Assessing Writing 54 (2022) 100669

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Validity evidences for scoring procedures of a writing assessment task. A case study on consistency, reliability, unidimensionality and prediction accuracy

Paula Elosua [*]

University of the Basque Country, San Sebastian, Spain

ARTICLE INFO

Keywords:
Assessing writing
Scoring procedures
Validity
Rater mediated assessment

ABSTRACT

Scoring is a fundamental step in the assessment of writing performance. The choice of the scoring procedure as well as the adoption of a discrepancy resolution method can impact the psychometric properties of the scores and therefore the final pass/fail decision. In a comprehensive framework which considers scoring as part of the validation process of the scores, the aim of this paper is to evaluate the impact of rater mean, parity and tertium quid procedures on score properties. Using data from a writing assessment task applied in a professional context, the paper analyses score reliability, dependability, unidimensionality and decision accuracy on two sets of data; complete data and subsample of discrepant data. The results show better performance of the tertium quid procedure in terms of reliability indicators but a lower quality in defining construct unidimensionality.

1. Introduction

Scoring is a key step in the development and further use of any assessment tool. While in many testing contexts scoring is almost undisputed, in rater-mediated assessment of writing, it is still a matter of discussion. The systematic biases and/or measurement errors which can be introduced by raters can impact the reliability and validity of the scores and therefore call into question the quality and equity of the assessment.

Operational scoring in writing assessment usually involves two stages. In the first stage two independent raters assess the production using analytical or holistic procedures and then assign a perceived value on a quantitative scale or on different assessment domains. Once the two raters' assessments are done, the concordance among outcomes is evaluated. In case of discrepancy a second round of reviews is activated; this second step is known as adjudication (Brennan, 1996) and this function is usually assigned to experienced raters or adjudicators (Myers, 1980; Wolcott, 1998). The most common scoring procedures are rater mean, parity, tertium quid, expert, and discussion (for a review see Kim, 2011; Penny & Johnson, 2011). The rater mean procedure averages the scores assigned by each of the raters; in the parity method a third rater is incorporated to the scoring design and his/her ratings are combined with the previous ones. The tertium quid also incorporates a new rater and the new rating is combined with the closest previous one, while the most distant is erased from the scoring system. This new rater can conduct a blind review or an evaluative report based on previous ratings. The fourth resolution method is based on an expertise judgment that replaces the original ratings. In the fifth method, discussion procedure, discrepancies are resolved by exchanging viewpoints until consensus is achieved. Except for the last scoring

[*] Correspondence to: University of the Basque Country, Avda. Tolosa, 70, 20018 San Sebastian, Spain.
E-mail address: Paula.elosua@ehu.es.

https://doi.org/10.1016/j.asw.2022.100669
Received 16 October 2021; Received in revised form 12 September 2022; Accepted 19 September 2022
Available online 1 October 2022

procedure, basically all of them reduce the variance of the score distribution by narrowing the differences among raters' scores.

The impact of the different scoring procedures has been analysed in a limited way in the literature (Wind & Walker, 2020). Papers authored by Johnson's team concluded that the choice of the resolution method influences the reliability of the operational scores; they also note that the rank of candidates may vary depending on the scoring and highlight the positive impact of including an expert in the scoring process (Johnson, Penny, & Gordon, 2000; 2001; Johnson, Penny, Fisher, & Kuhs, 2003; Penny & Johnson, 2011). Kim (2011) compares the tertium quid and the parity procedure and concludes that the first is associated with higher correlation values between operational scores and external criteria; that is, the external validity of the scores is higher applying the tertium quid technique. Wind and Walker (2019, 2020) pointed out that studies on the impact of resolution procedures on inter-rater agreement should be completed with analysis on individual students' rating. They do so by exploring the impact of score resolution procedures on person fit statistics derived from the application of the many-facets model. They found that including resolution procedures in the scoring system improve the psychometric quality of the assessment, but they also concluded that there were no differences in person fit indexes across the outcomes of the resolution methods.

Beyond these conclusions, some points remain under discussion, mainly from the increasing interest in the agreement measures in rater mediated assessment as part of the validation process of the rating procedures (Johnson et. al, 2003; Huang, 2012; Knoch & Chapelle, 2018; Wind & Walker, 2020).

Validity, which has been defined as the most fundamental consideration in developing and evaluating tests, is a concept constructed and referenced to scores (for a review of the evolution of the concept of validity see (De Boeck & Elosua, 2016), in the field of language testing, Chapelle & Voss, 2014). Messick (1989) defined it as an "evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). That influential position led to the argument-based approach as formulated by Kane (1992, 2006), who highlights the importance of the interpretative arguments for the plausibility and appropriateness of the proposed interpretation/use of test scores. Today that is the characterisation of validity adopted by the standards for the use of tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), and by the field of language testing (Bachman & Palmer, 2010; Knoch & Chapelle, 2018; Knoch & Macqueen, 2020).

