Article

LANGUAGE
TESTING

Language Testing
28(4) 543–560
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532211406422
ltj.sagepub.com
SAGE

The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters

Gad S. Lim
University of Cambridge ESOL Examinations, UK

Abstract

Raters are central to writing performance assessment, and rater development – training, experience, and expertise – involves a temporal dimension. However, few studies have examined new and experienced raters' rating performance longitudinally over multiple time points. This study uses operational data from the writing section of the MELAB (n = 20,662 ratings), an international exam of English proficiency, to investigate the rating quality of new and experienced raters over three time periods of 12 to 21 months. Rating quality was operationalized in terms of rater severity and consistency, and estimates of those modeled using multi-facet Rasch methodology. Results indicate that, within one particular rating context, (1) novice raters, where initially differing in performance, learn to rate appropriately relatively quickly, (2) raters are able to maintain rating quality over time, and (3) rating volume and rating quality may be related. Implications for rater preparation, rater certification, and the notion of expert rater are discussed.

Keywords

inter-/intra-rater reliability, multi-faceted Rasch analysis, rater development, rater expertise, rater training, writing assessment

Raters are central to the enterprise of performance writing assessment. Lumley (2005) lays out the process of this enterprise. On one end are test-takers' writing samples, which can be said to exhibit disordered complexity, as they are not in a form from which inferences can be drawn. On the other end is the institutional goal of measurement; what is desired is something standardized and reliable, a score, which the institution can use as a

Corresponding author:
Gad S. Lim, 1 Hills Road, Cambridge CB1 2EU, UK.
Email: lim.g@cambridgeesol.org

basis for making decisions. Raters occupy the middle position, the ones whose judgments turn performances into outcomes – the reason why raters are so consequential in this type of assessment. More than anything, raters need to be able to provide ratings that are appropriate, and to do so consistently. As Shaw and Weir (2007) put it, 'scoring validity is criterial because if we cannot depend on the rating of exam scripts, it matters little that the tasks we develop are potentially valid' (p. 143) in all other respects.

In aid of scoring validity, test programs typically make provision for rater training, which usually includes familiarization activities, practice rating, and feedback and discussion (Lane & Stone, 2006). From the literature on writing assessment, it appears that training tends to be relatively short, lasting no more than half a day in most cases (e.g. Congdon & McQueen, 2000; Weigle, 1998). Some studies indicate that this is sufficient for a difference in rating quality to be observed. In one such study, Shohamy, Gordon, and Kraemer (1992) compared 10 trained and 10 untrained raters' ratings of 50 compositions on three scales/criteria: holistic, communicative, and accuracy. The results showed that while inter-rater reliability coefficients were on the whole generally high, the trained raters were more reliable than the untrained raters (0.91–0.93 vs. 0.80–0.90). Other studies clarify that training helps not so much with inter-rater agreement as it does with intra-rater consistency (Engelhard, 1992; Weigle, 1994). Using a pre- and post-test design with one to three weeks in between, Weigle (1998) showed that training reduced but did not eliminate differences in severity between experienced and inexperienced raters. The consistency of inexperienced raters, however, showed much improvement after training. Taken together, these studies suggest that rater training is effectual in some limited fashion, at least in the period soon after the delivery of training.

Closely related to training are the notions of rater experience and rater expertise. Experience is being used in this paper to refer either to the length of time a rater has been rating or to the amount of rating a rater has done. Expertise, for its part, refers here not to people with particular competence in a subject area, for example Cumming, (1990), but to raters whose marking performance is consistently good. While experience and expertise are likely related (cf. Weigle, 1998), they are distinct; it is hoped that experience will lead to expertise, but that may not necessarily be the case. In a study by Wolfe, Kao, and Ranney (1998), raters were classified as competent, intermediate, or proficient, depending on how highly each individual rater's ratings correlated with those of the other raters. Using think-aloud protocols, they found that there were qualitative differences in the way these groups of raters rated. Proficient raters experienced fewer interruptions while reading and were able to withhold judgment until after they had finished reading. These raters also made more general comments, rather than comment on specific pieces of text. Finally, proficient raters considered all criteria equally and used more rubric-related language in justifying their ratings. Thus, while all raters in their study were trained and experienced, there remained qualitative differences in their approach to the rating task that affected the degree of their expertise. Huot (1993) and Barkaoui (2010) also arrived at similar findings. Other factors such as the language, educational, and professional background of raters and of those whose writing they rate may also have an effect on rater expertise, as does the context in which rating is done, and need to be kept in mind (Barkaoui, 2007; Cumming, Kantor, & Powers, 2002; Fox, 2003; Johnson & Lim, 2009; Pula & Huot, 1993; Santos, 1988; Shi, 2001).

