A comparison of holistic, analytic, and part marking models in speaking assessment

Journal:	*Language Testing*
Manuscript ID	LT-19-0034.R3
Manuscript Type:	Original Manuscript
Keywords:	marking models, holistic, analytic, marking by part, Many-Facet Rasch Measurement (MFRM), speaking assessment
Abstract:	This mixed methods study examined holistic, analytic, and part marking models (MMs) in terms of their measurement properties and impact on candidate CEFR classifications in a semi-direct online speaking test. Speaking performances of 240 candidates were first marked holistically and by part (phase 1). On the basis of phase 1 findings – which suggested stronger measurement properties for the part MM – phase 2 focused on a comparison of part and analytic MMs. Speaking performances of 400 candidates were rated analytically and by part during that phase. Raters provided open comments on their marking experiences. Results suggested a significant impact of MM; approximately 30% and 50% of candidates in phases 1 and 2 respectively were awarded different (adjacent) CEFR levels depending on the choice of MM used to assign scores. There was a trend of higher CEFR levels with the holistic MM and lower CEFR levels with the part MM. While strong correlations were found between all pairings of MMs, further analyses revealed important differences. The part MM was shown to display superior measurement qualities particularly in allowing raters to make finer distinctions between different speaking ability levels. These findings have implications for the scoring validity of speaking tests.

SCHOLARONE™
Manuscripts

A comparison of holistic, analytic, and part marking models in speaking assessment

Background and rationale

This study examines the impact of different marking models (MMs) – the approaches used for assigning ratings to performances – on candidate scores in the context of a semi-direct online speaking test. The test was originally developed to be part of a Cambridge Assessment English placement test battery as a quick measure of candidates' ability to speak in a variety of general everyday contexts. It elicits a range of language features and functions through four task types: interview questions about the candidates and their background, a description and comparison of two photographs, questions related to a scenario, and a one-minute monologue on an abstract topic (see Appendix A for further information). The test is designed to progressively increase in difficulty along some of the features identified in Robinson's (2001) framework of task complexity; e.g. more familiar/here-and-now topics in the initial parts of the test and more abstract, open-ended topics in the final parts. While there is progression in task difficulty, each task provides scope for performances at a range of CEFR levels. The test is rated holistically, i.e. a general overall evaluation of performance is given. The approach involves a rater assigning a single score to the candidate's performance on the whole test based on a balanced consideration of four criteria – *coherence and discourse management*, *language resource*, *pronunciation*, and *hesitation and extent*. This score is based on a six-level rating scale covering levels A1 to C2 on the Common European Framework of Reference (CEFR) (Council of Europe, 2001) – see Appendix B.

The advantages of holistic scoring are considered to be its efficiency, ease of reporting (Davies et al., 1999), and lower cognitive demand on raters (Xi, 2007). In our assessment context, the use of holistic marking, which is relatively quick, reflects the operational need for a placement test with a short results turnaround. However, in light of potential future uses of the test including the provision of diagnostic information to candidates, there was a need to empirically investigate alternative MMs that would allow the generation of more fine-grained information on speaking performance, maintain the test's scoring validity (Weir, 2005), and meet the practical demand for quick marking.

Choice of MM has been shown to affect rater marks (Barkaoui, 2011; Schoonen, 2005) and is thus an important aspect of a test's validity argument. There is, however, little empirical research on the impact of different MMs on scores, particularly in speaking assessment. While choice of MM might be context-dependent and related to test purpose, research into how models compare can be valuable in informing decisions on scoring approaches.

The rationale for our study is, therefore, twofold: to inform the operational validation of a speaking test, and to contribute to building theory around the relative strengths and limitations of different MMs based on empirical evidence.

Marking models

In this paper, we use the term marking model to refer to "methods [used] to form judgments" on a performance (Harsch & Martin, 2013, p.281). While the term can extend to automated approaches to marking, we will be narrowing our focus to human-mediated MMs. Our definition does not include methods used for ensuring or evaluating judgment quality, such as double marking or different measurement techniques, such as G-theory.

A review of the literature suggests holistic and analytic scoring as the most widely used human-mediated MMs in writing and speaking assessment, with definitions and discussions of their strengths and limitations extensively documented (see for example Davis, 2018; Hamp-Lyons, 1995; Lee, Gentile, & Kantor, 2009; Weigle, 2002). Part scoring is another MM that is operationally used by several international speaking tests such as *TOEFL*® and BULATS, as well as local tests such as the Oral English Proficiency Test for prospective international teaching assistants at Purdue University (Ginther, Dimova, & Yang, 2010; Yan, 2014). In contrast to the research on holistic and analytic MMs, empirical discussions of part scoring are less widely available in the assessment literature.

In the following sections, we discuss these three MMs in more detail.

Holistic

The practicality of holistic scoring is often seen as its main strength; theoretical arguments have also drawn attention to its suitability for the assessment of overall communicativeness (Weir, 1990), for representing "integrated higher-order skills" (Hunter, Jones, & Randhawa, 1996, p. 64), and as an antidote to "analytic reductionism" (White, 1984, p. 406). There are, however, limitations to the holistic approach, including limited transparency in the relative weighting of features which may be differentially applied by raters (Brown, 1995; Xi, 2007), the potential for raters to focus predominantly on what candidates can do well rather than on areas of weakness (Bacha, 2001), and an underlying assumption in holistic scoring that different features of performance develop at the same rate (Kroll, 1990), which is questionable from a second language acquisition point of view.

Analytic

Analytic scoring involves assigning separate scores to explicitly defined criteria (or dimensions) related to different aspects of performance (Davis, 2018; Xi, 2007). A strength of analytic MMs is that a multi-dimensional scale allows for a more systematic evaluation process where the criteria/dimensions and their relative weightings can be made explicit (Xi, 2007), which in turn provides raters with more clarity about the focus of

ratings, thus potentially increasing reliability (Goulden, 1994). The collection of a number of observations through analytic scoring is a further positive feature, since greater reliability can be achieved through multiple observations (Lee, 2006; Barkaoui, 2011).

A multi-dimensional scale, moreover, reflects the complex nature of language and is therefore more in line with theoretical models of communicative language ability (Bachman & Savignon, 1986). Analytic scoring can reveal differences in strengths and weaknesses of performances as learners go through developmental stages (Hamp-Lyons, 1995) and as such, better suits candidates with uneven profiles (Weigle, 2002). It can also offer diagnostic information to inform individual learning paths (Bacha, 2001).

The analytic MM is not without its limitations. Lee et al. (2009), for example, draw on high correlations between different analytic dimensions and holistic scores, to suggest that analytic scores may be psychometrically redundant. Other limitations include increased cognitive demand on raters (Underhill, 1987), difficulties in precisely defining scoring criteria (Douglas & Smith, 1997), and the need for rigorous rater training in reliably distinguishing between criteria (Xi, 2007).

Part

Part scoring involves assigning separate scores to different test parts on the premise that a single score covering a candidate's overall performance on a number of tasks may not be an accurate representation of language ability and "might be overly influenced by either good or poor performance on a particular task" (Nakatsuhara, 2011, p. 36). Similar to the analytic MM, the collection of multiple marks has the potential to enhance reliability. Lee (2006), for example, reported a "large impact" of increasing the number of tasks on score dependability in the *TOEFL*® speaking test. Another advantage is that ratings can be awarded by a single rater or by multiple raters each scoring different parts of a candidate's performance. A limitation is the shorter language samples elicited in each test part, which may not provide enough language for a valid and reliable evaluation. While this can be countered with longer tests with multiple tasks of adequate length for rating, this may not be always be possible due to practical constraints.