In this framework scoring validity can be seen as a wide umbrella that enables scoring issues to be studied by breaking the traditional approaches of reliability vs validity. (Deygers & Van Gorp, 2015; Stemler, 2004; Weir, 2005). Some psychometric models allow this kind of tradeoff between reliability and validity. The generalizability theory (Brennan, 2001), for instance, not only offers dependability (reliability) indexes but can be used to assess the discriminatory validity of the tasks being analysed (Marcoulides, 1989; Ohta, Plakans, & Gebril, 2018). The many-facets model (Eckes, 2015; Linacre, 2002) provides information about the reliability of the scores but also offers validity evidences which can be used as validation arguments about raters, domains and quality of the benchmarks. By defining measurement models which include raters, candidates and tasks, the influence of raters, task and scoring domain difficulty on observed variations in candidate ratings can be assessed, as well as some interactions between facets that may be the source of systematic errors. Those models are not new in the field of writing assessment and they are being used to assess performance (Wind & Peterson, 2018).

From this point of view, the quality of any operational scoring procedure should be analysed not only through reliability indexes, but also by studying how raters interpret and apply the scoring scales or rubrics. This idea is especially significant in the field of writing assessment (Weir, 2005), where it is not possible to separate reliability from validity, measurement error from systematic error, correlation among raters from domain interpretation, and ultimately, reliability from construct validity. In this context it is relevant to carry out systematic analyses of the operational scoring procedures and discrepancy resolution methods to evaluate their possible impact on writing performance assessment. In terms of Assessment Use Argument (AUA), the review and analysis of score resolution methods would be part of the generalization inference, and the study of raters' consistency and unidimensionality would serve to assess evaluation inference-related warrants through the study of scale properties and raters' reliability and consistency (Knoch & Chapelle, 2018).

2. Research questions

This study examined the effects of different operational scoring procedures and different discrepancy resolution methods in a wide framework of score validity by addressing the following research questions:

What is the effect of different operational scoring methods and discrepancy resolution procedures on candidates' order classification?

What is the effect of different operational scoring methods and discrepancy resolution procedures on estimates of the reliability indexes?

What is the effect of different operational scoring methods and discrepancy resolution procedures on estimates on dimensionality?

What is the effect of different operational scoring methods and discrepancy resolution procedures on estimates on classification accuracy and consistency?

3. Material and methods

3.1. Participants and instrument

The essay. The task consisted in writing a letter to the head of one service. The aim of the letter is to express the candidate's opinion about the service, explaining the most serious problems affecting it, as well as offering suggestions on aspects that need to be changed, and explaining the benefits these changes could bring. The minimum number of words was set at 180. The essay is part of a test to assess the C1 level in Basque as defined by the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). This test is managed by the Basque Institute for Public Administration (IVAP). It is a high-stakes exam whose purpose is to accredit knowledge of the Basque language. In the Basque Autonomous Community many public sector jobs require accreditation in Basque. The normalization process for the use of Basque in public administration was regulated in 1997 (Decree 86/1997).

The scoring. Essays were scored analytically using five domains: (a) Adequacy, (b) coherence, (c) cohesion, (d) richness and (d) accuracy. The definitions for each domain and the components are presented in Table 1. Each domain received a score based on a 5-point scale ranging from 0 (inadequate), 1 (minimal), 2 (sufficient) 3 (good), to 4 (very good). Detailed descriptors for each were the basis for the rating. Once each domain is assessed the scores are summed to provide the final score; a cut point for pass/fail decision was set at 10 points. Analytic marking is seen as providing more diagnostic information about students' writing competence, and according to Dunsmuir et. al (2015), it is easier to train raters to use analytic rather than holistic scales.

The examinees. Six hundred twenty-six public workers took part in this study. The participation in the exam was voluntary; there was no penalty or cost for failing the exam.

The raters. Thirty-nine raters took part in the scoring process. The number of essays scored per rater varied from 36 to 40, with 38 being the median value. In addition, ten raters took the role of adjudicators.

The rating procedure. In the scoring session, two raters independently scored each composition using an analytical approach. If two raters agreed on the fail/pass decision, scoring was complete for that essay. According to the legal decree regulating this test program, if the two raters assigned different pass/fail decisions the essay was treated as a discrepant paper and was reassessed by the adjudicator. In this testing program, the adjudicator was an expert selected from among raters with at least 3 years' experience in scoring and had demonstrated higher levels of accuracy. The experts were aware that discrepancy had occurred for the essays; however, they were not aware of the actual scores. The entire process is carried out in Basque.

3.2. Design

The rating design. The rating design was a non-complete design with sparse-rated data. Raters worked in groups of two, and each group assessed a number of essays. There was no linking across groups. To carry out the analysis we transformed this sparse-data into a rating matrix (see Fig. 1), assuming that statistically raters are random variables (Bachman, Lynch, & Mason, 1995). Under this common assumption, ratings (for example, the first, the second and the third ratings) instead of raters are considered random facets (Huang, 2012; Lee, 2005; Lin, 2017). That is, rating is the unit of analysis. For each of these ratings we analysed the complete data and the subsample composed only of discrepant data.

3.2.1. Study design. This study compares three operational scoring procedures

Rater mean (S1): The score is calculated as the averaged scores from two raters.