In any event, rater development – training, experience, expertise – involves a temporal dimension, yet there is a paucity of longitudinal studies beyond short-term pre- and post-test designs. In the absence of such studies, it is difficult to ascertain the persistence of training effects, the consistency of raters over time, or how raters develop from novices into experienced and/or expert raters. Congdon and McQueen (2000) and Myford and Wolfe (2009) both examined rater behavior over multiple rating sessions. However, each study spanned only a relatively brief period of time – nine days in the case of the former and four days in the case of the latter. Congdon and McQueen's sample of raters also included novice raters, but these raters were not identified or separately analyzed in their study. Cho (1999) conducted a study where raters re-rated compositions four times over a longer period of time, after periods of four to six weeks. Most raters appeared to have high levels of internal consistency. However, because those raters were rating the same compositions multiple times, the study is confounded by possible memory effects. To date, Lumley (2005) provides the best picture of rater behavior over time, providing quarterly estimates of rater severity and consistency for four raters over a period of two years. His study showed that the raters performed with remarkable stability over time, maintaining their relative rank-order throughout. However, the raters in Lumley's study were chosen for their experience, and it is unclear whether the findings hold for inexperienced raters.

In sum, there is at present no study which sheds light on the initial and long-term development of novice raters into experienced and/or expert raters. Studies that address these would be useful as they can provide insight into questions such as: At what point are people ready to act as raters? How long can raters continue as raters (e.g. in contexts where raters are certified for a period of time) with no additional training, intervention, or support? Do training effects last? On a more theoretical level, such studies can illumine the possible category of 'expert rater'. Apart from any other definitional requirements, the category of 'expert rater' depends on the ability to maintain a certain level of rating quality over time. If it is shown that raters' performance varies widely all the time, then that category cannot be said to exist in any meaningful way. On the other hand, if it is shown that rating quality can be maintained by at least a few raters over time, then the category potentially exists, and a host of other questions can then be asked: How are these raters different from other people? What characteristics do they possess? How does one become such a rater? An understanding of how raters' rating quality develops and maintains over time can clearly benefit the theory and practice of writing performance assessment. This longitudinal study of novice and experienced raters aims to contribute to that understanding.

Context of the study

The testing context of this study is the writing section of the Michigan English Language Assessment Battery (MELAB). The MELAB is an international examination of English proficiency used for various high-stakes academic and professional purposes (for reviews of the test, see Chalhoub-Deville, 2003; Purpura, 2005). The writing section is a timed, impromptu writing test (Hamp-Lyons, 1991; Weigle, 2002); examinees have 30 minutes to write a composition on one of two prompts, which they do not see in advance. The prompts can call for narrative, expository, or argumentative modes of writing. In the time period covered by this study, 60 different prompts were used. The characteristics and comparability of the prompts are detailed in Lim (2010). Examinees' compositions are independently read and rated by two raters using a holistic, 10-point scale. Where the two ratings are more than one scale-point apart, a third rater is introduced. The final score is the average of the ratings that are either equal or different by one scale point (English Language Institute, 2005). As examinees are allowed to request a rescore, there are potentially up to six ratings for each test taker.

The MELAB is unique in that while it is a large-scale exam, it is also at the same time small enough to allow all compositions to be rated by a small team of in-house raters. Unlike other similar exams, whose raters could be freelance workers working from diverse locations, all MELAB raters are language testing professionals and regular employees of the English Language Institute, University of Michigan, who work out of a single location and who interact with each other on a daily basis. This setup may well have an effect on rater behavior (e.g. a small group of people working together may find it easier to have a shared understanding of a rating scale), and should be kept in mind in interpreting the results or generalizing based on the findings of this study.

All MELAB raters go through a standard training process, and in the case of this study, all novice raters were trained by the same trainer, thus addressing one possible source of variability. As with other rater training programs (Lane & Stone, 2006), the training process includes familiarization activities, several rounds of practice rating, and extensive discussion with and feedback from the trainer throughout the process. But in the case of the MELAB, the completion of these activities merely permits the rater to do provisional live rating. During this provisional period, these new raters receive continual feedback from fully certified raters regarding their performance; the feedback can be oral or written and highlight aspects of new raters' marking that may be problematic. Only after new raters provide a total of 80 consecutive ratings with no more than 10% off ratings do they become fully certified raters. In total, it can take up to several months for a rater to be certified. For additional quality assurance, all raters are also monitored continually to ensure that they are still performing at an acceptable level of accuracy (Johnson, 2005, 2006, 2007; Johnson & Song, 2008).