Part marking can be used in holistic and analytic models, although it may be too practically cumbersome and cognitively demanding for raters to assign a full range of analytic marks to each test part (Taylor & Galaczi, 2011). To the best of our knowledge, the speaking tests mentioned earlier that use part marking involve raters assigning holistic scores to each part. Nevertheless, this MM lends itself to different configurations, such as different criteria for different test parts and/or some parts marked holistically and others analytically.

Relationships between MMs

In this section, we look at empirical research on the relationships between MMs, drawing from studies on both writing and speaking assessment. It is

worth emphasising that while most of this research focuses on measurement properties of different MMs, there are broader conceptual issues at play. MMs are representations of what is considered important in performance. For example, a part MM attributes more importance to variation in performance across tasks and aligns more closely to task-centred approaches to construct definition (Norris et al., 2002, Brown et al., 2002). An analytic MM, on the other hand, places the primary focus on performance against different linguistic criteria; an approach that has closer affinity with trait-based approaches to construct definition (Chapelle, 1998). A full discussion of these different paradigms is beyond the scope of this paper but it is important that they are taken into account when considering different MMs.

In the context of investigating the relationship between holistic and analytic scales, Bacha (2001) used a stratified sample of essays by L1 Arabic students of English and found high inter-rater agreement and strong correlations between the different analytic criteria, as well as the holistic and analytic scores. Further analyses, however, showed that students' performance on the different analytic criteria were statistically distinct. The diagnostic element of the analytic MM was seen as more "informative" for learning (Bacha, 2001, p. 381).

The effects of holistic and analytic MMs on writing were further investigated by Barkaoui (2011) and Wiseman (2012). Barkaoui's (2011) data consisted of essays on two prompts written by adult learners of English from three proficiency levels. Findings from the Many-Facet Rasch Measurement (MFRM) analysis showed comparable ability estimates for the candidates across the MMs, suggesting that the two MMs "may be regarded as measuring the same underlying construct from a measurement point of view" (Barkaoui, 2011, p. 285). Nevertheless, observed differences between the two approaches indicated that the analytic MM resulted in lower standard errors for candidate ability estimates, separation of candidates into more statistically distinct levels, a higher proportion of candidates with acceptable infit[1] values, and increased rater leniency. Interestingly, inter-rater agreement was lower with the analytic approach, which the author attributed to this model better capturing the raters' "diversity of opinions and values" (Barkaoui, 2011, p. 288). In contrast, the analytic scale in Wiseman (2012) was associated with increased severity for raters. In both studies, the analytic approach was shown to display "greater measurement precision" (Barkaoui, 2011, p. 287) and to be more "sensitive" to differences in candidates' writing abilities (Wiseman, 2012, p.169). As highlighted by Barkaoui (2011), in an analytic MM, there are multiple observations for each candidate – as

---

[1] The term "infit" stands for "information weighted fit statistic"; it is one of the fit indices produced by FACETS (Linacre, 2018a) which "enable diagnosis of aberrant observations and idiosyncratic elements" (Linacre, 2018b, p.14). This index has an expected value of 1 and a range from 0 to infinity with values between 0.5 and 1.5 generally considered "productive for measurement" (Linacre, 2018b, p.278).

opposed to a single observation with a holistic MM – with this additional information likely to contribute to increased measurement precision.

An additional perspective on the relationship between MMs was offered by Harsch and Martin (2013). In this study, raters applied the same rating scale to a sample of scripts in three increasingly fine-grained methods: firstly, scores were assigned holistically to an overall performance; secondly, scores were awarded for each of the four criteria in the scale; and lastly, scores were assigned with reference to the descriptors within each of the criteria. Findings showed that raters' levels of agreement decreased as the scale granularity increased, suggesting that the two more holistic approaches "masked deviances in how the raters applied the descriptors defining a criterion" (Harsch & Martin, 2013, p. 296).

In one of the few studies focusing on MMs in speaking, Xi (2007) explored the viability of analytic scoring in a large-scale holistically marked speaking test. Descriptors were extracted from the holistic scale to create three separate analytic scales, and performances covering a range of proficiency levels and L1s were subsequently marked by raters. Scores were awarded holistically to each task and then averaged to calculate an overall holistic score. The same performances were also marked analytically for each task. Findings showed high correlations between scores on different analytic criteria, which were taken to suggest that, from a psychometric perspective, the different dimensions were not sufficiently distinct from one another. Results also indicated varied profiles at the individual task level in some cases, but these profiles were flattened once scores were averaged across tasks, leading the author to conclude that analytic scores "would not provide additional information beyond what the holistic scores could offer for most examinees" (Xi, 2007, p. 281).

Also in the context of speaking assessment, Nakatsuhara (2011) focused on part marking in the IELTS Speaking test, and found differences in candidate scores on two of the test parts. While these differences did not reach statistical significance, Nakatsuhara (2011) argued that results provided empirical support for a part MM. Similar to Xi (2007), the part scores were calculated by awarding analytic scores on the four IELTS criteria for each test part and subsequently aggregating the scores; an approach which, as Nakatsuhara cautions, may prove impractical in operational settings and result in "increased burden on the examiners" (2011, p. 36).

Key issues from the literature

The following points emerge from our review of the literature:

Firstly, while research suggested a trend of awarding comparable scores across holistic and analytic MMs, differences also emerged. For example, in Barkaoui (2011), candidates tended to be awarded higher scores with the analytic MM, whereas the opposite was observed in Wiseman (2012). Findings also diverged on the extent to which performances on different criteria were sufficiently distinct to warrant an analytic MM. A possible explanation is the

5

differences in methodology and scales. For example, while Harsch and Martin (2013) and Xi (2007) used the *same* scales/descriptors and applied them in different ways, other studies such as Wiseman (2012) applied *different* rating scales/descriptors in their comparisons. We also found terminological inconsistencies which may have contributed to these contradictory results. For example, what Harsch and Martin (2013) refer to as "holistic-criterion" scoring is seen as analytic scoring by others (Barkaoui, 2011; Wiseman, 2012). Similarly, Xi's (2007) holistic and analytic scoring is what O'Sullivan and Nakatsuhara (2011) would list as part-holistic and part-analytic scoring respectively.

Secondly, in reporting on the empirical relationships between different MMs, most studies have relied on correlations; however, this may disguise the effect of MMs on individual candidate marks. It is therefore essential to consider alternative ways of examining MMs, with an explicit focus on their impact on candidate final scores and classifications; in other words, their "practical significance" (Fulcher, 2003, p. 65). What constitutes practical significance is context-dependent. Given the use of the CEFR for score reporting and decision-making, practical significance in often defined in terms of CEFR levels, i.e. cases where candidates receive higher/lower CEFR levels as a result of the MM applied.

Thirdly, the majority of empirical research on MMs is in writing assessment. There is, we believe, a need to better understand the impact of MMs in speaking assessment.

Fourthly, while the part MM has been used operationally in several speaking tests, there is little reported research on the comparison between the part MM and other marking approaches. Given that speaking and writing tests are typically designed with a range of task types aimed to tap into different aspects of the construct of interest, a systematic examination of the part MM is warranted.

Research aims

Our study aims to empirically evaluate the theoretical assumption that differences between MMs can influence performance scores. We attempt to address some of the issues raised in our literature review by (a) using the same scale/set of descriptors and applying them in three different ways in order to control for the potentially confounding effects of variations in these, (b) focusing on the practical impact of MMs on candidate CEFR classifications, (c) selecting speaking assessment for the context of our study, and (d) including the part MMs in addition to holistic and analytic approaches.