Parity (S2): The score is calculated as the averaged scores from two raters and the adjudicator for the situations with discrepancy among raters in the pass/fail decision.

Tertium quid (S3): The score is based on the tertium quid resolution procedure.

For each of the scoring design we differentiate among two sets of data:

Complete data.

Subsample of discrepant data.

3.3. Data analysis

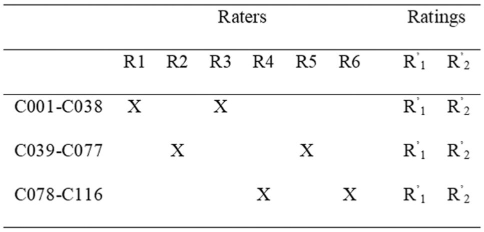
For each of the six conditions (3 scoring procedures × 2 data samples) the following analytical scheme was applied:

Descriptive analysis. (a) Score means, variances, skews and kurtosis indexes were calculated. (b) Density plots for each condition were drawn. (c) Using ANOVA for repeated measures, differences among score procedures were analysed.

Inter-rater agreement. Following the classification proposed by Stemler (2004), consistency and reliability indexes were estimated. To assess consistency among raters, percentages of pass/fail decision and Spearmen rank-correlation index were computed. The reliability of the data was estimated using two frameworks: a) the classical test theory and, b) the generalizability model. For the first, alpha coefficient of internal consistency was calculated, and for the latter, the phi or Φ index of dependability which emphasises the

Table 1
Definitions and components of the five analytical domains.

Adequacy	Approached the subject. Explanations and arguments, not mere examples and anecdotes.
Coherence	Explain the ideas in a clear, concise and appropriately organised manner. The structure of the text is appropriate.
Cohesion	Related ideas and sentences well and appropriately explained; Use of elements of cohesion to give unity to the text.
Richness	General lexicon correctly and appropriately used. Ideas expressed using appropriate complex structures.
Accuracy	Correct spelling and grammar.



	Raters						Ratings	
	R1	R2	R3	R4	R5	R6	R'_1	R'_2
C001-C038	X		X				R'_1	R'_2
C039-C077		X			X		R'_1	R'_2
C078-C116				X		X	R'_1	R'_2

Fig. 1. Raters/rating design.

multifaceted nature of the measurement error (Brennan, 2001).

Unidimensionality of the data: The dimensionality of the data is an essential component of any construct validity study. Summing scores across analytical domain to establish an indicator of writing competence is only acceptable if their internal structure is unidimensional. Furthermore, the unidimensionality of the data is a basic condition for many item response models (Linacre, 1989). We assess the unidimensionality of each data set using a confirmatory approach for ordinal data. Robust weighted least squares (WLSMV) estimators were used as implemented in the R lavaan package (Rosseel, 2012). Model fit was assessed by the Chi-square statistic, the RMSEA and the Comparative Fit Index (CFI). RMSEA less than.08 and CFI more than.90 indicate acceptable fit (Hu & Bentler, 1999). In order to gain more information about the internal structure of the data and taking into account the nature of the tasks, two competitive models were fit to the data: (a) the unidimensional model, and (b) the unidimensional model with correlated errors. The correlated measurement error represents systematic rather than random measurement error due to the similarity between the analytic domains used to score the essay. The path of the two models for the rater mean and tertium quid are drawn in Fig. 2.

Classification accuracy and consistency. Classification consistency is defined as the degree to which examinees would be classified into the same performance categories over parallel replications of the same assessment (Lee, 2010). Classification accuracy is the degree to which observed classification would agree with “true’’ classifications assuming known cut-scores on a single assessment. In the estimation of these indexes, we used Rudner’s approach based on the Item Response Theory (Rudner, 2001, 2005; Zhang, 2010).

4. Results

4.1. Descriptives

Of the total number of essays (N = 626), 123 passed to the second stage of scoring; that is to say, 19.64% of the essay generated discrepancy among raters in the pass/fail decision. The arithmetic mean of each of the data sets in the complete sample varied from 8.71 to 8.83, with the tertium quid giving the smallest mean and standard deviation values. The repeated measures ANOVA showed significant impact of the scoring on the means (F(2,1250)= 22.31; p < .01) but the effect size was zero (η^2_G=0; Olejnik & Algina, 2003). The univariate values of skewness and kurtosis for each score resolution procedure were acceptable in terms of normality of the scores (Gravetter & Wallnau, 2014). That is, the scoring procedure does not alter the group descriptive results of this writing competence test.

In the data sets containing discrepant data the arithmetic mean values ranged from 9.16 to 9.74, the highest value being for the parity scoring procedure. The differences among scoring procedures were significant although the effect was close to 0 (F(2244)= 25.96; p < .01; η^2_G= 0.08). The univariate values of skewness and kurtosis for each score resolution procedure were acceptable in all the cases to assume normal univariate distribution (Table 2; Fig. 2). Although these results do not guarantee the validity of the scores, they are important in terms of face validity of the writing assessment test. Fig. 3.

