching*, *2*(3), 39–43.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.
Wheeler, D. J. (2006). *EMP III (Evaluating the measurement process): Using imperfect data*. SPC Press.
Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252.
Zhao, Z. (2013). Diagnosing the English speaking ability of college students in China – Validation of the Diagnostic College English Speaking Test. *RELC Journal*, *44*(3), 341–359.

Publisher's Note