Our study is guided by the following research question:

> *How do the MMs under examination – holistic whole test (henceforth holistic), holistic by part (henceforth part), and analytic whole test (henceforth analytic)– compare in terms of (a) impact on candidate scores and CEFR classifications and (b) measurement properties?*

Method

Design

This is a two-phase study comprising quantitative score data and qualitative rater comments, integrated in a concurrent mixed methods design (Creswell, 2013). We opted for a 'competition' design where, in the first phase, our speaking test's operational MM (holistic) would be compared to an alternative MM. Subject to empirical evidence, the stronger model in terms of measurement properties would be compared to our second alternative MM in the subsequent phase. To allow for a direct comparison of all three MMs simultaneously, the design included a linking of the data sets from the two phases through common raters and performances. Such an approach served as a practical and cost-effective solution to addressing the study's research question.

In phase 1, we compared the holistic and part MMs. We limited our investigation to a holistic by part MM (as opposed to an analytic by part MM), due to its likelihood for adoption in operational settings. On the basis of phase 1 results (see findings), phase 2 focused on a comparison of the part and analytic MMs. Figure 1 provides a snapshot of the study's design, data collection and analysis procedures.

[Insert Figure 1]

Participants

Four raters participated in phase 1. Given the larger data set in phase 2, an additional six raters participated in this phase. All raters (six female, four male) were L1 speakers of English, had over five years' experience teaching ESL/EFL, and were trained/certified for a number of different Cambridge Assessment English speaking tests (see Table 1). All had worked as examiners for the BULATS online speaking test, which shared important similarities with the speaking test in our study in terms of format, task types, and assessment scale descriptors.

[Insert Table 1]

All raters worked independently in order to limit the potential for collusion between them.

Speech data

Data for the study were selected from a pool of available speaking tests (approximately 2500 candidates at the time of data collection), and comprised 240 performances in phase 1 and 400 in phase 2. Each performance had an associated holistic mark, assigned by a single examiner using standard operating procedures. A stratified sampling approach was used to select performances that covered a range of ability levels and L1s. Table 2 shows the breakdown of CEFR levels by dataset according to operational holistic ratings. The distribution of the sample dataset was designed to closely approximate that of the test population at the time of

7

research. 30 different L1s were represented, with Chinese (29.5%), Arabic (15.3%), and Portuguese (11.8%) the most frequent.

[Insert Table 2]

Rating scales

The assessment scale is a six- level holistic scale covering four criteria: *coherence and discourse management*, *language resource*, *pronunciation*, and *hesitation and extent* (see Appendix B). For the purpose of our study and similar to Xi (2007), we created an analytic scale by extracting the descriptors from each of the four criteria in the holistic scale and displaying them separately. No changes were made to the content of the descriptors.

Data collection

Data collection took place in two phases with a three-month interval in between. Within each phase data collection was completed in two rounds with a one-week interval in between in a counter-balanced design to minimise any order or halo effects.

Phase 1 focused on a comparison of holistic and part MMs; a rating matrix was designed to ensure that (a) each speaking test was marked holistically and by part, (b) there was a link between MMs with each rater marking the same candidates' performances using part and holistic MMs, and (c) there was a link between raters through a common batch of performances. This design feature created a link between candidates and raters in order to meet the requirements of MFRM. The performances not in the common batch were single scored; i.e. once scored holistically and once by part. On the basis of the rating matrix, speech files were allocated to raters along with the assessment scale and detailed instructions.

Phase 2 was informed by the results of phase 1 and focused on a comparison of part and analytic MMs. Data collection for phase 2 closely followed phase 1 procedures, with MM as the main difference.

Upon completion of each phase of marking, raters were invited to provide open comments to a short questionnaire on their experiences and views of applying each pair of MMs. The questionnaire focused on raters' preferences regarding MMs, ease of marking in each model, and the feasibility of part or analytic marking in operational conditions.

Data analysis

Scores awarded by raters were analysed using MFRM with FACETS (Linacre, 2018a). MFRM provides a technical solution to the well-documented rater effect in performance assessment (McNamara, 1996) by allowing different facets of the testing situation to be measured independently and then mapped onto a common linear scale measured in "logits". Importantly, candidates' ability measures from the analysis are estimated independently of the particular rater or task assigned to them, with their raw scores adjusted

for the effects of the facets of performance. The resulting candidate fair-average mark is a more objective estimate of the candidate's ability. MFRM is also robust against missing data as long as there is enough linking between different facet elements (Linacre, 2018b). This is particularly important from a practical perspective, as a fully-crossed design may not be possible. We opted for a connected design where "a network of links exists through which every element that is involved in producing an observation is directly or indirectly connected to every other element of the same assessment context" (Eckes, 2009, p. 39). Our rating matrix was designed to ensure a linking of MMs through the same candidate performances, the linking of raters through a common batch of candidates, and finally a linking of the two phases with a common set of raters.

To address the study's overall research question, several MFRM models were examined. Firstly, separate analyses were run for each MM: a two-facet (candidate, rater) model for analysing the holistic scores, a three-facet (candidate, rater, test part) model for analysing the part scores, and a three-facet (candidate, rater, criteria) model for analysing the analytic scores. In each phase, candidates' fair-average marks were correlated and their rankings compared using a Wilcoxon Signed-Rank test in SPSS. Additionally, the fair-average marks from the MFRM analyses were converted into CEFR levels; marks were rounded down to the nearest integer given that operationally a candidate needs to meet all the descriptors in a level to be awarded that level. The percentage of candidates receiving the same CEFR classifications across MMs was then calculated. A range of statistics (as explained below) for the two pairings of MMs were compared. A three-facet (candidate, rater, MM) model was also run in each phase to allow for a direct comparison of each pair of MMs on the same logit scale. Lastly, a two-facet (candidate, MM) analysis was run, where fair-average marks from the MFRM analyses of different MMs were combined and analysed simultaneously for a direct comparison of all three MMs. In all analyses, the candidate facet was allowed to float with all other facets centred at zero.

Data analysis drew on a range of statistics that are generated in MFRM for each facet. These included parameter estimates for each facet and corresponding reliability indices, i.e. the standard error index, the separation statistics, which are useful for summarising observations and drawing inferences about group trends, and the separation indices and strata[2] which estimate the number of statistically distinguishable performance levels and their associated reliability (Linacre, 2018b). We interpreted these indices mindful of their facet-dependency; for example, a high separation index with associated high reliability is desirable for candidates and can show that the

---

[2] The choice of separation statistic to report depends on whether the extreme scores or outliers in the sample are "accidental" or whether they represent "extreme performance levels" (Wright & Masters, 2002, p. 888). In the case of this study, there are no accidental outliers and therefore (any) extreme scores relate to persons of relatively high or low speaking abilities which is why the strata statistic serves as a more accurate measure of spread of ability, difficulty, or severity.

test has successfully distinguished between different ability levels, whereas a low separation index and reliability is desirable for raters, who should be similar in measures.

Apart from group-level statistics, we considered fit statistics which "enable the diagnosis of aberrant observations and idiosyncratic elements" (Linacre, 2018b, p. 14) within each facet. Specifically relevant are the infit and outfit mean statistics, which can indicate misfit. They have an expected value of 1 and a range from zero to infinity where "the higher the (…) mean-square index, the more variability we can expect" in the rating patterns (Myford & Wolfe, 2000, p. 15). Here we only report infit, as it less sensitive to outliers compared to outfit and because it is broadly viewed as the more important statistic to be considered in evaluating fit of the data to the model (Eckes, 2009; Myford & Wolfe, 2003). Values below 1 are considered to be "overfitting" the model and too predictable, while values above 1 are considered to be "underfitting" and too unpredictable (Linacre, 2018b) with the latter generally raising more cause for concern (Eckes, 2009; Linacre, 2018b). In line with Linacre (2018b), the current study adopted lower and upper control limits of 0.5 to 1.5 for the infit mean square index. A summary of results for all MFRM models can be found in Appendix E (phase 1) and Appendix F (phase 2).