4.2. Inter-rater agreement

Consistency. In terms of pass/fail decisions, the rater mean scoring procedure generated the highest passing rate (23.32%), whereas the tertium quid produced the least percentage of candidates passing the test (20.44%). The results were replicated in the subsample of discrepant data (see Table 2), where the difference among percentages was higher in favour of the rater mean procedure (39.83%) and to the detriment of the tertium quid resolution method (25.20%). These results are consistent with the expected ones, since in terms of score differences among raters, the largest are related to the rater mean procedure and the smallest are associated with the tertium quid, where the discrepant value is excluded from the final computation. From this partial analysis, we conclude that the third rater (adjudicator) generates a double effect; on the one hand, the score distribution is reduced (smaller SD values) and on the other, the arithmetic means decrease. The adjudicators tend to correct the scores downwards.(Table 3).

The information was completed with the rank correlations among scoring procedures. The correlation values were significant for the complete data set (0.93,0.96 and.99). In the subsample of discrepant data, the correlation values were not significant between the rater mean and the tertium quid procedures (r = 0.10; p = .26), showing important differences between candidates rank order. As the parity scoring procedure included data from the three ratings, the correlation with the rater mean and tertium quid were significant (r = 0.81, r = 0.57; p < .01).

Reliability. The coefficient alpha for the entire sample ranged from .73 and .81. These values decreased in the subsample of discrepant data. Among the three subsamples the tertium quid discrepancy resolution procedure had the highest alpha coefficient

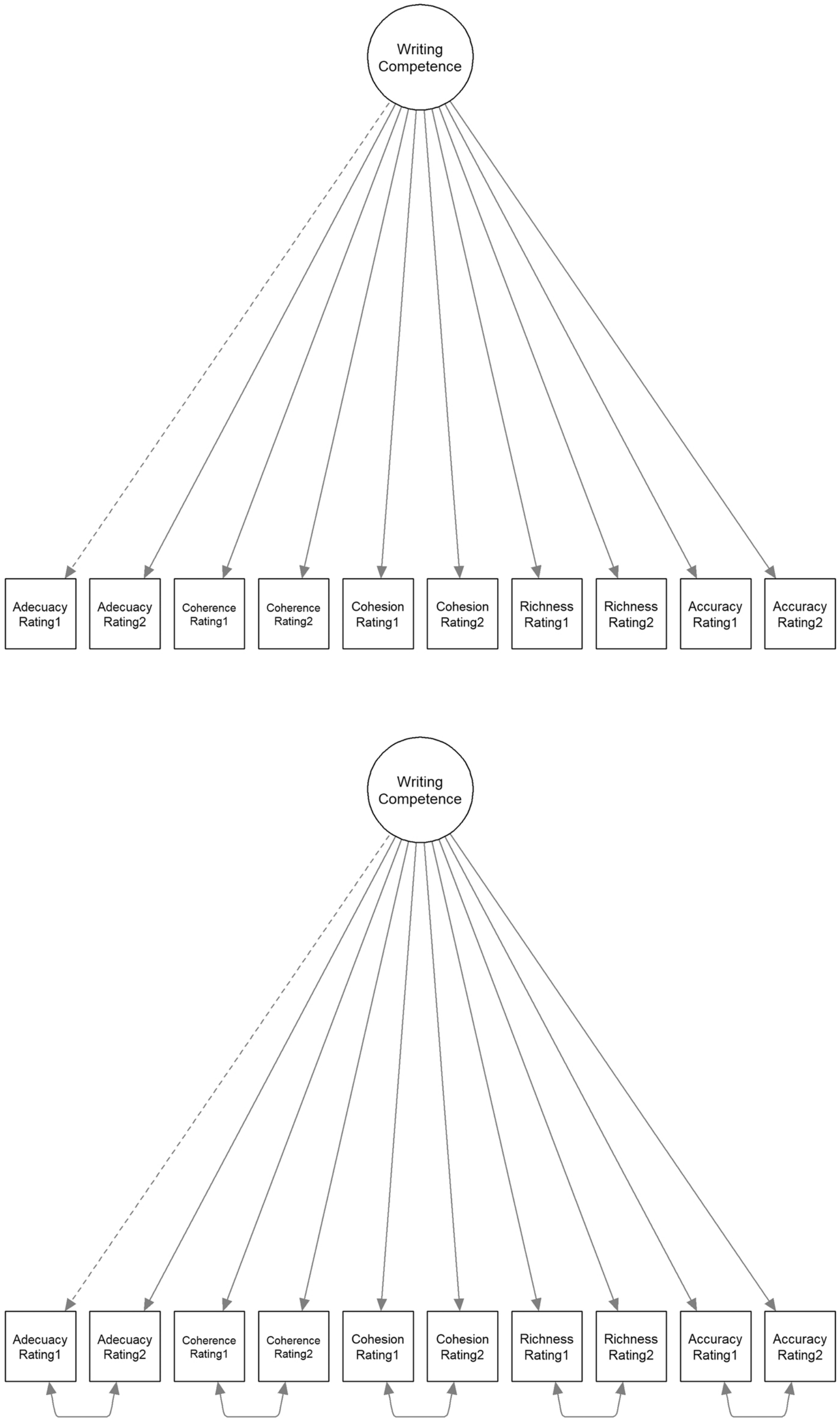


Fig. 2. Unidimensional models Note: The models for the parity scoring are similar; instead of 10 variables they include 15: 5 evaluation criteria by 3 raters.