Research questions

Within the context described above, using actual operational test data, this study seeks to answer several questions regarding rater quality, which for the purposes of this study will refer to raters' severity and consistency in rating. The research questions are as follows:

1. How does novice raters' rating quality in one English language proficiency testing context develop over time?
2. To what extent do raters in this context maintain their rating quality over time?

Method

Measures of raters' severity and consistency can be obtained through the use of multifacet Rasch analysis (Linacre, 1989, 2006), a method that puts variables of interest – for example raters, rating scale, examinee scores – onto a common, interval scale, thus facilitating meaningful comparisons among them. The method has been used in numerous previous investigations into rater behavior in writing assessment (e.g. Johnson & Lim, 2009; Kondo-Brown, 2002; Lumley, 2005; O'Sullivan & Rignall, 2007; Weigle, 1998).

Data

The data for this study come from a larger study that covers four and a half years (Lim, 2009). MELAB test takers in this time period had an average age of just under 29 years old (SD = 11.1), and came from more than 115 different first-language backgrounds. (For more details about test takers in these time periods, see Johnson, 2005, 2006, 2007, and Johnson & Song, 2008.) The present study covers three specific time periods (Table 1), and comprises all operational ratings (n = 20,662) provided by the raters (n = 11) selected for the study – including off ratings that did not finally figure in determining examinees' scores. The ratings represent 83% of all MELAB ratings given in the time periods covered. The 11 raters in the study are identified by alphabetical characters as Raters A to K.

Two raters at the beginning of each time period are new raters, giving the study a total of six new raters. As previously mentioned, the six new raters are all regular employees of the English Language Institute, University of Michigan who work as language testing professionals. Two of the six are male (Raters G and K). All six have an undergraduate background in linguistics, with the exception of Rater G, who has undergraduate and graduate degrees in English literature and education. The six raters are 'new' in that prior to their first appearance in this data, they had no experience rating compositions for MELAB. Of the six, only Rater G had previous experience rating compositions in any

Table 1. Time periods, raters, and number of ratings

Rater	Time period			*Total*
	1	2	3	
	(1/04–12/04)	(9/04–5/06)	(1/07–1/08)	
A	1101	2648	2091	5840
B	527			527
C	507	1450		1957
D	719*			719
E	419*	1307		1726
F		632*	407	1039
G		2623*	1418	4041
H			1903	1903
I			1336	1336
J			940*	940
K			634*	634
Total	3273	8660	8729	20662

*Denotes new rater at beginning of time period context. Over time, new raters become experienced raters. Thus, for example, Rater E is classified as a new rater in Time Period 1, but is no longer classified as such in Time Period 2.

The data were divided into three time periods and include only a selection of raters – those who rated throughout each time period – due to the need for a fixed frame of reference for the analysis, that is, that the severity and consistency measures in each time period be based on the performance of the same group of raters. In any event, as previously mentioned, the 11 raters selected for inclusion in the analysis provided approximately 83% of all MELAB ratings in the time periods covered by the study. That is to say, the data still cover almost the entire population of MELAB test takers in those time periods.

Data connectivity

To do multi-facet Rasch analysis, it is important that the data be connected; that is, all the elements need to be linked in some way so that there are no 'disjoint subsets' so as to yield unambiguous measures (Linacre, 1989). This analysis benefits from MELAB compositions being double-marked, creating a data structure with strong connectivity among raters and robust estimates for the same. However, it also is the case that the MELAB writing test asks test takers to choose between two prompts and to respond to just one. This creates a connectivity problem with regard to prompts; if each test taker responds to only one prompt, it is impossible to tell if any scoring differences observed are due to prompt difficulty or to test-taker ability.

The approach taken by other studies (e.g. Breland, Lee, Najarian, & Muraki, 2004; Broer, Lee, Rizavi, & Powers, 2005; Lee, Breland, & Muraki, 2004) to solving this problem is by creating matching variables – usually some overall language ability variable based on test-takers' scores in other skill areas – and then matching different test takers according to their similarity in that regard. This is arguably an imperfect solution, as it requires making certain arguments regarding the relationship between writing and the other skills. As well, identical overall scores can well mask differing skill profiles.

