IELTS RESEARCH REPORTS

IELTS Writing Scales Review and Update – summary overview (November 2023)

Introductory note

Research for IELTS is conducted collaboratively between the three IELTS test partners – British Council, IDP IELTS and Cambridge University Press & Assessment (Cambridge). The IELTS Research Group (IRG) is a cross-partner team of 10–12 experienced researchers based globally. It is overseen by several internationally established members of the language assessment field.

The purpose of the IRG is to collaborate on test development, trialling of tasks or any work that goes beyond the day-to-day IELTS requirements (e.g., designing new task types or exploring how the band descriptors for the productive skills may be updated and optimised, as described in this document).

This summary overview draws upon three lengthy internal reports that were compiled over the 4-year period of the project (2019–2023) to review and update the criteria and band descriptors for IELTS Writing for both the Academic and General Training tests. The original internal reports were written up by members of the IRG from all three partner organisations. Extracts were then selected and edited to provide this much shorter and accessible summary overview for release on the IELTS website.

Executive summary

This summary overview reports on a systematic review of the IELTS Writing criteria and band descriptors conducted in response to ongoing monitoring of test performance, including anecdotal feedback from IELTS Writing examiners over several years. The cross-partner project was conducted between March 2019 and May 2023 and entailed multiple phases.

Phase 1 investigated quantitative and qualitative data collected via a survey questionnaire and focus groups from at least 400 IELTS Writing examiners worldwide. Data analyses from Phase 1 suggested a need to remove perceived ambiguities in some scales and band descriptors through adjustment of the existing wording.

Building on the Phase 1 findings, Phase 2 involved a team of IELTS senior examiners and researchers making targeted adjustments to the existing IELTS scales and band descriptors to prepare a draft revised version for field trialling.

The Phase 3 field trial collected quantitative scoring data from 103 examiners marking 882 responses using the adjusted band descriptors, accompanied by questionnaire and focus group feedback. Score data analyses, using Linear Mixed Effects (LME) and Multi-Faceted Rasch Measurement (MFRM) models, indicated that the adjusted scales demonstrated appropriate technical performance qualities, while examiner feedback showed that raters perceived the adjustments as a positive improvement.

Although work in Phases 1, 2 and 3 had been completed by February 2020, the onset of the Covid-19 pandemic in March 2020 made the planned implementation of the adjusted scales and band descriptors for IELTS Writing difficult at that time. Instead, the IELTS partners prioritised innovations in the existing test system which could respond to and keep pace with rapidly changing market requirements due to the pandemic (e.g., IELTS Indicator).

Phase 4 of the review project was initiated in January 2022 in preparation for operational implementation of the adjusted scales and band descriptors. Between May and July 2022, a cross-partner team of subject matter experts made further refinements to the descriptor wording and developed revised examiner training materials. The appropriate functioning of the revised scales was confirmed in a study that examined score data from 1,000 IELTS Writing responses marked twice by 100 examiners, once using the existing descriptors and then using the adjusted descriptors. Feedback from examiners was also gathered for analysis. The outcomes of the Phase 4 study showed no cause for concern and the adjusted band descriptors were introduced operationally from May 2023.

PHASE 1: REVIEW (March to May 2019)

Original internal technical report authors: Clark, T., Tasviri, R., Lopes, S., & Galaczi, E. (2019).

Introduction

Assessment criteria and scales are recognised as key elements of performance tests, impacting on test reliability, scoring validity, and ultimately test fairness. For some years, the IELTS Writing criteria and scales (*Task Achievement, Task Response, Coherence & Cohesion, Lexical Resource, Grammatical Range & Accuracy*) had not been subjected to in-depth review and revision. Renewed internal interest with how the scales performed and were being applied by IELTS examiners to evaluate test taker performance triggered an investigation in 2019 to discern whether relatively minor revisions might be needed to address ambiguities in wording or whether more extensive changes might be required.

Methodology

Research design and research questions

The Phase 1 study adopted a multi-phase *mixed methods design* (Creswell, & Plano Clark. 2011): qualitative and quantitative survey data were collected first and then used to inform focus group data gathering. The multiple data sources were then integrated and interpreted to inform any recommendations for scale revision. Research questions (RQs) were as follows:

RQ1: How do IELTS examiners perceive the functioning of the scales in terms of clarity and confidence in using them to award marks?

RQ2: Which features of the scales do examiners perceive as affecting their ability to score confidently and accurately?

Data collection: sources and participants

A large-scale worldwide *online survey* was used to explore IELTS examiners' views and experiences of using the existing scales. The survey content was based on anecdotal feedback from IELTS examiners about issues to explore, together with input from the internal research team. The instrument included selected-response, 5-point Likert-scale questions plus open-ended questions giving the respondents the opportunity to add further comments. It also included selected-response background questions to gather relevant demographic data. In total, the survey consisted of 33 questions. Following drafting and internal piloting, the finalised survey was delivered via SurveyMonkey to all existing British Council and IDP examiners. Completion was voluntary and no incentives were offered for completing it.