Raters' open comments were analysed for common themes and insights that could further inform the study's quantitative findings. Given the small number of raters involved and short questionnaire, the qualitative part of the project was comparatively limited. It involved the collation of rater comments in Excel and analysis for common themes which emerged from the data. The authors completed this stage collaboratively in order to ensure appropriate interpretation of data.

Findings

A comparison of holistic and part MMs

To compare the holistic and part MMs and explore their potential impact on candidate scores/CEFR classifications, we first estimated candidate ability levels from independent MFRM analyses of scores as awarded by the two MMs. This ensured that candidate measures were adjusted for the effects of raters. The holistic MM was associated with a slightly higher mean (M=3.70; SD=0.90) than the part MM (M=3.64; SD=0.83), although the results of a paired-sample *t*-test showed no statistically significant differences and a Wilcoxon Signed-Rank test indicated no significant difference in candidate rankings. We also correlated these fair-average measures and a strong statistically significant correlation emerged (*r*=0.88, *p*<0.01).

These fair-average scores were then converted to CEFR levels; results in Table 3 show a similar distribution of CEFR levels, with a slightly higher percentage of candidates at the A1/A2 levels for the part MM and generally small differences at the group level.

[Insert Table 3]

We then focused on the percentage of candidates that received the same CEFR classification across the two MMs, including the size and direction of the differences. Results showed that 68.6% of candidates were awarded the same CEFR classification regardless of the MM while 30.9% fell within an upper or lower adjacent level. Less than 1% received more than a level difference (see Appendix C for a contingency table with details of specific occasions where the two MMs converged/diverged). Candidates were likely to receive a higher CEFR level when the holistic MM was used. Possible explanations for this trend are drawn from rater feedback: one rater commented that when marking holistically he was "less likely to pay attention to a below par performance" for any particular part and another rater "tended to use what appeared to be the predominant level of language over the whole test".

Raters also referred to "jagged profiles" in candidate performances, i.e. differential performance on different test parts. "Candidates rarely fit a single band [level]", one rater noted, adding that an advantage of a part MM was in "capturing candidate performances that sometimes varied widely in different parts of the test".

These findings confirmed that the choice of MM has a practical impact on the final CEFR classifications of more than 30% of candidates. We therefore proceeded to compare the two MMs in an attempt to identify the model exhibiting superior measurement qualities.

We combined the two data sets, defined MM as an additional facet, and reran the MFRM analysis. The results indicated that the two MMs were not statistically distinct, with the separation indices confirming that the two MMs could not be reliably divided into different strata (H=0.33; R=0.00). A closer look at other indices, however, revealed differences: although the examinee statistics showed comparable distribution of speaking abilities – albeit slightly wider for the holistic MM (1.23 to 6.0) compared to the part MM (1.55 to 5.99) – the part MM separated candidates into more statistically distinct levels (H_{Part}=6.54; R_{Part}=0.90) than the holistic MM ($H_{Holistic}$=3.46; $R_{Holistic}$=0.85). In other words, the use of the part MM resulted in more reliable distinctions between candidates. Secondly, although the rater statistics showed similar severity rankings and acceptable levels of consistency based on individual and average infit mean square statistics for the four raters, the holistic MM showed overfit for the two most lenient raters, with infit mean square values close to zero. While overfit may be "less productive for measurement", it is not considered "degrading" (Linacre, 2018b, p.279). Overfit is typically an indication of central tendency or restriction of range, and in this case, is in line with previous research on holistic MMs (Barkaoui, 2011; McNamara, 1996). In order to ensure that these two raters were not unduly affecting the results, we re-ran the MFRM by removing the problematic raters from the analyses and examined the impact on candidate separation indices. While results showed a slight improvement in the separation indices for candidates (H=3.55; R=0.87), this difference was small and did not affect the interpretations.

1

Moreover, the same two raters did not show any underfit/overfit in the remaining analyses and we therefore retained them in the analyses.

Additionally, we considered the percentage of unexpected responses flagged by FACETS[3]; this figure was higher for the part MM (1.27%) compared to the holistic MM (0.77%). A closer look revealed that it was the differential performance of the same three candidates on different test parts that resulted in an increased number of unexpected responses. This suggests that candidate performances on different test parts are varied enough to be distinguished by raters. The test part results substantiate this finding: different parts exhibited a difficulty range of 0.4 logits from -0.23 logits (Part 1– easiest) to 0.17 logits (Part 4– most difficult). The separation indices suggested that the different parts can be divided into a minimum of 2.64 statistically distinct difficulty strata (R=0.85). Raters' open comments indicated that they believed the part MM to be a more "fair" and "reliable" approach to assessment. Moreover, all four raters agreed that marking by part is feasible in operational conditions.

In summary, phase 1 findings suggested that choice of MM has a practical impact on the CEFR classifications of at least 30% of candidates, with a pattern of higher CEFR levels with the holistic MM. Amongst the two MMs under examination, we argue that there is a case to be made for adopting the part MM given its enhanced measurement properties in reliably distinguishing between candidates from different ability levels and separating them into almost twice as many ability strata compared to the holistic MM. Qualitative support for the part MM derives from raters' expressed preferences for this model in allowing more reliable and fair assessment of candidates. We therefore selected the part MM from this phase to be compared against the analytic MM in phase 2.

A comparison of part and analytic MMs

In Phase 2 we compared the part and analytic MMs and explored their impact on candidate scores/CEFR classifications. Similar to phase 1, we first estimated candidate ability levels on the basis of the MFRM analyses as measured by the two MMs. The analytic MM resulted in a higher mean (*M*=3.54; SD=0.74) than the part MM (*M*=3.31; SD=0.84) and a paired-sample *t*-test revealed that these differences were statistically significant (*t* (395) = 9.23, *p*<0.05, *r*=0.42) with a moderate effect size. These were confirmed in the results of a Wilcoxon Signed-Rank test that showed statistically significant differences in the ranking of candidates across the two MMs, with candidates receiving different (higher) rankings with the analytic MM (*z*=-8.17, *p*<0.01, *r*=-0.42). We also correlated these fair-average measures; results showed a strong statistically significant correlation (*r*=0.80, *p*<0.01).

---

[3] The percentage of unexpected responses can be used for evaluating fit of the data to the model; according to Linacre (2018b, p.171), "when the data fit the model, about 5% of standardized residuals are outside ±2, and about 1% are outside ±3". In both phases, results suggested acceptable fit of the data to the Rasch model.

These fair-average measures were then converted to CEFR levels. Table 4 shows a broadly similar distribution of CEFR levels; aligning with the above findings, the analytic MM resulted in a slightly higher percentage of candidates at the upper B and C levels, whereas the part MM resulted in a higher percentage of candidates at the A levels.

[Insert Table 4]

We then focused on the percentage of candidates receiving the same CEFR classification across the two MMs, taking into account the size and direction of any differences. Results showed approximately 52% of candidates receiving the same CEFR classification, with the remaining 48% falling within an adjacent upper/lower level. There was once again a systematic trend of lower CEFR levels with the part MM (see Appendix D for a contingency table with details of specific occasions where the two MMs converged/ diverged). Two comments by raters help explain this trend: "there was no criterion relating to task achievement. I found myself applying this anyway, almost instinctively" and "even though there is no task completion, when marking by part, it's easier to mark down answers that are not relevant to the task".

These findings confirm that choice of MM could impact candidates, resulting in a drop or increase in their overall CEFR classifications. We then proceeded to further compare the two MMs to identify the model exhibiting superior measurement qualities.