In the present study, the data permitted making a weaker assumption. A large number of test takers took the MELAB more than once; some up to 10 times (Johnson, 2004). In the data, these repeat test takers are treated conservatively as distinct individuals, as their ability may have changed over time, and it was no longer the same 'person' taking the test. In many cases, however, these test takers' scores did not change much, if at all (Johnson, 2004). Thus, unlike other studies where matching depended on similarities in test scores alone, this study could make matches according to similarities in test scores *and* the fact that those being matched were in fact the same person. In addition, elapsed time between test sittings was taken into account, providing an additional control for the matching; the less time between sittings, the less likely a person's true ability had changed. Taking the above together, there can be greater confidence that matches being made are warranted.

Multi-facet Rasch analysis is very robust with regard to missing data, and estimates for the entire data can be made so long as some minimum overlap and connection exists among them (Bond & Fox, 2007). In matching, a procedure was followed that maximized stringency while minimizing matches required. Elapsed time and difference between listening scores and between reading scores on the two sittings to be matched were each set at the minimum (i.e. 1 month; score difference on each of the two receptive skills = 0) and gradually increased until the FACETS software (Linacre, 2006) indicated that connection had been achieved. Connection was achieved with the parameters set at three months between test sittings, and score differentials on the listening and reading tests of no more than two points between sittings. (For reference, the standard error for the listening and reading tests are approximately 3 points and 4 points, respectively.) In total, using the above parameters, only 214 pairs of test-taker performances needed to be matched for data connection to be achieved. The mean score difference between the second and first sittings for matched performances is 0.24 of a scale point (SD = 0.94 scale point). The connection created through this matching allowed unambiguous analysis of the entire data of over 20,000 ratings using multi-facet Rasch. Full details of the matching procedure can be found in Lim (2009).

Data analysis

The software FACETS (Linacre, 2006) was used to perform multi-facet Rasch analysis. First, the complete, original four and a half year data set was run through FACETS to produce difficulty estimates for each prompt and ability estimates for each examinee L1 background. Then, with prompt and L1 background anchored to the previously arrived at estimates, separate runs of FACETS were specified for each month covered by the present study, including only the raters selected for the study. Prompt and L1 background were anchored as the reduced monthly data would no longer be connected otherwise, and also as quality controls to ensure that differences in these have no effect on resulting estimates for raters. Following the above procedures, month-on-month estimates are produced for each rater, making it possible to observe their performance over time.

Rater severity is reported by FACETS in terms of distance from the group mean, in log-odds units or logits. While these units provide probabilistic information that is useful for certain purposes, reporting in terms of logits is not very helpful in the present instance. Recall that within each run, the raters and the rating scale are placed on the same logit scale. Let us say that for a hypothetical month, a rater had a severity estimate of 1 logit, and further, that each MELAB scale point spanned a range of exactly 4 logits each. In this example, as shown in the equation

$$1\ \text{logit} \times \frac{1\ \text{MELAB scale point}}{4\ \text{logit}} = 0.25\ \text{MELAB scale point}$$

the rater is more severe than the average rater by a quarter of a MELAB scale point. Let us say further that for the following month, the same rater had a severity estimate of 1 logit, but that in that month's analysis, each MELAB scale point only spanned 2 logits. In this second month, the rater had the 'same' severity estimate of 1 logit, but the rater was actually more severe than average by half a MELAB scale point. Clearly, 1 logit in the first month is not the same as 1 logit in the second month. In this longitudinal study, expressing the severity estimates in terms of MELAB scale points thus makes sense, as rater performance from one month to another can then be compared more meaningfully.

(The MELAB scale is not a perfect interval scale, so the conversions were based on the average logits of MELAB scale points in each month's analysis.) The measure remains such that positive values indicate harshness while negative values indicate leniency. Where rater consistency is concerned, this study reports the infit mean square residual statistic. Compared to the outfit mean square, this measure is generally considered more useful because it is weighted to favor on-target observations that are more accurately measured (Henning, 1992). The results for rater severity for all three time periods will be presented first, followed by the results for rater consistency over the same three periods.