The survey was completed by 436 examiners. The sample population was made up of a range of examiners in terms of their teaching and examining experience, examiner role, examining location and partner affiliation. The majority of participants were experienced ELT teachers and IELTS examiners: 80.5% had 11–20+ years of teaching experience and 97% had more than one year of examining experience, with the largest group (35%) falling in the 5–9 years range. In terms of examiner role, 82.6% were examiners, 13.4% senior examiners and 2.3% assistant principal examiners. Examiners were working in 17 countries, with the majority based in the UK (43.3%), Australia (25.4%) and Canada (22.0%). Though not perfectly representative of the total population of IELTS examiners, the survey was believed to provide good representation of the views of the IELTS examiner population worldwide.

Following the survey, three *focus groups* were conducted: one in Australia with IDP assistant principal examiners and senior examiners, and two in the UK with British Council examiners. All three focus groups were conducted online using Zoom software, recorded and professionally transcribed. Each group comprised five examiners, a moderator (a member of the research team) and a note-taker to ensure that no observations were missed. Discussions were semi-structured in nature, allowing participants some freedom to express what they wished to, given their status as expert stakeholders. The focus group protocol questions probed deeper into the high-level trends that had already emerged in the earlier surveys (according to preliminary analysis of the data) in order to precisely target unfolding trends. One British Council group comprised more senior and experienced assistant principal examiners, while the other was largely formed of 'regular' examiners. This aimed to solicit perspectives from both those who had been examining for a considerable time period (generally considered as highly accurate 'experts') and those regular practitioners who are perhaps more representative of IELTS examiners globally.

Data analysis: examiner feedback survey and focus group discussions

Descriptive and inferential statistics were generated in SPSS (version 25) for the selected-response questions, in order to understand examiner perceptions about the functioning of the scales (RQs 1 and 2). All questions and options were converted into an 'item' in SPSS, totalling 181 items. Summary tables and graphs were generated for each response option, capturing the percentage of respondents endorsing each one, together with the mean/SD/min/max, depending on the item. Wilcoxon Signed Rank (Paired) tests were run for the nine IELTS score bands and across the five IELTS Writing criteria to investigate differences in the mean between two observations. Non-parametric tests were used as the scores were not normally distributed and exhibited a negative skew.

For the qualitative strands, Braun and Clarke's (2006) model of thematic analysis was used. Open-ended comments were analysed for key themes using NVivo Pro (version 11). In order to develop a uniform and consistent coding approach and to adequately capture the main themes in the examiner comments, two researchers independently coded the survey data and subsequently compared notes to verify that the emerging themes of relevance to the RQs overlapped. When the codes were compared, they were found broadly to agree. In cases where one theme was not exactly the same between the two sets of researchers' notes, a discussion brought them into line. Any smaller themes that were deemed to be irrelevant or less than critical to the RQs were discarded, due to the large dataset involved and the need for the analysis to remain highly focussed.

NVivo software and the same coding method as above were used to analyse all three focus group transcriptions in order to identify key topics and opinions discussed by the examiners. Researchers' independent notes on the main themes emerging from the focus groups were also compared to provide a full and accurate account of participants' perspectives.

Results

This section summarises aspects requiring further attention that were identified across the five IELTS scales (*Task Achievement, Task Response, Coherence & Cohesion, Lexical Resource, Grammatical Range & Accuracy*). Some aspects are common to all scales while others relate to specific scales. Results on assessing candidates at higher bands are presented first, followed by discussion of each of the five scales in turn, with a focus on:

- confidence using the top bands within that scale
- confidence applying the descriptors in that scale
- perceived ambiguities in the wording of some scale descriptors.

In addition, findings relating to the *Instructions to Examiners* and to computer-delivered (CD) IELTS are presented below.

Assessing higher bands in IELTS Writing (7, 8 and 9)

Some survey questions focused on examiners' overall level of confidence in discriminating between the 0–9 IELTS bands. Examiner confidence levels were used as an indicator of potentially challenging areas in and around the band descriptors.

Establishing why examiners report a lack of confidence is less straightforward than revealing that a problem itself exists. Interestingly, however, low confidence in

applying top bands was an outcome of previous research on the IELTS Writing scale carried out in 2012 (Galaczi, Lim, Khabbazbashi, & Vice 2012).

Task Achievement (TA)

Survey responses for most TA indicators suggested that the level of examiner confidence was reassuringly high, with 'confident'/'very confident' choices ranging between 71% and 84%. The following indicators appeared to be more challenging for examiners: 'key features covered', 'key features vs. details' and 'details vs. minor details'.

Qualitative data revealed that certain aspects of the TA band descriptor wording were repeatedly identified by examiners as challenging, possibly due to perceived ambiguities. Some examiners found it harder to differentiate between descriptors in the top bands (supporting findings from an earlier internal survey conducted in 2012).

Thematic evidence highlighted some difficulties for examiners associated with: wording about addition of relevant information to the response (Band (B)9); the notion of when/whether something is 'fully developed' or 'covering the requirements of the task' (B7, 8 & 9); identifying what the key features are (B7 & 8 – also at B3, 4 & 5); defining what constitutes details vs minor details (B6); and overviews (B5, 6 & 7 AC).

Task Response (TR)

A similar trend was evident in the quantitative data for TR: the higher, less frequent B8 and B9 were seen by examiners as somewhat more challenging, with only 37% of examiners feeling confident applying B9 and 44% B8. This was in marked contrast to the high level of confidence in applying most of the other bands (ranging from 64% to 84%).