Similar to phase 1, we combined the two data sets and defined MM as an additional facet. The results indicated that the two MMs were statistically distinct (X^2=105.9, p<0.01) with the separation indices (H=9.94; R=0.98) showing that the two MMs could be reliably divided into approximately nine difficulty strata. These findings suggest that regardless of the strong correlations between candidate ability measures across the two marking conditions, the part and analytic MMs distinguish between candidates in different ways and potentially tap into distinct aspects of the construct.

We subsequently considered other indices to help evaluate and compare the two MMs, and examined the part and criteria statistics to see whether candidates' speaking abilities on the different test parts and criteria are sufficiently distinct to merit part or analytic marking. Similar to phase 1, the results of the part statistics showed that the four test parts in the speaking test exhibit a range of difficulty levels, from -0.20 (Part 1 – easiest) to 0.19 (Part 4 – most difficult), with the separation and reliability indices (H=4.32; R=0.90) suggesting that the different test parts can be reliably separated into a minimum of four statistically distinct difficulty strata. The null hypothesis that all parts exhibit similar difficulty measures was rejected (X^2=39.3, p=0.00). This serves as empirical evidence that candidates may be displaying speaking abilities on the different test parts that are sufficiently distinct, thus justifying a part MM. Similarly, the criterion measurement report of the analytic score data showed that the four criteria in the scale exhibited a range of difficulty measures, with *coherence and discourse management* as the easiest category (-0.31 logits) and *pronunciation* as the most difficult (0.13 logits). The

separation strata and reliability indices (H=5.02; R=0.95) indicated that candidate performances on the different criteria are sufficiently varied to be reliably distinguished by raters thus justifying the use of an analytic MM.

Candidate group-level statistics showed that the part MM separated candidates into slightly more statistically distinct levels than the analytic MM (H_{Part}=6.62, R_{Part}=0.96; $H_{Analytic}$=5.58; $R_{Analytic}$=0.94). The rater statistics generally showed acceptable levels of consistency for the ten raters in the study across the two MMs. However, two of the raters exhibited slight underfit with the analytic MM (infit mean square values >1.5) while none of the raters in the part MM exhibited misfit. Lastly, the percentage of unexpected responses (those with residual values > |2|) was 1.30% and 1.10% for the analytic and part MMs respectively.

Raters' views regarding preference for the two MMs were mixed; six out of ten raters preferred the part MM, two raters were equally happy with either MM, and two raters preferred the analytic MM. Preference for the latter was based on the ability to capture uneven profiles: "I found it hard to balance the different aspects of assessment to give an overall mark in some cases, e.g. very good pronunciation but with limited vocabulary and grammar". Another rater noted that "each part has its limits and there may not always be enough evidence to mark each part".

Raters believed that marking the test by part or analytically was feasible in operational settings. One rater distinguished between face-to-face tests (where the examiner serves the dual role of interlocutor and rater) and computer-delivered tests (where the examiner is only focused on rating). In the latter, the time pressure is removed and rater cognitive load is lower, which supports the feasibility of awarding multiple scores particularly when "there is more control over the audio files and you can pause and move on as you like".

To summarise, our findings from phase 2 suggest that the choice of MM has an impact on candidates' CEFR classifications. Both part and analytic MMs exhibited precision in measurement, as expected given the number of observations per candidate (four in each case). Raters' perceptions also provided support for both models in terms of feasibility of use in operational settings. There was, however, a slight advantage for the part MM on three grounds: (a) it allowed for finer distinctions between candidates, (b) raters showed higher consistency, and (c) the percentage of unexpected responses were lower. Additionally, more raters preferred the part MM, although given their limited number, this finding should be treated with caution.

A comparison of all MMs

Each of the two phases focused on a detailed comparison of pairings of MMs. Given the study's linked design we were able to compare all three MMs simultaneously; candidate ability estimates from the independent FACETS

analyses were analysed with a two-facet model consisting of candidates (*n*=637) and MM (*n*=3).

The MM map is presented in Figure 2 and the MM measurement report is summarised in Table 5 with MMs arranged in ascending order of difficulty.

[Insert Figure 2]

[Insert Table 5]

Results show a logit range of 1.7, with the analytic MM as the easiest (logit value=-0.95) and the part MM as the most difficult (logit value= 0.75). All infit mean square values fall within an acceptable range. The separation and reliability indices (H=9.65; *R*=0.98) suggest that the different MMs can be reliably separated into a minimum of nine statistically distinct strata. The null hypothesis that all MMs exhibit similar difficulty measures is rejected (X^2=199.4; d.f.=2; *p*=0.00). This can be interpreted as follows: the probability of the *same* candidate receiving a *different* ability estimate as a function of choice of MM is statistically significant.

We also explored the extent to which these statistically significant results translate into practical significance in terms of CEFR classifications. In doing so, we considered the fair-average results associated with each MM and calculated the maximum difference between the easiest (analytic) and most difficult (part) MM (3.78-3.26=0.52). The value of 0.52 is approximately half a CEFR level, which can have a practically significant impact, particularly for borderline candidates. This effect is attenuated when comparing the two easiest or the two most difficult MMs. Nevertheless, these results demonstrate the potential impact of choice of MM on candidates and in this particular validation investigation have lent support to a change to an alternative MM.

We also ran a candidate x MM bias analysis; this analysis allows for an examination of whether each MM maintains a consistent level of difficulty across candidates. From a total of 1264 interaction terms, there were 389 bias terms with Z values > |2|. However, no single interaction displayed statistical significance (*p* > 0.05) with the summary statistics (X^2=1037.5; d.f.=1264; *p*=1.00) suggesting that the different MMs do not disadvantage different candidates in a statistically distinct way. These somewhat contradictory results can be explained by the small number of observations for each candidate which is critical for statistical significance testing (Eckes, 2009; M. Linacre, personal communication, October 2019). We therefore considered alternative approaches for identifying large and meaningful interaction terms. The first approach, following Eckes (2009), was to calculate the percentage of *t* values larger than |2| as an indication of large bias. In our data set, this was 3.1% (*t* value range: -2.96 to 2.46). The second approach was to consider substantial differences between observed and expected averages for each interaction term. Given that half a CEFR level can be of practical significance in our context, we identified any cases where Observed – Expected Average> |0.5|. In our data set, this was 4.7%; the largest absolute value was |0.88|, which is less than one CEFR level. When

both approaches are combined, this percentage is 4.8%. We can therefore conclude that in general MMs display a uniform level of difficulty across candidates; however, for a small percentage of the candidature (<5%), there is evidence of bias.

Discussion

In this study, we compared different MMs in terms of their impact on candidate CEFR classifications and measurement qualities in a specific speaking assessment context. This was done with the view to potentially switch to a model that would allow for the generation of more fine-grained information on performance while maintaining and/or enhancing the test's scoring validity and preserving the practical demands for a quick results turnaround.

There was strong evidence from both phases to suggest a significant impact of choice of MM; approximately 30% and 50% of the candidates in phases 1 and 2 respectively were awarded different (adjacent) CEFR levels depending on the MM. When all MMs were considered together in a single analysis, results showed half a CEFR level difference between the easiest and most difficult MM. Taken together, findings suggest that the probability of the same candidates receiving different scores/CEFR levels as a function of choice of MM is statistically and practically significant and should therefore be an important test validity consideration and carefully aligned with the purpose of that assessment.