Results and discussion

Rater severity over time

The raters' month-on-month severity estimates in each of the three time periods are presented here in graphical form. As previously stated, the measures are presented in terms of the original MELAB scale; thus, a value of one means the rater is more severe than the average rater to a magnitude equal to one scale point. The magnitude of raters' severities is not the focus of this study, but rather maintenance and change, improvement and deterioration over time. However, an operational definition of quality is required to address issues such as when novice raters are ready to rate and whether expert rater can be a meaningful category or not. Judgment is of course required in doing this. For the present, the paper will employ +/−0.5 of the average as the cutoff for acceptable quality, on the argument that the rater who falls within that limit is still more likely to give the appropriate score than the next higher or lower score.

Figure 1 shows that in Time Period 1, the two novice raters were initially more lenient than the experienced raters. However, the magnitudes of their leniency were both less than half a scale point of the average. On the other hand, the three experienced raters were clustered tightly together at the beginning of the time period, with estimates of around 0.2. If we assume that theirs is the proper interpretation of the rating scale, and following that, if the analysis were anchored to their level of severity, then new Rater E would be more than half a band more lenient. However, Rater E did not remain more lenient for long. By the second month, at least in terms of severity, there is no longer a clear separation between novice and experienced raters. It can also be seen that over the course of the year, the raters' severities all do manage to stay within half a band of the average. Rater B did seem to be moving away from the average toward the end of the period (−0.49 in December), as was Rater C.

Figure 2 shows that in Time Period 2, the new raters were quite extreme when they started provisional live rating. Rater F was very lenient, while Rater G was very harsh. Both became more moderate relatively quickly, however. By the second month, Rater G was at less than 0.5, and then more gradually moved ever closer toward the middle. Rater F took slightly longer, becoming less than half a scale point of the average only in her fourth month of rating. But as previously mentioned, experience can be quantified according to length of time rating and volume of ratings given. Table 2 shows that Rater G read more compositions more quickly compared to Rater F. Both raters came within 0.5 of the average at around the same volume of compositions rated; for Rater F, somewhere between 94 and 121, and for Rater G, somewhere between 35 and 130. Unfortunately, the data could not be broken down more finely to determine the volume at which each rater crossed the threshold. Of note about Time Period 2 is that after the initial few months, from about February 2005 onward, this group of raters' severities stayed very close to each other, with no value exceeding +/−0.2.

The new raters in Time Period 3 were very close to the average from the moment they started (Figure 3). Rater K was the most severe in January 2007, but was only 0.05 away from the average. It can also be seen that for the first five months in this time period, the two new raters were somewhat more severe than the other raters, and tracked each other's severities quite closely, but later converged with the experienced raters. At no point in the course of the year did any rater exceed the 0.5 limit, though at the end of the time period the raters seem to divide into two groups and began to diverge in terms of severity.

To summarize, where severity is concerned, it appears that novice raters may or may not be distinguishable from experienced raters. There were new raters who were much more severe or lenient compared to the average, but there were also new raters who performed similarly to existing raters from the moment they began marking. Where novice raters were more severe or more lenient, in this data, they learned to moderate their rating behavior relatively quickly and converge with the more experienced raters. The amount of time it took them to get to within 0.5 of the average varied, but they reached that point at about the same number of compositions rated. Once new raters