Survey questions for TR also focused on examiner confidence in applying band descriptors/indicators and indicator clarity. The 'development of position' indicator attracted the lowest rating of 60%. Three other indicators also emerged as somewhat less straightforward, with examiner confidence only reaching two thirds: 'main ideas-relevance' (65%), 'clarity of position' (66%), 'main ideas – support' (67%).

Certain aspects of the TR band descriptor wording were repeatedly highlighted. Many examiners apparently struggled to differentiate between B7, B8 and B9, and largely attributed this challenge to the following points: the difference between a prompt being 'fully explored' (B9), 'appropriately and sufficiently addressed' (B8) or 'appropriately' addressed (B7) and support/main ideas (B6, B7, B8 & B9); 'nothing can reasonably be added' (B9) – something can always be added (same as TA above); addressing the prompt (all bands, but particularly higher ones – B7, B8 & B9); the difference between 'a clear and fully developed position' (B9), 'a clear and well-

developed position' (B8) and 'a clear position' (B7); and the provision of evidence (B4, B7, B8, B9).

Coherence & Cohesion (CC)

Only 43% and 52% of examiners felt confident in assessing B9 and B8 respectively, compared to confidence ranging between 62%–78% for the other bands. This was slightly higher than TA and TR (discussed above), but still lower than might have been expected.

The indicators with the lowest percentage of examiner confidence were 'progression', 'overuse/underuse of cohesive devices' and 'sense of coherence', with just under two thirds of examiners in all three cases feeling 'confident' or 'very confident': 61%, 64%, and 65% respectively.

In the open-ended survey responses and focus group discussion, certain aspects of the wording of the CC band descriptors were identified as challenging, albeit to a slightly lesser extent compared with TA and TR. Examiners were again struggling to differentiate between B7, B8 and B9, and largely attributed this to the following points: the difference between 'the message can be followed effortlessly' (B9) and 'the message can be followed with ease' (B8); the difference between 'cohesion is used in such a way that it attracts no attention' and 'all aspects of cohesion are well-managed'; managing cohesion and cohesive devices (B6, B7 and B8); paragraphing, 'skilfully managed' (B9) vs sufficiently and appropriately' (B8) and the paragraphing ceiling (B8); no clear progression (B4, B5, B6 & B7) and logic (B7 and B8).

Lexical Resource (LR)

Confidence applying band descriptors for LR was higher than in TA, TR and CC, with confidence ranges in the 60%–84%. The lowest values were observed for B8 (65%) and B9 (60%).

The indicators which showed the lowest degree of examiner confidence in applying them were 'impact of errors' (65%), with several other indicators associated with relatively low degrees of confidence as well: 'adequacy of resource' (67%) and 'frequency of errors' (67%).

LR was not a major element in any of the focus group discussions or survey responses. However, certain aspects of the wording of the LR band descriptors were identified as challenging. Examiners were again struggling to differentiate between B7, B8 and B9, and largely attributed this to the following points: distinguishing between errors– extremely rare (B9), rare (B8) and occasional (B7); defining what constitutes the various forms of 'flexibility and precision' at each band (B7, B8 and B9); impact of

errors (difficulty vs strain, spelling, accuracy) (B4, B5, B6, B7); and the concept of a 'risk-taker' (B6).

Grammatical Range & Accuracy (GRA)

Examiner confidence in applying the GRA band descriptors across bands was lowest at B8 and B9, with fewer than two thirds of examiners expressing confidence: B9 (55%) and B8 (61%). For the other bands confidence ranged between 67% and 93%.

The GRA indicators with the lowest degree of examiner confidence were 'range' (65% and 64% respectively were 'confident' or 'very confident') and 'flexibility'.

GRA was not a large part of extended survey comments or focus group discussions. Even though this scale was perceived as functioning adequately for the most part, some comments of note did emerge. In the findings from the qualitative data, examiners repeatedly mentioned the following points: errors – determining 'frequency' (B6, B7, B8 and B9) is not straightforward and 'non-systematic errors' (B8) is a difficult concept; distinguishing between 'full flexibility and control' (B9) and 'flexibly and accurately used' (B8); and a max/range/variety of structures (B6, B7, B8 and B9).

Instructions to Examiners (ITE)

The ITE booklet was mentioned throughout the data described above. Respondents either referred to the existing help in the booklet (commenting that it was more or less useful to them in its current form) or expressed the view that some aspect could be added, in order to make it more straightforward to apply some of the more challenging aspects of the descriptors already described. In some cases, examiners admitted that the instructions required were there, but they did not always refer to them, perhaps because they needed to be more efficiently signposted in certain cases. Although there was a relatively even spread between categories, it may be the case that Key Features (15% of respondents felt the instructions on this were 'not' or 'somewhat' helpful) and Overviews (14%) require further attention in particular, which supports the survey data described above. Some revisions to the ITE were made in 2018.

Computer-delivered (CD) IELTS issues: short scripts, further ITE improvements, spelling and typos

Marking of the CD scripts was a recent development in IELTS and the band descriptors and ITE (along with other training and support documents available to examiners) were produced with paper-based (PB) scripts in mind. Both survey comments and focus group discussions highlighted additions that were needed to the existing ITE resources to help examiners address challenges arising with CD scripts. These

included how to rate under-length responses and how to deal with spelling errors/typos, punctuation, paragraphing, spacing and indentation in the CD format.