(0.62); for the rater mean and parity procedures, the alpha coefficients were.0 and.15, showing the lack of consistency among data.

The dependability values derived from the generalizability theory are displayed in Table 4. The phi coefficients of dependability were 0.72, 0.70 and.71 for the entire sample data sets. Analysing only the subsample containing discrepant data, the results showed the

Table 2
Descriptive statistics by scoring procedure.

Scoring procedure		Mean	SD	Skew	Kurtosis	Pass	%Pass
Rater mean (S1)	Complete data	8.82	1.56	0.89	1.04	146	23.32
	Subsample	9.74	0.78	0.94	0.86	49	39.83
Parity (S2)	Complete data	8.78	1.53	0.95	1.34	133	21.24
	Subsample	9.53	0.65	0.41	-0.43	36	29.68
Tertium quid (S3)	Complete data	8.71	1.53	1.10	1.55	128	20.44
	Subsample	9.16	0.98	0.82	0.19	31	25.20

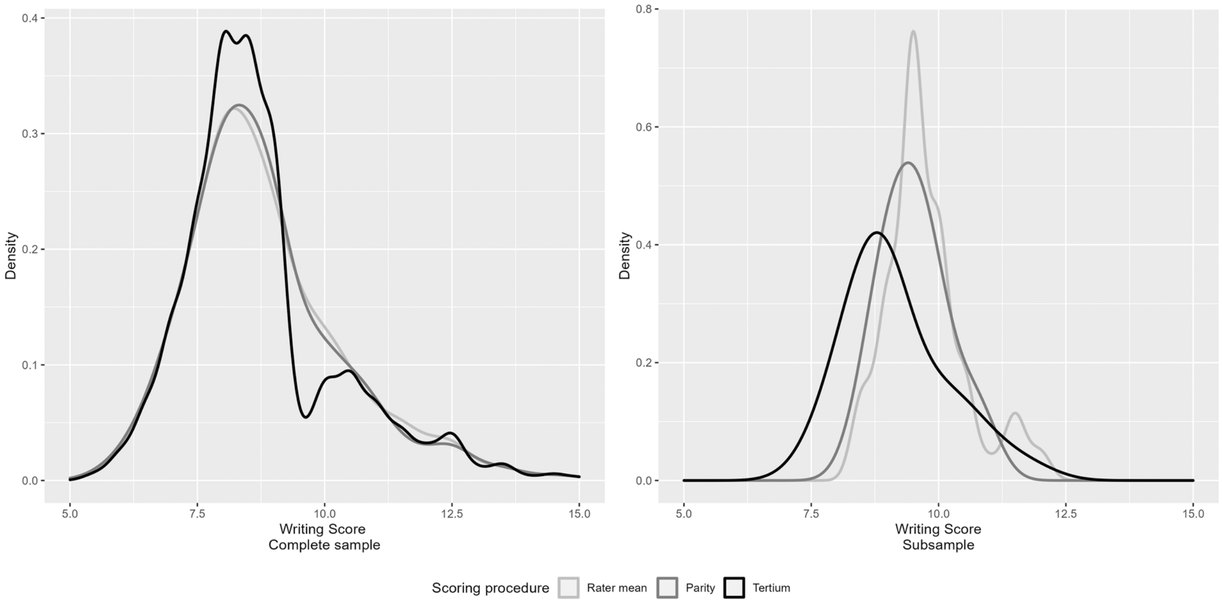


Fig. 3. Density plots for operational scoring methods.

Table 3
Rank order correlations among scoring procedures.

	Parity	Tertium quid
Rater mean	0.99*(.81)*	.93*(.10)
Parity		0.96*(.57)*

Note. Subsample correlation values in parenthesis
[*] p < .01

higher consistency for the tertium quid subsample data (Φ_{S3}=0.73) comparing the rater mean and parity resolution methods whose dependability coefficients were very low (Φ_{S1}.=27, Φ_{S2} =0.29). That is, in terms of variance components, in the tertium quid discrepancy resolution method the variance attributable to persons was 73.4%, whereas a smaller percentage of the variance, 11.1%, was explained by the five analytical scoring domains. Those percentages were 27.77% and 15.1% for the parity model and 70% and 10.1% for the rater mean scoring method.(Tables 5 and 6).

Table 4
Reliability, dependability and AVE by scoring procedure.