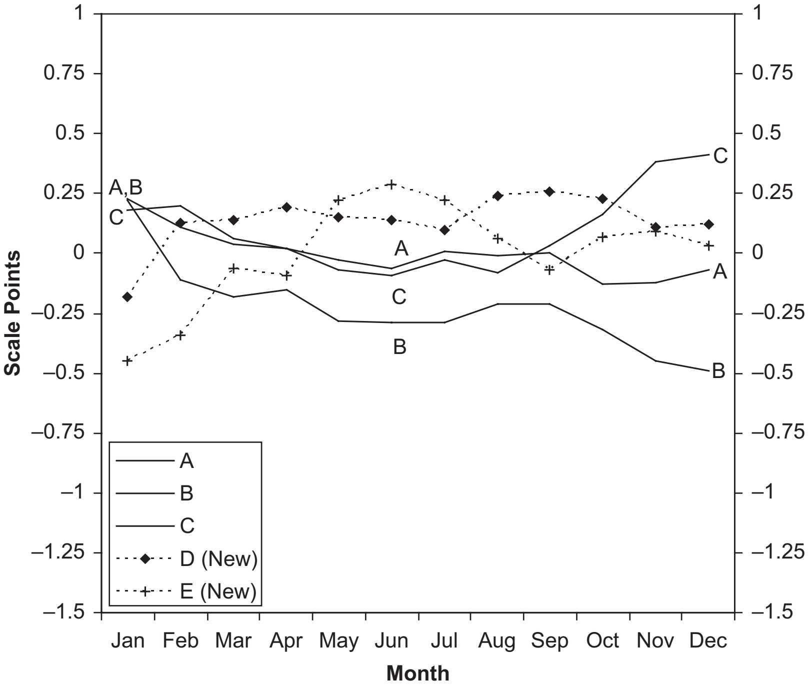

Figure 1. Rater severity, Jan–Dec 2004

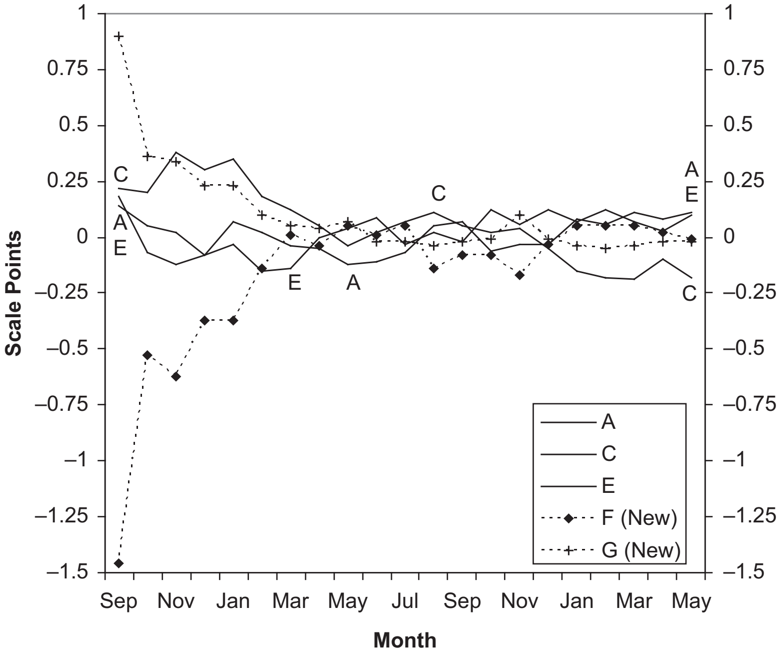

Figure 2. Rater severity, Sep 2004–May 2006

Table 2. Cumulative compositions rated, first four months of Time Period 2

Rater	Sep	Oct	Nov	Dec
F (new)	33	88	94	121
G (new)	35	130	281	428

became acceptably moderate, they stayed moderated throughout the periods of time they were tracked. The experienced raters in the study all began and stayed within acceptable limits of severity throughout.

Rater consistency over time

The infit mean square statistic provides an indication of rater consistency. The statistic has an expected value of one, with higher values indicating more variation than expected (i.e. inconsistency) and lower values indicating less variation than expected. There are no hard and fast rules on what constitutes acceptable fit, and what is acceptable can depend on the type of test being analyzed. That being said, infit values between 0.4 and 1.5 are generally considered to be acceptable (Linacre, 2002; Wright, Linacre, Gustafsson, & Martin-Loff, 1994). In the present study, raters with infit statistics within that range were considered to be sufficiently consistent and not overly predictable.