Recommendations from Phase 1

The Phase 1 study indicated that, overall, the scales were performing well, and that no wholesale changes were required. At the same time, lower-than-desirable confidence levels on certain points suggested a need to remove ambiguities in existing scale wording and a number of recommendations were made for next steps.

High-level recommendations: Initiate a second project phase (Phase 2) to explore revising the wording for some band descriptors. This would first identify the priority issues to be addressed, and the means for doing so, i.e. through revised scales, or training and standardisation scripts. Consult with a range of IELTS experts – both internal and external. Use score data as statistical confirmation of issues identified and points for revision. Validate the revised scales through trialling and the collection of quantitative and qualitative data from a range of sources, including surveys, focus groups and scores.

Micro-level recommendations: Use the detailed Phase 1 feedback on individual scale categories and band descriptor wording to input to the commissioning and drafting of revised scales in consultation with IELTS experts with a view to future field trialling (Phase 3).

PHASES 2 AND 3: DEVELOPMENT AND VALIDATION (Sept 2019–Feb 2020)

Original report authors: Clark, T., Tasviri, R., Schmidt, E., Galaczi, E., Dunlea, J., Spiby, R, Westbrook, C., & Dunn, K. (2020).

Introduction

This section of the summary overview reports on the empirical work that took place in Phases 2 and 3. The information obtained in Phase 1 (see previous section) formed the basis for targeting adjustments to the wording of the IELTS Writing scales and band descriptors in Phase 2, and provided a baseline for comparing examiners' perceptions of the adjusted scales in a Phase 3 field trial.

Methodology

Research design and research questions

The Phase 2/3 study adopted a multi-phase *mixed methods design* (Creswell, & Plano Clark, 2011). In Phase 2 the findings and insights from Phase 1 were combined with expert input to make written adjustments to the band descriptor wording. In Phase 3 the adjusted band descriptors were used in a field trialling exercise with IELTS examiners. A survey instrument and focus groups were used to gather both quantitative and qualitative and data in order to answer the following RQs:

RQ1: Have the adjustments to the scales affected the overall levels of examiner confidence?

RQ2: Do examiners feel that the changes have made the scales clearer than before?

RQ3: Have any remaining ambiguities or unclear points been identified?

RQ4: What is the change in band scores awarded by raters using the adjusted scales?

RQ5: How do raters perform when using the adjusted rating scales in terms of rater severity and consistency?

RQ6: How do the adjusted rating scales function in terms of empirical difficulty, consistency and scale step differentiation?

Data collection: sources and participants (Phase 2)

Phase 2 of the IELTS Writing Scale Review project involved a cross-partner Working Group to prioritise critical aspects for attention and to make the necessary adjustments by rewording the existing scales. Based on the findings of the earlier report (Clark, Tasviri, Lopez, & Galaczi, 2019) and lengthy practical experience in the examining field, the group's task was to determine the critical areas to be addressed.

These emerged after a series of meetings in which each potential point of revision was discussed in depth, and rewording of the scales (including modifications of, additions to and subtractions of existing wording) were subsequently made. In between meetings, teams of senior examiners were informally consulted to ensure that their perspectives were included beyond the initial data collection in Phase 1. This iterative and in-depth process ensured that sufficient rigour was employed in making decisions.

Most of the adjusted descriptors for IELTS Writing Task 1 and Writing Task 2 were located around the top bands, which had been identified as particularly challenging in the Phase 1 examiner consultation. Adjustments to TA and TR were most numerous, with fewer changes made to CC, and fewer still to LR and GRA. Adjusting the 'key features' element of the band descriptors was a significant part of the changes made, in addition to rewording the top bands (B8/B9) in an attempt to encourage greater use of them. It was decided that any remaining – less critical but still important – areas could be addressed through revisions to the ITE Booklet.

Two draft documents were produced – one for each set of reworded band descriptors for Writing Task 1 and Task 2 (both Academic and General Training) – and the iterative consultation approach to the adjustments continued. In September 2019, three highly experienced principal examiners received a document explaining the project, detailing the changes made to the scales and the rationale behind each change. They were also given a copy of the report by Clark et al. (2019). In addition to looking at the draft adjustments made by the Working Group, the principal examiners were asked to use the draft of adjusted band descriptors to mark a series of nine scripts, which had been pre-scored by other senior examiners, across the bands, for both Task 1 and Task 2. This process was completed online. The principal examiners were then asked to respond to a series of general questions about the adjusted scales, and to comment in detail on each aspect. These documents were sent back to the Working Group for analysis, and individual follow-up meetings conducted with each principal examiner.

Once finalized, the draft adjusted scales were ready for trialling with a larger examiner cohort. A *mixed methods design* was chosen to gather quantitative and qualitative data which would provide detailed insights from multiple perspectives into how the adjustments to the scales functioned in practice.

Data collection: field trialling (Phase 3)

Phase 3 of the review project involved a range of participants selected to provide different perspectives and experience to inform the questions of interest. Participants comprised: 103 examiners who provided scores on 884 responses using the adjusted scales; 91 examiners who completed the online survey (out of the 103 examiners who scored); and 11 examiners (a sub-set of the survey/scoring sample) who participated in three focus groups. Participating examiners were drawn from the IDP and British Council cadres to ensure cross-partner representation. All examiners were fully trained and certified and had responded to an invitation to participate, sent out to the entire population of British Council and IDP examiners. They represented a broad range in terms of gender, partner affiliation and years examining.