Scoring procedure		α	Φ	AVE
Rater mean	Complete data	0.79	0.72	0.45
	Subsample	0	0.27	0.48
Parity	Complete data	0.73	0.70	0.32
	Subsample	0.15	0.29	0.36
Tertium quid	Complete data	0.81	0.71	0.46
	Subsample	0.62	0.73	0.30

4.3. Unidimensionality analysis

The results for the complete data clearly show the improvement of the fit values when we allow correlated errors between the same correction domains. In the first scoring procedure, rater mean, the CFI increased from.94 to.95 and the RMEA and χ^2 decreased. The same behaviour was observed for the parity scoring procedure and for the tertium quid method. Among the three procedures the best indexes are related with the procedures which involve only two raters: the rater mean and the tertium quid. For the rater mean the CFI was.95, and the RMSEA and SRMR were.10; the CFI for the tertium was high (CFI=0.99) and the RMSEA and SRMR were both below the cut point of.08 (RMSEA=0.04; SRMR=0.05).

By analysing only the subsample of discrepant data, the results showed a better fit for the rater mean scoring procedure (CFI=0.98; RMSEA=0.06; SRMR=0.11) than for the tertium quid resolution method (CFI=0.84; RMSEA=0.12; SRMR=0.11). In terms of average explained variance (Table 4; AVE; Fornell & Larcker, 1981), the values were.48 and.30 for the two procedures. The first one is close to the accepted cut point of.50, and for the tertium quid the value moves away from that point. Given the percentage of discrepancy data in the entire sample, the impact of the resolution procedures on the unidimensionality was not significant. However, by analysing the subsample of discrepant date, better performance of the rater mean procedure over the tertium quid can be concluded.

4.4. Classification accuracy and consistency

The accuracy indexes for the complete sample ranged from.91 to.94 for the three scoring procedures. That means that a randomly selected candidate would be correctly classified 91%, 92% or 94% of the cases. The consistency indexes were 0.87, 0.89 and .91.

For the subsample of discrepant data, the results were lower. The number of items (domains) remains the same, 10 for the rater mean and tertium quid model and 15 for the parity scoring procedure, but the number of candidates decreases from 626 to 123. In these conditions the accuracy values were 0.58, 0.67 and.88, and the consistency of the classification was 0.52, 0.58 and .83. That is, the procedure related with the highest decision accuracy and consistency was the tertium quid.

5. Discussion and conclusions

Scoring procedure is a key topic in writing assessment, and its analysis implies a joint study of score reliability and validity. Although reliability is usually identified with measurement errors and validity with systematic errors, that is errors affecting the meaning of the scores (De Boeck & Elosua, 2016), there are many trade-offs between the concepts and therefore it makes no sense to strictly separate one from the other (Clifton, 2020; Weir, 2005). The study of scoring procedure is part of any test system validation process and it would be worthwhile to go beyond the inter-rater agreement analysis by providing information about the construct to be measured, classification accuracy and rater behaviour.

This paper set out to analyse the effect of different scoring procedures on score quality by gathering validity evidences of score unidimensionality, candidates' rank order, or classification accuracy in order to justify the use of the scores as indicators of writing competence for a pass/fail decision. In terms of validity arguments (Kane, 2006), we focus on inferences for evaluation and, generalization (Knoch & Chapelle, 2018). The evaluation inference related warrants say that the scale properties are as intended and that raters rate reliably at task level. We analyse those warrants using confirmatory factor analysis and studying rater consistency. The generalization inference regarding rating is constructed on the warrant that different raters assign the same rating to responses. We analyse this warrant by reviewing the methods of score resolution and studying the rater consistency.

To offer a better picture of the impact of the scoring and resolution procedures our study design included two groups of data: data from the entire sample and data from the subsample who received a second round of ratings. For these two samples we compare the evidences related to the rater mean scoring procedure, the parity resolution method and the tertium quid method for discrepant scores. In terms of descriptive analysis the percentage of essays which passed to the second correction round was 19.64%. To put the results in context it is important to note that the study is based on a non-complete rating design with non-connectivity among raters. The raters are paired up and work together in blocks with non-connection among them. The lack of connectivity among raters-pairs makes it

Table 5
Unidimensionality models by scoring procedure.

Scoring procedure			χ^2	d.f.	CFI	RMSEA	SRMR
Rater mean	Complete data	M1	346.76	35	0.94	0.11	0.11
		M1CE	266.27	30	0.95	0.10	0.10
	Subsample	M1S	54.52	35	0.97	0.06	0.11
		M1SCE	44.15	30	0.98	0.06	0.11
Parity	Complete data	M1	562.89	90	0.91	0.09	0.20
		M1CE	455.67	75	0.93	0.09	0.20
	Subsample	M1S	268.75	90	0.83	0.12	0.18
		M1SCE	152.94	75	0.92	0.09	0.18
Tertium quid	Complete data	M1	192.97	35	0.97	0.08	0.08
		M1CE	68.38	30	0.99	0.04	0.05
	Subsample	M1S	142.01	35	0.72	0.15	0.13
		M1SCE	90.65	30	0.84	0.12	0.11

Notes: Model: M1 = Complete data; M1S=Subsample; M1CE: Model with correlated errors; M1SCE= Subsample with correlated errors

Table 6
Classification accuracy and consistency.