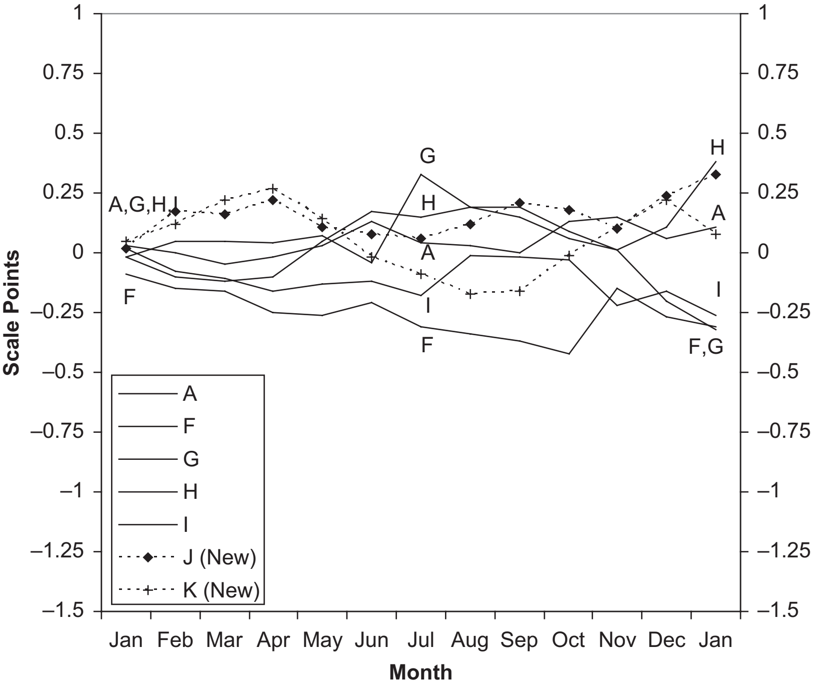

Figure 3. Rater severity, Jan 2007–Jan 2008

Figure 4 shows the fit statistics for the raters in Time Period 1. Overall, the raters were all generally consistent, including new Rater D, with fit statistics mostly between 0.6 and 1.2 throughout the time period. The notable exception is Rater E, a new rater, who was somewhat inconsistent for a few months, but had acceptable fit statistics anyway (1.41 at the highest, in March). More interesting was the way this rater switched from initially exhibiting very little variation to exhibiting too much variation, before settling down and becoming consistent. This would seem to indicate that Rater E was still developing her approach to rating in that time period.

In Time Period 2, we see that new Rater F was initially very inconsistent, with an infit mean square of 2.49 (Figure 5). Her improvement was relatively rapid however, with a slight regression in the third month. This coincided with the fact that she rated just six compositions that month (cf. Table 2), which may partially explain the deterioration in her performance. Rater F was able to self-correct subsequently, however. As with her degree of severity, Rater F was within acceptable bounds by her fourth month rating. New Rater G showed a good degree of consistency from the beginning.

In Time Period 3, as in the previous time periods, one of the two new raters was inconsistent at the beginning. Rater J started with an infit of 2.27, but was rating consistently by her second month, no later than her 50th composition read. Of note is Rater F, by this time an experienced rater, who at the end of the time period became inconsistent for a short while (1.67 in November and 1.65 in December). A look at the number of ratings