Findings also indicated trends in the direction of these differences. In contrast to Barkaoui (2011) and in line with Wiseman (2012), our study found higher overall scores and CEFR levels with the holistic MM. Supported by raters' open comments, it appears that the holistic MM lends itself to a benefit-of-the-doubt policy, with scores awarded on the basis of candidates' best performance, which was also stipulated by Bacha (2001). A halo effect across performance on different tasks is possibly in operation with the holistic MM. The trend of lower scores/CEFR classifications with the part MM was also observed in the comparison with the analytic MM. We drew on raters' comments to suggest that, in the absence of a task achievement criterion, the part MM seems to serve as an implicit alternative that allows raters to award lower scores for task-irrelevant responses or unsuccessful attempts at handling more complex tasks.

Focusing on a comparison of the MMs from a measurement perspective, our findings showed strong correlations between all pairings of MMs; however, further analyses revealed important differences. The two MMs that showed the strongest correlation with no statistically significant difference in means and candidate rankings were the holistic and part MMs, suggesting that the two are tapping into a similar construct. This is not surprising given that the same scale and criteria were used in both models, with the holistic MM applied to a whole performance and the part MM applied to individual test parts. Nevertheless, the part MM was shown to display superior measurement

qualities particularly in separating candidates into more ability strata. A possible explanation is that the part MM includes multiple observations for each candidate and this additional information may result in increased measurement precision (Barkaoui, 2011). To verify this, we used the average of the four test parts instead of the four independent scores in an additional MFRM analysis (not reported here due to space limitations) and observed a drop in measurement precision. This implies that candidates are indeed displaying differential performance across the four test parts and the part MM allows raters to make these finer-level distinctions. These results align with Nakatsuhara's (2011) argument that a single score is not necessarily a good representation of a candidate's ability on different test parts; they also substantiate Barkaoui (2011), who found that the holistic approach may not be sufficiently sensitive to differences in performances. This was echoed in rater comments who noted that the part MM allowed them to take a more objective and critical stance and award scores that better represented candidate performances on each test part. For this reason, the part MM was selected as a stronger contender to be compared to the analytic MM in phase 2.

Results of the second phase also showed strong correlations between the analytic and part MMs; further statistical analyses, however, pointed to significant differences in means and ranking of candidates, suggesting that the two models may distinguish between candidates in distinct ways and may be tapping into different aspects of the speaking construct. A possible reason is that the analytic approach focuses raters' attention more explicitly on individual features of performance such as fluency or grammar, whereas the part MM invites raters to consider language use holistically but in dynamic interaction with individual tasks.

Unlike the holistic MM, the analytic and part MMs both included multiple observations, i.e. four score points per candidate performance. As expected, they displayed increased measurement precision. The part MM, nevertheless, exhibited marginally higher measurement precision, evidenced in the separation of candidates into more statistically distinct ability levels and higher levels of rater consistency. Raters were mixed in their preference for either approach, with the majority favouring the part MM. Both MMs were considered to be practically feasible in operational settings.

Taking all results together, we believe that a strong case can be made for adopting the part MM for the assessment context investigated in this study on the following grounds: firstly, it allows for more measurement precision compared to the current operational holistic model and the analytic MM; secondly, it can provide more diagnostic information to candidates in terms of their performance on different test parts; thirdly, there is evidence that candidate performances on the different test parts are varied enough to merit part scoring; fourthly, there is support from raters regarding its feasibility of application in operational settings; and lastly, it has the practical benefit of requiring minimal changes to the current rating scale and the possibility of implementation within a short timeframe.

Implications

The findings shed light on the influence of MMs in a semi-direct online speaking test, with evidence of significant impact on candidate scores/CEFR classifications. It is therefore essential for test development and validation activities to take MM into account and clearly justify the choice of MM.

All MMs under examination were successful in distinguishing between candidates from different ability levels; however, raters showed the highest sensitivity to differences in performance when using the part MM. As discussed in the literature review, the part MM is not commonly investigated or reported on. In speaking tests that consist of a variety of task types, there is an underlying assumption that candidate performances on different tasks may vary and as such, a MM that attunes to these variations is appropriate. We would like to argue that the part MM is a feasible alternative to the more commonly used methods of analytic and holistic scoring. Adopting a part MM can also facilitate the inclusion of a task achievement criterion, which is evaluated at the part level. The downside is that the part MM cannot necessarily reflect "jagged profiles" in terms of linguistic features of performance (e.g. fluency vs. grammar use). Possible solutions to address these limitations are discussed in the next section. Nevertheless, in light of these findings, a recommendation was made for the operational test in question to adopt a holistic by part MM, which included a task achievement descriptor per part.[4]

An additional implication relates to the choice of MM when developing, training and evaluating automated marking systems. Automated assessment technologies largely rely on human-awarded marks as the gold standard for the training and evaluation of their systems (Chen et al., 2018). Our study has shown that choice of MM has an impact on those human-awarded marks, and as such it directly influences the source data for machine learning purposes and system evaluations. When reporting the results of human-machine agreement levels, it is important to be transparent about how those human marks are derived.

Future directions

The scope of our study was limited to three specific MMs and did not allow for an examination of other MMs. Hybrid MMs, for example, can be considered in addressing the limitations of the part MM; to illustrate, a task achievement criterion can be applied to each test part, alongside analytic criteria applied to the whole test. Alternatively, marking criteria can be tailored to the task focus; for example, an extended monologue task could be rated for discourse organisation, grammar, and vocabulary; a question-and-answer

---

[4] This recommendation has not been implemented at the time of publication as it is undergoing further research as part of the validity evidence needed to support scale revision.

task in the same test could focus on grammar and vocabulary, and a read-aloud task on pronunciation only. A full picture of ability can thus be provided through the complementary use of specific assessment criteria for specific tasks. This suggestion aligns with Taylor and Galaczi (2011) who advocate such an approach, although they caution that the choice of assessment criteria per test part would need to be empirically justified.

The wide use of automated scoring technologies also offers the potential for different hybrid models of machine and human scoring (Isaacs, 2018; de Jong, 2018). For example, an auto-marker can focus on features of speech that it can most reliably assess, such as fluency and pronunciation, while raters can focus on more complex elements such as coherence or task achievement. Alternatively, machines could be used to assess more predictable routine tasks with human raters focusing on extended spontaneous speech on less predictable topics. Such complementary approaches can increase scoring reliability and minimise the limitations of both human-mediated and automated MMs. An investigation into different hybrid MMs can therefore be an exciting avenue for future research.

Limitations

There are three main limitations in our study: firstly, the number of raters was small, which can affect the robustness of the statistical analyses and restrict generalisations of the qualitative data; secondly, the analytic scale used here was not independently piloted; and thirdly, there were differences between the CEFR classifications in the operational data and those observed across the three MMs, which is likely to be attributable to noise in the operational single-scored data. These limitations notwithstanding, we believe we have provided evidence-based insights which can inform future theoretical discussions and operational considerations on MMs.

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System,29*(3), 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal,70*(4),380–390. https://doi.org/10.1111/j.1540-4781.1986.tb05294.x

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

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1),1–15. https://doi.org/10.1177/026553229501200101

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessments*. Honolulu, HI: University of Hawai'i Press.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language*

acquisition and language testing research (pp. 32–70). New York: Cambridge University Press.

Chen, L., Zechner, K.,Yoon,S., Evanini, K.Wang,X., Loukina,A., Jidong, T., Davis,L., Lee, C. M., Ma,M., Mundkowsky,R., Lu,C.,Leong, C.W., & Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRater^{SM} v. 5.0 engine*. ETS Research Report Series. https://doi.org/10.1002/ets2.12198

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, California: Sage Publications.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Studies in Language Testing volume 7. Cambridge: UCLES/Cambridge University Press.

Davis, L. (2018). Analytic, holistic, and primary trait marking scales. *The TESOL Encyclopaedia of English Language Teaching,*1–6.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the test of spoken English revision project*. Princeton, NJ: Educational Testing Service.