Script selection: A total of 502 writing scripts were initially selected across IELTS Bands 4 to 9, including at half-integer bands. Scripts included both PB and CD test versions, with both Task 1 and Task 2. At each band and half-band up to level 7.5, 48 scripts were selected, with roughly half the sample comprising Academic (Ac) and half General Training (GT) scripts. Additionally, at these specific bands scripts from the GT module were absent so most scripts included at these bands were Academic. No scripts at Bands 1–3 were included due to the low incidence of candidates at these levels and the non-critical function of these bands in IELTS test use contexts. All scripts were confirmatory marked. The majority of scripts had been used in an internal previous IELTS validation study, for which they had gone through confirmatory marking. Other scripts had had their marks confirmed for standard setting purposes; the remaining scripts were sent for confirmatory marking by three principal examiners. For scripts to be included in the study, scores from examiners could not differ by more than half a band. The range of variability in the old scores is therefore very narrow and does not reflect normal variability found between examiners. Additionally, half the sample was chosen as borderline scripts, while the other half fell squarely within the band/half-band.

Script allocation: Each of the 110 recruited examiners was allocated to one of 15 groups, with six or seven examiners per group. The 442 scripts were divided into 16 batches – a common batch for all examiners and a batch for each group, with each batch consisting of 28 or 29 scripts, and a common link between each batch (e.g., Script 29 was common to Batches 1 and 2). Each batch included scripts across the range from Bands 4 to 9, including borderline and middle-of-the-band scripts. Each examiner group was assigned two batches – the common batch and the individual group batch, resulting in each examiner marking 48 or 49 pseudo-randomised scripts and each script getting marked at least six times.

Score collection: Each examiner was sent their allocated scripts and an Excel spreadsheet in which to record their assigned scores and add comments if necessary. Before marking the assigned scripts examiners went through a self-access familiarisation process with the revised scales and adjusted descriptors which involved a small number of scripts, associated marks (given by three principal examiners during Phase 2 of the project) and commentaries providing a rationale for the marks. On completion of marking, they e-mailed their completed Excel spreadsheets.

Data analysis: examiner scores from field trialling (Phase 3)

Linear Mixed Effects (LME) models were chosen as the most suitable method for this analysis because they allow for *examiner* and *response* to be included as random factors. This makes it possible to account for the effects that individual examiners and individual scripts can have. Since not all examiners marked all scripts, it is possible that some scripts were more straightforward to mark, perhaps because the revised scale descriptors were not applicable to those scripts. Equally, the person marking the script can have an effect on the final outcome. Adding examiner as a random factor made it possible to account for variability introduced by individual examiners.

LME models were used for the averaged overall band score and each of the averaged criterion scores. The Writing band score or the respective criterion scores (*Task Achievement, Task Response, Coherence and Cohesion, Lexical Resource, Grammatical Range and Accuracy*) were included as response variables, and *Outcome* ('new' scores higher, same, or lower than 'old' scores, where 'new' refers to the adjusted scales and 'old' to the operational scales), *Module* (Academic vs. General Training), *Experience* (examiner experience in years), and *Confidence* (confidence applying the 'new' scales for each criterion, averaged across bands, on a 1–5 Likert scale) as predictor variables. Interactions between *Outcome* and *Experience*, and *Outcome* and *Confidence* were also included.

Multi-Faceted Rasch Measurement (MFRM) was used to place the variables of interest, or facets, onto a common measurement (logit) scale, using the FACETS program. According to the script allocation described above, each examiner group rated a common batch of scripts in addition to a unique set of scripts, enabling robust linking in the data set across all examiners so that relevant facets could be analysed on a single scale. Using MFRM with the FACETS program, it was possible to take account of both the relative severity of raters and the difficulty of items in the final estimates of test-taker ability. These estimates are referred to as Fair Averages, since elements of the target facet are calculated according to the averaged values of all other facets. The Fair Average thus provides an adjusted estimate, whereas the Observed Average is

simply the average of all the 'raw' ratings within each facet. FACETS also allows the inclusion of 'dummy facets' within the analysis, which do not contribute to overall measurement, but are used to explore interaction effects between these variables and ratings. FACETS provides a useful quality assurance measure of rater and item consistency, the infit mean square statistic. This indicates to what extent the data fit the Rasch model, with greater sensitivity to inlier responses. A higher fit statistic represents underfit, or unpredictability. Levels greater than 1.5 indicate that those raters are not rating the items in the same relative order of difficulty, although fit statistics in the range of 1.5 to 2.0 are still considered useful in some contexts and are not degrading to measurement. Underfit is usually considered more problematic than overfit, or low infit (less than 0.5), which represent response patterns that are too predictable, yet are not degrading for measurement.