Scoring procedure		Accuracy	Consistency
Rater mean	Complete data	0.91	0.87
	Subsample	0.58	0.52
Parity	Complete data	0.92	0.89
	Subsample	0.67	0.58
Tertium quid	Complete data	0.94	0.91
	Subsample	0.88	0.83

inadvisable to obtain individual performance measures as many-facets models do. To overcome this design shortcoming, we worked on rating rather than raters, since this approach allowed us to carry out many analyses based on classical test theory, generalizability, and item response theory. It is also true that more investigation is needed to analyse the assumption that raters are random variables without significant differences among them and to study the impact of the violation of that assumption on results. We would like also to mention that for this research paper discrepancy has been defined as differences in the pass/fail decision. We have not studied any other kind of discrepancies such as differences in scores; this decision is based on the design of the test analysed, which defines the discrepancy between raters only at the pass/fail level.

One of the main result of this research shows no significance differences in the total sample descriptives among the scoring procedures; but as expected, in the subsample of discrepant data, the rater mean procedure generated the highest arithmetic mean of the scores. In terms of inter-rater agreement, we worked with the Spearman rank-order correlation. The coefficients estimated for the entire sample were higher than.90 for the three correlations. But the correlation values estimated in the subsamples showed discrepant results. The correlation between the rater mean procedure and the parity model was high ($r = 0.81$), but the rank order between the rater mean and the tertium quid was not significant ($r = 0.10$). This result was expected since the parity model values are constructed on the two first raters' scores but the tertium quid only include one of the original scores.

In terms of consistency among analytical domains, the highest alpha coefficient was for the tertium quid scoring procedure ($\alpha = 0.81$), and the lowest value was associated with the parity model ($\alpha = 0.73$). The dependability coefficients showed a similar pattern, with the highest value in the tertium quid ($\Phi = 0.86$). These results are in line with previous research; it is important to note that at this point some authors alert us that the tertium quid score procedure can introduce an artificial inflation of the reliability estimates (Johnson et al., 2000).

The confirmatory factor analysis allows us to study the quality of the writing score as indicator of a latent construct composed of five domains: adequacy, coherence, cohesion, richness and accuracy. The fit indexes for the entire sample were good for the rater mean and for the tertium quid procedures, showing a valid argument to consider that the sum score across domains is a good indicator of the writing assessment. As expected, the error-correlated models improved the fit of the original models. The parity model did not perform well in terms of fit indexes. Although the alpha coefficient is not an unidimensionality index, we found a similar pattern for both indicators. The two-rater methods performed better in terms of unidimensionality than the scoring that included three raters. The results were different in the analysis of the subsample of discrepant data. The tertium quid discrepancy resolution method performed worse than the rater mean operational scoring procedure. The fit indexes of the models and the percentages of extracted common variance showed problems related to the quality of the measure defined through this discrepancy resolution method.

Regarding the classification accuracy, while there are not many papers that address this issue in writing assessment (Zhang, 2010), the three scoring procedures clearly perform very well with accuracy values above.90. According to the standards for test use (AERA, APA & NCME, 2014), it is not appropriate to dictate a minimal level of classification accuracy, but Subkoviak (1988) points out the importance of guaranteeing an agreement coefficient exceeding.85. If the focus of the analysis is the subsample of discrepant data, the accuracy values decreased, the tertium quid procedure being the one keeping the best indicators. Classification accuracy can be seen as part of the evaluation inference, since the assumption about the raters' ability to identify differences in performance levels implies that decisions are accurate.

At this point, and following the design of the study, it is important to differentiate between the two blocks of data defined at the beginning of this work. The total sample shows that the impact of the tertium quid and parity models can be read as similar in terms of reliability (alpha, phi, rank correlation), internal validity (AVE, CFA models), and classification accuracy. The values are slightly better for the tertium quid but the indicators for the rater mean scoring procedure are also psychometrically good. If we focus only on the subsample of discrepant data the results are unclear since they do not always go in the same direction. Traditional reliability values (alpha, phi), rank order correlations, and classification accuracy support the use of the tertium quid discrepancy resolution method over the parity and rater mean scoring procedure. However, the internal structure of data for this subsample does not fit well to the unidimensional model, which could justify summing scores across domains as indicator of writing competence. As an explanation for these results it could be said that in the tertium quid design the adjudicator is an expert whose decision plays a vital role in determining the pass/fail decision, and he/she knows his/her role in the assessment process. However, this circumstance could be controlled using a rating design in which the adjudicator is not aware of his/her role and so the design could be to keep the statistical condition of independence between raters.

The goal of any score resolution procedure is to increase the reliability of the scores without introducing any bias or systematic error which could affect their validity. In the balance among those key concepts there are many trade-offs that the researcher must consider in favour of a psychometrically and ethically strong assessment. The studies comparing scoring procedures are not categorical since many aspects must be evaluated, and some of them are not strictly psychometric. As Penny and Johnson (2011) point out, the application of the parity model in high-stakes programs could be problematic because combining discrepant scores may not be acceptable for policy makers; that is, parity and rater means could be associated with lack of face validity; the use of experts' judgments as in the tertium quid procedure might also introduce systematic errors in the scoring process, which needs to be evaluated and controlled. In any case, it is necessary to include operational scoring designs and discrepancy resolution methods in the assessment of writing tasks to help assure the standards of scientific and ethic quality in an activity that impacts candidates from their school years though their professional life.