Eckes, T. (2009). Section H: Many-facet Rasch measurement. *Reference Supplement to the Manual for Relating Language Examinations to the CEFR for Languages: Learning, Teaching, Assessment,*1–52. Retrieved from: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a23

Fulcher, G. (2003). *Testing Second Language Speaking*. Harlow: Pearson Education.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–99. https://doi.org/10.1177/0265532210364407

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research & Development in Education,27*(2), 73–82.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly,29*(4), 759–762. https://doi.org/10.2307/3588173

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

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation,11*(2), 61–85.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly,15*(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264

De Jong, N. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3),237–254. https://doi.org/10.1080/15434303.2018.1477780

Kroll, B. (1990). *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131–166. https://doi.org/10.1191/0265532206lt325oa

Lee, Y., Gentile, C., & Kantor, R. (2009). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*(3), 39–417. https://doi.org/10.1093/applin/amp040

Linacre, J. (2018a). Facets Rasch measurement computer program (version 3.81). [software]. Available from: https://www.winsteps.com/facets.htm

Linacre, J. (2018b). A user's guide to FACETS: Rasch model computer programs Retrieved from: https://www.winsteps.com/a/Facets-Manual.pdf

McNamara, T. F. (1996). *Measuring second language performance*. Reading, Mass: Addison Wesley Longman.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement*, *5*(2), 189 – 227.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the test of spoken English assessment system. *ETS Research Report Series, 2000*(1),i–51. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x

Nakatsuhara, F. (2011). *The relationship between test-takers' listening proficiency and their performance on the IELTS speaking test*. IELTS Research Reports Volume 12.

Norris, J. M., Brown, J. D., Hudson, T., & Bonk, J. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing*, *19*(4), 395–418. https://doi.org/10.1191/0265532202lt237oa

O'Sullivan, B., & Nakatsuhara, F. (2011). Quantifying conversational styles in group oral test discourse. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 164–185). Basingstoke: Palgrave MacMillan.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics*, *22*(1), 27-57. https://doi.org/10.1093/applin/22.1.27

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

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking* (pp.171–233). Cambridge: UCLES/Cambridge University Press.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation*. Basingstoke: Palgrave McMillan.

Weir, C. J. (1990). *Communicative Language Testing*. Hemel Hempstead: Prentice Hall.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400–409.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150–173. https://doi.org/10.1016/j.asw.2011.12.001

Wright, B.D. & Masters, G.N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16* (3), 888.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Language Testing, 24*(2), 251-286. https://doi.org/10.1177/0265532207076365

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501–527. https://doi.org/10.1177/0265532214536171

[Insert Appendix A]
[Insert Appendix B]
[Insert Appendix C]
[Insert Appendix D]
[Insert Appendix E]
[Insert Appendix F]

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

Month 1

Month 4

Phase I (Holistic vs. Part MM)

Phase II (Part vs. Analytic MM)

Test Set
- Speaking Tests (n=240)
- Range of L1s/CEFR levels
- Raters (n=4)

Phases linked through common raters

Test Set
- Speaking Tests (n=400)
- Range of L1s /CEFR levels
- Raters (n=10)

Rating Matrix
(a) candidates in test set receive *at least* one **holistic** mark and one mark for each test **part**
(b) raters linked through common performances
(c) ratings counter-balanced for MM

Rating Matrix
(a) candidates in test set receive *at least* one **analytic** mark for each criterion and one mark for each test **part**
(b) raters linked through common performances
(c) ratings counter-balanced for MM

Round 1
n_1=120 (Holistic MM)
n_2=120 (Part MM)

Raters linked through common batch

Round 2
n_1=120 (Part MM)
n_2=120 (Holistic MM)

Round 1
n_1=200 (Part MM)
n_2=200 (Analytic MM)

Raters linked through common batch

Round 2
n_1=200 (Analytic MM)
n_2=200 (Part MM)

Rater questionnaire

Rater questionnaire

Analysis
Quan (MFRM of score data)
Qual (thematic analysis of open comments)

Part MM selected for Phase II

Analysis
Quan (MFRM of score data)
Qual (thematic analysis of open comments)

```
+--------------------------------------------+
|Measr|+Candidates  |-MM          |Scale|
|-----+-------------+-------------+-----|
|   11 + **.          +             +  (6) |
|   10 + .            +             +  --- |
|    8 + *.           +             +   5  |
|    6 + ***.         +             +  --- |
|    5 + *******.     +             +   4  |
|    3 +              +             +      |
|    1 + ********     + Part        +  --- |
*    0 * **********.  * Holistic    *   3  *
:      :              : Analytic    :      :
|   -2 + *******.     +             +  --- |
|   -4 + ******.      +             +   2  |
|   -5 + ***.         +             +  --- |
|   -7 + **.          +             +      |
|   -9 + **.          +             +   1  |
|  -10 + .            +             +  --- |
|  -12 + *.           +             +  (0) |
|-----+-------------+-------------+-----|
|Measr|  * = 11       |-MM          |Scale|
+--------------------------------------------+
```

Figure 2. Map of all MMs

Table 1. Rater examining/teaching experiences.

Number of Years	Speaking examining experience	EFL/ESL teaching experience
<1	0	0
2-5	5	0
6-10	3	2
10+	2	8

Table 2. CEFR level breakdown for phases 1 and 2 datasets.

CEFR Level	Phase 1		Phase 2	
	N	%	N	%
A1	30	12.5	50	12.5
A2	60	25	100	25
B1	60	25	100	25
B2	60	25	100	25
C1	25	10.4	45	11.2
C2	5	2.1	5	1.3
Total	240	100	400*	100

*Three candidates were later dropped from the analysis due to audio quality.

Table 3. Distribution of CEFR levels by holistic and part MM (%).

CEFR Level	Marking Model	
	Holistic (%)	Part (%)
A1	3.7	6.2
A2	13.3	14.2
B1	48.3	47.9
B2	25.4	22.1
C1	4.7	6.7
C2	4.6	2.9
M	3.70	3.64
SD	0.90	0.83

n=240

Table 4. Distribution of CEFR levels by part and analytic MM (%).

CEFR Level	Marking Model	
	Part (%)	Analytic (%)
A1	9.9	4.0
A2	21.9	17.9
B1	44.1	47.9
B2	21.7	25.6
C1	2.3	4.1
C2	0.3	0.5
M	3.31	3.54
SD	0.84	0.74

n=397

Table 5. MM measurement report.

Marking Model	Difficulty (logits)	Fair-M Average	Infit Mean Square
Analytic	-0.95	3.78	1.05
Holistic	0.20	3.52	0.90
Part	0.75	3.26	0.96

Appendix A. Description of test parts with example tasks.

Part	Focus	Example
1	Short interview questions on personal topics *General interaction and social language, giving personal information.*	How often do you use English? In your country, how important is it to learn English?
2	Describing and comparing two photographs *Organising a lager unit of discourse, describing and comparing two pictures.*	
3	Questions on a familiar topic *Responding appropriately to a series of questions on a common theme, giving information, expressing and justifying opinions, suggesting, comparing and contrasting.*	Where is a good place to eat in your town? In your country, how popular is eating out?
4	Long turn on abstract topic *Organising a coherent and extended opinion/argument with examples and justification, speculating, and describing.*	Some people say that going to a concert to listen to live music is better than listening to recorded music. What do you think?

Appendix B. Assessment scale.