A 5-facet analysis was performed with candidates, raters and items (i.e., rating scales) – the key variables contributing to measurement. Institution (British Council or IDP) and delivery mode (CB or PB) were included as 'dummy facets' to investigate potential bias interactions with ratings. A total of eight rating scales (the 'items' facet) were specified for each of the Ac and GT modules. Despite overlaps in scale descriptor categories, all rating scales were treated as discrete elements to account for some differences in wording and potential for differences in scale performance between Task 1 and Task 2. Similarly, the Ac and GT modules were analysed separately, since linking occurred through common raters rather than common candidate performances across modules. Consequently, for scale analysis, evaluation of item and rater performance and bias interactions, Tasks 1 and 2 for Ac were investigated in one analysis, and Tasks 1 and 2 for GT were dealt with in a parallel analysis. In order to compare the ratings with old and new scales, the raw component scores as established in confirmatory marking with principal examiners were totalled, weighted and rounded according to scoring procedures in place. To find the new scores, separate analyses were conducted for Tasks 1 and 2 for Ac and GGT, respectively. The Fair Averages for each task performance were then totalled, weighted and rounded in the same way. It was then possible to estimate the differences in overall writing band score for each candidate taking the relative severity of raters into consideration, producing a more accurate picture of potential shifts in scoring behaviour as a result of using the adjusted scales.

Data collection: survey of examiner perceptions (Phase 3)

The survey (Survey 2, November 2019) was designed to be as similar as possible to the one used during the examiner consultation stage (Survey 1, March 2019), in order to ascertain how examiners' perceptions of using the 'new' as opposed to 'old' scales had changed. As before, the SurveyMonkey platform was used. There were 30 items, a mixture of selected response (24 questions) and open response (six comment boxes to elaborate on answers given). Participating examiners were asked to complete the survey immediately after finishing the marking, so that their impressions did not diminish as time passed. Demographic data on participants was not the focus of this survey and asking for as few details as possible about respondents was intended to make them feel comfortable with sharing their views on the changes as candidly as possible.

Data analysis: survey of examiner perceptions (Phase 3)

In order to understand the difference between examiner perceptions about the functioning of the scales in Surveys 1 and 2, descriptive and inferential statistics were generated in SPSS (version 25). In total, 164 items were compared across the two surveys for the selected-response questions. Summary tables and graphs were generated for each of the response options, capturing the percentage endorsing each response option, and the mean/SD/min/max, depending upon the item. The Mann-Whitney U test was used to explore the differences between the two surveys across the nine bands and across the five criteria. For the qualitative analysis, Braun and Clarke's (2006) model of thematic analysis was used. Open-ended comments were analysed for key themes using NVivo Pro (version 11). In order to develop a uniform and consistent coding approach and to capture the main themes in the examiner comments adequately, two researchers independently coded the survey data and subsequently compared notes to verify that the emerging themes of relevance to the RQs overlapped. When the codes were compared, they were found broadly to agree. In cases where one theme was not exactly the same between the two sets of researchers' notes, a discussion brought them into line. As part of this process, any smaller themes that were deemed to be irrelevant or less than critical to the RQs were discarded, due to the large dataset involved and the need for the analysis to remain highly focussed.

Data collection: focus groups with examiners (Phase 3)

Focus groups were conducted in January 2020 with a subset of examiners who had participated in the field trial and applied the adjusted band descriptors. The focus groups were online and semi-structured, and focused on probing areas in the scales that had been shown to require improvement or further clarification. Three focus

groups were conducted with 12 examiners from the three IELTS partners located in New Zealand, Australia, UK and Canada. Participants represented different levels of experience and seniority (examiner, senior examiner and principal examiner) to reflect potentially differing perceptions about the clarity of the adjustments depending upon seniority. Sessions were conducted using Zoom, lasted 60–90 minutes, and were recorded for analysis. Questions for the focus group sessions focused on adjustments in the band descriptors around which the trial examiners had expressed lack of clarity or felt were challenging to apply.

Data analysis: focus groups with analysis (Phase 3)

Thematic analysis was used to extract high-level findings. Three members of the research team – one from each partner organisation – coded the focus group interviews according to comments made relating to the themes. For each interview, at least two of the three researchers were involved in the coding, with some interviews coded by all three researchers.

Results

Examiner perceptions of the scale adjustments

The RQs of interest here were:
RQ 1: Have the adjustments to the scales affected the overall levels of examiner confidence?
RQ 2: Do examiners feel that the changes have made the scales clearer than before?
RQ 3: Have any remaining ambiguities or unclear points been identified?

Overall, the adjustments to the band descriptors were well received by examiners. At Bands 8 and 9, the 'new' adjusted descriptors were perceived as more effective than the 'old' descriptors, as seen in improved examiner confidence levels. Additionally, some specific wording was flagged up as needing further attention and clarification through training, adjustments to the ITE, worksheets or tweaking of the modified wording in order to help examiners apply the modified scales.

Score changes using the adjusted scales

The RQ of interest here was:
RQ 4: What is the change in band scores awarded by raters using the adjusted scales?

This question was explored through two different but complementary analyses: Multi-Level modelling and Multi-Faceted Rasch Measurement [MFRM).

Multi-Level modelling: Results showed that the changes to the Writing band descriptors had a minimal effect on the scores. The values for overall band scores, as well as the eight criterion scores (the four criterion scores were analysed separately

for Task 1 and Task 2), increased slightly. Importantly, the experience of examiners (the number of years they had examined IELTS papers) did not affect the scores.

Multi-Faceted Rasch Measurement (MFRM): Results confirmed those produced by the Multi-Level modelling, with 63% having no change or a slight increase (0.5 bands) in score, and just under 20% experiencing a drop of 0.5 bands.