Declaration of interests

none.

Data Availability

The data that has been used is confidential.

Acknowledgements

Thanks to Basque Institute for Public Administration and the research project PID2019–103859RB-I00.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257. https://doi.org/10.1177/026553229501200206

Brennan, R.L. (1996). Generalizability of performance assessments. In G. W.Phillips (Ed.), Technical issues in large-scale performance assessments (pp.19–58). Washington, DC: U.S. Department of Education.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1079–1097). Chichester: Wiley. https://doi.org/10.1002/9781118411360.wbcla110.

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods, 25*(3), 259–270. https://doi.org/10.1037/met0000236

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching and Assessment. Strasburg.

De Boeck, P., & Elosua, P. (2016). Reliability and Validity: History, Notions, Methods, Discussion. In F. T. L. Leong, D. Bartram, F. Cheung, & K. Geisinger (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 408–421). New York: Oxford University Press.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, 32*(4), 521–541. https://doi.org/10.1177/0265532215575626

Dunsmuir, S., Kyriacou, M., Batuwitage, S., Hinson, E., Ingram, V., & O' Sullivan, S. (2015). An evaluation of the Writing Assessment Measure (WAM) for children's narrative writing. *Assessing Writing, 23*, 1–18. https://doi.org/10.1016/j.asw.2014.08.001

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. New York: Peter Lang,.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement errors. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312

Gravetter, F., & Wallnau, L. (2014). Essentials of Statistics for the Behavioral Sciences. *CA* (8th ed.). Belmont: Wadsworth.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*, 123–139. https://doi.org/10.1016/j.asw.2011.12.003

Johnson, R., Penny, J., & Gordon, B. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*, 121–138. https://doi.org/10.1207/S15324818AME1302_1

Johnson, R., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication, 18*, 229–249. https://doi.org/10.1177/0741088301018002003

Johnson, R., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education, 16*, 299–322. https://doi.org/10.1207/S15324818AME1604_3

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M.T. (2006). Validation. In R. Brennen (Ed.), Educational measurement, 4th ed. (pp. 17–64). Westport, CT: Greenwood.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000

Kim, B. (2011). Resolving discrepant ratings in writing assessments: The choice of resolution method and its application. *English Teaching, 66*(2), 211–230.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes with an argument-based framework. *Language Testing, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Knoch, U., & Macqueen, S. (2020). Assessing English for professional purposes. Abingdon: Routledge.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*, 1–17. https://www.jstor.org/stable/25651533.

Lee, Y.W. (2005). Dependability of Scores for a New ESL Speaking Test: Evaluating Prototype Tasks. ETS Monograph Series, 28. ETS.

Lin, C. (2017). Working with sparse data in rated language tests: Generalizability theory applications. *Language Testing, 34*(2), 271–289. https://doi.org/10.1177/0265532216638890

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Marcoulides, G. A. (1989). Performance appraisal: Issues of validity. *Performance Improvement Quarterly, 2*, 3–12.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. Urbana, IL: National Council of Teachers of English.

Ohta, R., Plakans, L., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing, 38*, 21–36. https://doi.org/10.1016/j.asw.2018.08.001

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Penny, J. A., & Johnson, R. K. (2011). The accuracy of performance task scores after resolution of rater disagreement: A monte Carlo study. *Assessing Writing, 16*(221), 236. https://doi.org/10.1016/j.asw.2011.06.001

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Rudner, L.M. (2001). Computing the expected proportions of misclassified examinees. Practical Assessment Research & Evaluation, 7(14). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=14.

Rudner, L.M. (2005). Expected classification accuracy. Practical Assessment Research & Evaluation,10(13). Retrieved from http://pareonline.net/pdf/v10n13.pd.

Stemler, S. E. (2004). A Comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Pract. Assesm. Research, and Evaluation, 9*(*4*). https://doi.org/10.7275/96jp-xz07

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*, 47–55. https://doi.org/10.1111/j.1745-3984.1988.tb00290.x

Weir, C. (2005). Language testing and validation. New York: Palgrave Macmillan.

Wind, S. A., & Walker, A. A. (2019). Exploring the correspondence between traditional score resolution methods and person fit indices in rater-mediated writing assessments. *Assessing Writing, 39*, 25–38. https://doi.org/10.1016/j.asw.2018.12.002

Wind, S. A., & Walker, A. A. (2020). Exploring the impacts of different score resolution procedures on person fit and estimated achievement in rater-mediated assessments. *Language Assessment Quarterly, 17*(4), 362–385. https://doi.org/10.1080/15434303.2020.1783668

Wolcott, W. (1998). An overview of writing assessment: Theory, research, and practice. Urbana, IL: National Council of Teachers of English.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*(1), 119–140. https://doi.org/10.1177/0265532209347363

Paula Elosua Ph.D. in Psychology and professor of psychometrics at the University of the Basque Country. She heads a psychometrics research group specialising in language, educational and psychological assessment, test and questionnaire construction/adaptation, and validation studies.