Band	Global Descriptors
C2 • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Fully operational command of the spoken language** • Able to express both simple and complex ideas with ease; coherent extended discourse. • Consistently, displays wide range and accurate use of grammar and vocabulary. • Pronunciation is easy to understand; stress, rhythm and intonation are used to express meaning effectively. • Responds promptly with only natural hesitation; makes effective use of the allowed response time.
C1 • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Good operational command of the spoken language** • Able to express simple and complex ideas; generally extends discourse coherently. • Generally, displays wide range and accurate use of grammar and vocabulary. • Pronunciation is easy to understand; stress, rhythm and intonation are used to express meaning well. • Generally responds promptly, with only natural hesitation; generally makes good use of the allowed response time.
B2 • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Generally effective command of the spoken language** • Able to express simple ideas and makes some attempt to express complex ideas; mostly coherent, with some extended discourse. • There is an adequate range of grammar and vocabulary which is sufficiently accurate to deal with the tasks. • Pronunciation can generally be understood; stress, rhythm and intonation are used to express meaning adequately. • May be some hesitation while searching for language; generally makes adequate use of the allowed response time.
B1 • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Limited but effective command of the spoken language** • Able to express simple ideas; little extended discourse; some incoherence. • The range of grammar and vocabulary used is sufficient to complete tasks in a limited way. Some language in simple utterances is accurate but basic inaccuracies may impede communication of ideas and achievement of the tasks • Pronunciation can generally be understood but L1 features may cause strain; an attempt is made to use aspects of stress, rhythm and intonation to express meaning. • Hesitation may demand patience of the listener; use of the allowed response time may not always be adequate.
A2 • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Basic command of the spoken language** • No extended discourse • The range of language is sufficient to respond to simple prompts but not to complete complex tasks. Some utterances (single words or short phrases) may be accurate but inaccuracies in grammar and vocabulary limit achievement of the tasks and restrict coherence and communication of ideas. • Pronunciation of single words may be intelligible but L1 features may make understanding difficult; little attempt is made to use aspects of stress, rhythm and intonation to express meaning. • Hesitation is excessive; use of the allowed response time is adequate on only a few occasions.
A1 • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Minimal command of the spoken language** • Utterances may be limited to single words. • The range of language is limited and inadequate to complete the tasks. Some accurate language but frequent inaccuracies may mean the message is not communicated. • Pronunciation of single words may be intelligible but L1 features may cause excessive strain to a listener; no attempt is made to use aspects of stress, rhythm and intonation to express meaning. • Hesitation is excessive; use of the allowed response time is generally inadequate.
0	• Throughout the task, responses are not attempted, OR consistently no meaning is conveyed, OR responses are consistently unrelated to the rubric.

Appendix C. Contingency table - % of agreements between holistic and part MMs.

Marking Model	Holistic MM	Row Total (%)

Part MM	CEFR Levels	A1	A2	B1	B2	C1	C2	
	A1	3.6	0.1					3.7
	A2	2.6	8.1	2.6				13.3
	B1		6.0	36.4	5.9			48.3
	B2			8.4	15.9	1.1		25.4
	C1			0.5	0.3	2.8	1.1	4.7
	C2					2.8	1.9	4.6
	Column Total (%)	6.2	14.2	47.9	22.1	6.7	2.9	100

*Due to rounding, percentages may not always appear to add up to 100%.

Appendix D. Contingency table - % of agreements between analytic and part MMs.

Marking Model		Analytic MM						Row Total (%)
Part MM	CEFR Levels	A1	A2	B1	B2	C1	C2	
	A1	3.8	0.2					4.0
	A2	6.1	8.6	3.2				17.9
	B1		13.1	27.3	7.5			47.9
	B2			13.6	11.1	1.0		25.6
	C1				3.0	1.1		4.1
	C2					0.2	0.3	0.5
	Column Total (%)	9.9	21.9	44.1	21.7	2.3	0.3	100

*Due to rounding, percentages may not always appear to add up to 100%.

Appendix E. Summary results for phase 1

Phase 1 Focus on **Holistic MM**: 2-facet model (Candidate x Rater)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.70	0.90	1.74	2.85	0.07	3.46	0.85	240
Rater	3.71	0.36	N/A	0.37	0.60	12.56	0.98	4

Phase 1 Focus on **Part MM**: 3-facet model (Candidate x Rater x Part)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.64	0.83	1.38	0.69	0.84	6.54	0.90	240
Rater	3.63	0.3z8	N/A	0.09	0.96	21.11*	1.00	4
Part	3.60	0.03	N/A	0.09	0.99	2.64	0.85	4

Phase 1 Focus on **Holistic and Part MM comparison**: 3-facet model (Candidate x Rater x MM)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.70	0.88	1.39	1.43	0.72	6.58	0.96	240
Rater	3.69	0.49	N/A	0.18	0.87	22.06*	1.00	4
MM	3.61	0.00	N/A	0.12	0.88	0.33	0.00	2

*The rater separation strata in phase 1 (and also in phase 2 – see results in Appendix F) are relatively high. While the rater separation indices are expected to be large when the "number of observations per rater […] is large" (Myford & Wolfe, 2004, p.197), these figures are higher than expected. A closer look at the rater data revealed that one extreme – severe but not misfitting – rater contributed substantially to these figures. This was evidenced in a large drop in the separation strata once this rater was removed from the analyses: in phase 1, figures dropped from 12.56 and 21.11 to 9.08 and 11.02 for holistic and part analyses; in phase 2, figures dropped from 20.78 and 18.27 to 12.93 and 11.68 for analytic and part analyses. Removing the rater, however, had minimal impact on the results of other facets. Moreover, the maximum difference between the most severe and most lenient rater (in Fair-M Average) in the case of the highest rater strata measure of 21.11 (phase 1 part analysis) was 1.01, which is one CEFR level. Given that the raters in our study were not trained or standardised to use the rating scale with the analytic and part MMs, these differences in severity are to be expected. We therefore made a decision to retain all raters in the analyses on the basis of a number of considerations: (a) the extreme rater was not showing misfit and was consistently severe, (b) the impact of removing the extreme rater was minimal on the remaining analyses, (c) differences in rater severity levels are to be expected when raters have not been trained/standardised with new MMs, and (d) when raters are consistent within themselves, the use of MFRM allows for other facet measures to be adjusted for differences in rater severity levels. Should any of the suggested MMs become operational, however, it would be important to conduct thorough training and standardisation to reduce any large differences in severity

Appendix F. Summary results for phase 2

Phase 2 Focus on **Analytic MM**: 3-facet model (Candidate x Rater x Criteria)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.54	0.74	1.00	0.50	0.76	5.58	0.94	397
Rater	3.57	0.52	N/A	0.07	1.11	20.78*	1.00	10
Criterion	3.65	0.06	N/A	0.04	1.00	5.02	0.95	4

Phase 2: Focus on **Part MM**: 3-facet model (Candidate x Rater x Part)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.31	0.84	-0.28	0.60	0.78	6.62	0.96	397
Rater	3.45	0.66	N/A	0.08	1.07	18.27*	1.00	10
Part	3.55	0.03	N/A	0.05	1.05	4.32	0.90	4

Phase 2 Focus on **Analytic and Part MM comparison**: 3-facet model (Candidate x Rater x MM)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.35	0.78	-0.20	0.68	0.79	5.97	0.97	397
Rater	3.50	0.35	N/A	0.09	0.98	17.41*	0.97	10
MM	3.51	0.09	N/A	0.04	1.02	9.94	0.98	2

Comparison of all three MMs: 2-facet model (Candidate x MM)	Summary results							
	Fair-M Average	SD	Logit (M)	SE	Infit Mean Square (Average)	Strata	Separation R	n
Candidate	3.48	0.80	0.21	1.40	0.95	5.72	0.94	637
MM	3.55	0.10	N/A	0.10	0.97	9.65	0.98	3

* See note in Appendix E.