Use of MFRM allowed investigation of two other RQs:

RQ5: How do raters perform when using the adjusted rating scales in terms of rater severity and consistency?

The relative severity or leniency of the raters was evaluated according to their rater severity measure on the logit scale. The amount of variability in rater severity was found to be within an accepted range of +1/-1 logits for more than 95% of raters. For Ac, only three raters and for GT, only four raters out of a total of 103 fell marginally outside this range. This indicates relative consistency in rating 'to standard', suggesting that training received by raters in the new scales was adequate and that adjustments to the scales were not disruptive to the marking process. For Ac, 79% of examiners fell within +/- 0.25 logits of the median Fair Average severity of 6.24 and for GT 71% of examiners fell within +/- 0.25 logits of the median Fair Average severity of 6.37, indicating a very tight range of severity for the majority of the rating cohort in this study. In terms of their level of fit to the model, overall rater performance was strong, with over 95% of raters displaying 'good fit' within the range 0.5–1.5 Infit Mean Square. For Ac, 99 raters and for GT 98 out of 103 raters were within this category. There was no underfit on either module and only limited rater overfit (Infit Mean Square >1.5), indicating that only a small number of raters were not marking consistently. Only three of these raters might be considered a matter for concern, displaying values slightly greater than 2.

The MFRM results thus indicated a high degree of consistency in the allocation of scores according to candidate ability level, supporting the interpretation that this group of operational raters demonstrated acceptable levels of quality in using the adjusted scales. Bias analyses were also conducted in the analysis, with delivery mode and rater institution included as 'dummy facets' to investigate possible interactions between these and rater behaviour. In terms of the institutional background of the raters, the bias analysis indicated that the performance of British Council and IDP raters was very similar.

RQ6: How do the adjusted rating scales function in terms of empirical difficulty, consistency and scale step differentiation?

The criteria associated with Task 1 showed a higher level of difficulty than those for Task 2, across both Ac and GT. Consultation with the project team indicated that this trend was consistent with rating with the current, unadjusted scales, and was not a result of the introduction of adjustments. Some misfit (>1.5) was observed on the TA for Ac and both TA and TR scales for GT, indicating some inconsistency. However, these values were not excessive, being still below 2. Changes to rater training were expected to have a positive impact on this in the future.

Recommendations from Phases 2 and 3

The integration of qualitative and quantitative data collected through this project suggested that the adjustments made to the IELTS Writing scales and piloted as a part of the project would be beneficial and the project team therefore recommended implementation based on the evidence from the survey and focus group data and the score data. A number of additional recommendations were made concerning planned implementation of the adjusted scales, including: the need for face-to-face and self-access training and standardisation materials, including information on the rationale for the adjustments; cross-partner examiner training and standardisation to ensure uniform and consistent standards; recalibration of existing reference items; and continued involvement by the research team to ensure ongoing operational monitoring of the adjusted scales post roll-out.

PHASE 4: IMPLEMENTATION (Jan 2022–May 2023)

Original internal report authors: Lee, D. & Boden, R. (2023)

Introduction

The first version of the revised band descriptors for IELTS Writing was signed off by the IRG in February 2020, following a comprehensive programme of review, revision and field trialling (see Clark et al., 2019, and Clark et al., 2020). Although implementation was recommended, the follow-up phase had to be put on hold due to the impact of the Covid-19 pandemic. In January 2022 the project was restarted by the IELTS partnership and the go-ahead for implementation was given as Phase 4 of the project.

During the period May–July 2022, a cross-partner working group of subject matter experts further refined the wording of the descriptors and the advice contained in the IELTS Examiner Booklet (IEB), and produced a training package and materials designed to familiarise examiners with the revised descriptors. As part of the implementation efforts, a new *marking exercise* was included to assess the efficacy of

the training package, to support implementation and to complement the findings of the original 2020 marking trial. In addition, examiners were asked to complete a *feedback survey* to gather their perceptions on the training and revisions to the descriptors and the IEB.

The aim of the Phase 4 study was to investigate the impact of the revised descriptors on the marks awarded to a set of Writing responses, and to identify differences in the functioning of the criteria scales which could inform ongoing examiner monitoring actions post-implementation. The scope of the study was limited to an analysis of the raw scores at the response level (Task 1 or Task 2). The analysis reported the direction (increase or decrease) and significance of any observed score changes. The MFRM analysis of the key scoring criteria (TA, TR, CC, LR and GRA) focused on any changes in relative difficulty, fit to the model, and band score distribution. Conclusions and recommendations were presented in the context of implementation, aiming to provide useful advice on short and medium-term monitoring of examiner performance and, in combination with the qualitative data gathered on examiner perceptions, to guide the development of supplementary examiner support materials on the revised descriptors.

Methodology

Marking exercise

The design of the marking exercise aimed to allow for as direct as possible a comparison between how marking of the IELTS Writing tasks changed when using the current and the revised descriptors. To enable a direct comparison, a common set of scripts were marked twice by the same group of examiners, with a break of seven weeks between marking rounds. 1,000 responses in total were selected for marking and a response was defined in this study as one Task 1 or Task 2 paper. 500 responses were selected by each IELTS delivery partner (i.e., British Council and IDP). The responses were selected to cover as much of the IELTS scale as possible, with a focus on responses covering the levels where score changes were most likely to be observed. Script selection included samples of both Ac and GT Task 1 and Task 2.

The responses were arranged into 12 packs of 100 and distributed across the participating examiners. Each pack was assigned to a minimum of eight examiners, resulting in all scripts being marked a minimum of eight times in each round of marking. In addition, to ensure a fully linked rating design, 18 common anchor scripts were included in all 12 packs. The rating plan involved 100 examiners (50 from each delivery partner), with each being assigned a 100-script pack.

A mixed-model approach to the analysis was used to interpret the data. A comparison of the raw marking data was undertaken to identify if and how the marks changed from one round to another across the individual responses, by task type and criteria. The data were analysed using MFRM with the Facets software package in order to compare the relative difficulties of the criteria (conceptualised as 'items' in the Rasch model) and their scale structures. The Rasch model was also used in this study to examine the rater facet (the examiners) to validate fit to the model and identify any relevant anomalies that might impact measurement of the functioning of the revised descriptors upon implementation. The facets included in the analysis specifications were: *Script*, *Task*, *Module*, *Examiner*, *Organisation*, *Criteria*.

Examiner feedback survey

Upon completion of the session introducing the planned revisions to the writing band descriptors, 99 examiner participants were asked to complete an online questionnaire, comprising 20 questions, 13 of which were Likert-scale, four were open-ended, and the remainder were closed. Responses were received from 52 British Council and 47 IDP examiners.

Results

Result from the analyses broadly aligned with the original aims of adjusting the descriptors and making additions to the IEB. Changes in scores observed between the two rounds of marking were almost exclusively increases. One desired outcome was for examiners to gain the confidence to award the higher ratings when merited by increasing the clarity of the relevant wording in the descriptors. The results of the analysis suggested that the rating of individual responses would not increase by more than 1 criterion point. In terms of the functioning of the test and the key criteria, the results were pleasing. Fit improved in most cases, indicating a more predictable, consistent marking pattern. In terms of the test's ability to discriminate between the categories (band scores), the revised descriptors functioned well, although in the case of Task 1 Ac, the category probability curves for CC, LR and GRA were not quite as well ordered as the current descriptors, something that could be addressed by training and further familiarity with revised descriptors. For the other tasks, there was little discernible difference. The raters appeared to adapt to the revised descriptors well. Rater fit in the second round of marking was very good, and the spread was tighter. Examiner feedback via the post-training survey was positive, with examiners expressing increased confidence in using the revised descriptors.

Recommendations from Phase 4

Before implementation – and to mitigate some the potential issues raised in the above sections – raters were given training on the use of the new band descriptors.

Implementation

The results of the marking exercise supported the conclusions of the initial trial study, which was to proceed with implementation. The inclusion of a comprehensive, standard training exercise between the first and second rounds of marking produced positive outcomes. The revised descriptors generally produced a small, tolerable increase at the response level, though short and medium-term monitoring of scores was recommended to track the impact on overall test-taker writing scores post-implementation. There were some unexpected observations, such as the changes in ratings observed with the Task 1 GT criteria, and perhaps the slightly unusual category probability curves for some criteria in Task 1 Ac. However, it must be kept in mind that this was the first time examiners had used the revised criteria, and further training and support would help to refine the cohort's interpretation of the descriptors.

Conclusion – ongoing monitoring and support

It was acknowledged that examiner support materials would be required post-implementation. The set of responses used in this study was an ideal bank from which to source suitable responses to be further developed into self-access or trainer-facilitated materials to support examiners in the correct interpretation of the revised descriptors.

The changes implemented to the IELTS Writing scales will be kept under review and further adjustments made as deemed necessary in light of ongoing monitoring. This will involve periodic revisitation of their use beyond score monitoring in isolation, including at a qualitative level (for example, interviewing raters on their perceived use of the adjusted scales after the changes have 'bedded in' is also planned). The ITE document (used to assist raters in their tasks) and the training they receive will be updated as required.

REFERENCES

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Clark, T., Tasviri, R., Lopez, S., & Galaczi, E. (2019). *IELTS Writing Scales Review: Phase 1*. Internal technical report.

Clark, T., Tasviri, R., Schmidt, E., Galaczi, E., Dunlea, J., Spiby, R, Westbrook, C., & Dunn, K. (2020). *IELTS Writing Scales: Developing and validating adjustments to band descriptors*. Internal technical report.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research* (2nd edition), Thousand Oaks: Sage Publications.

Falvey, P., & Shaw, S. D. (2006). IELTS Writing: Revising assessment criteria and scales (Phase 5). *Research Notes*, *23*, 7–12.

Galaczi, E., Lim, G., Khabbazbashi, N., & Vice, M. (2012). *IELTS Speaking and Writing Assessment Scales Review*. Cambridge: Cambridge English internal report.

IELTS. (2010). *Writing Test: Instructions to Examiners*. IELTS.

Lee, D., & Boden, R. (2023) *Marking Exercise Implementation Report*. IELTS.

Shaw, S.D., & Falvey, P. (2008). The IELTS Writing Assessment Revision Project: Towards a revised rating scale. *Research Reports, 1*, 1–295.

Weir, C. J., & O'Sullivan, B (2017). *Assessing English on the global stage: The British Council and English language testing, 1941–2016*. British Council & Equinox Publishing.