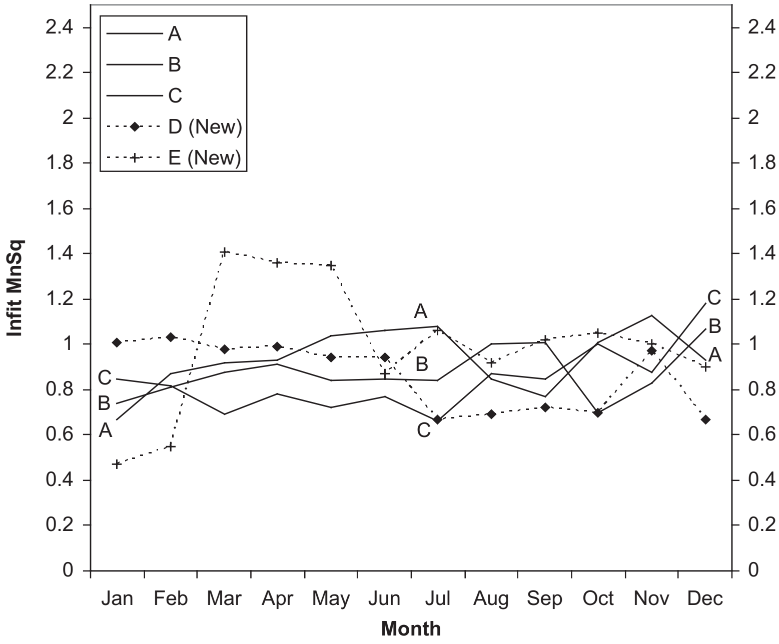

Figure 4. Rater consistency, Jan 2004–Dec 2004

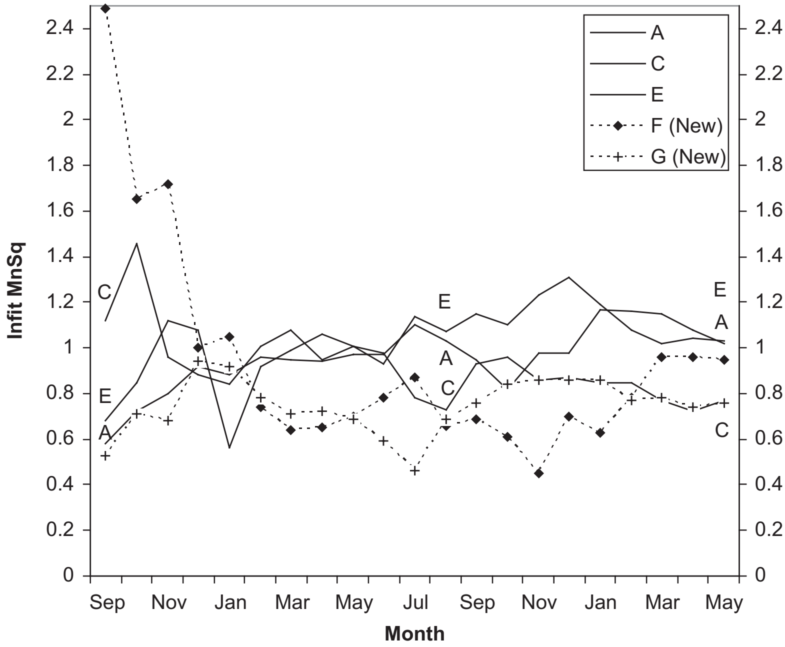

Figure 5. Rater consistency, Sep 2004–May 2006

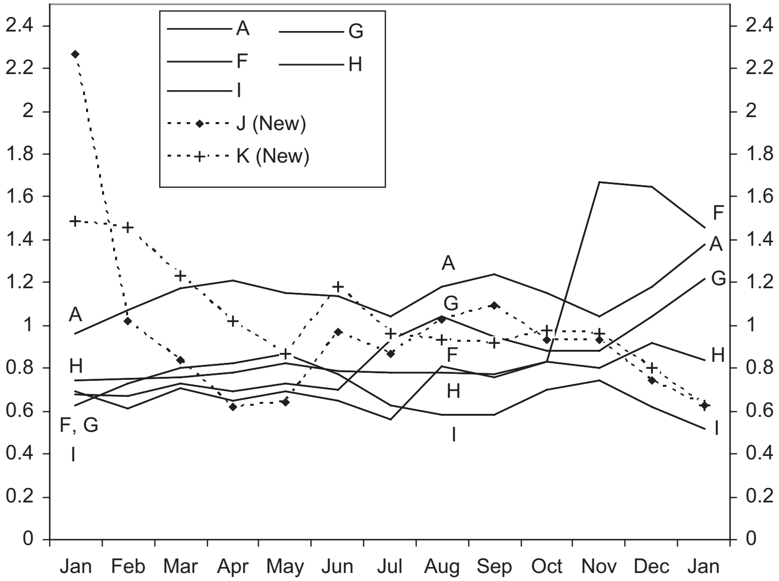

Figure 6. Rater consistency, Jan 2007–Jan 2008

she assigned again provides a possible explanation for this. In the period from September to December, she read 12, 10, 18, and 9 compositions respectively, the small numbers of compositions rated possibly affecting the quality of her ratings. The precise reason is difficult to ascertain in the absence of more qualitative information. In any event, the rater appeared to have self-corrected by January.

To sum up, the findings regarding rater consistency appear to be very similar to the findings for rater severity. New raters may or may not be inconsistent when they begin rating, and those who are do not stay that way for very long. With one brief exception, experienced raters all began and stayed within acceptable bounds of fit throughout the time periods tracked. As with rater severity, there are indications that rater consistency may be related in some way to rating frequency.

Conclusion

The purpose of this study was to shed light on the initial and long-term development of raters of L2 writing, focusing on rating quality. The first research question asked how novice raters' rating quality developed over time. In the context of this data, the results showed that novice raters' severity and consistency were not always distinguishable from their experienced counterparts. In those instances where they were distinguishable, their rating quality was always worse. However, it was also the case that these novice raters' rating quality improved relatively quickly. What contributed to this improvement cannot be determined conclusively from this study; the results suggest, however, that frequency or volume of rating done may be one factor among several that influences rating quality. The second research question addressed the degree to which raters maintained their rating quality over time. In the context of this data, raters stayed within acceptable limits of quality, with the exception of one rater who for a brief (2 month) period rated inconsistently. Raters being able to maintain their rating quality over time indicates that the category of expert rater can in theory exist.

Several limitations need to be considered in interpreting the findings of this study. The study included only a small number of novice raters, and investigated only one testing context, which had particular distinctives (e.g. rater pool, rating scale, training program, language being assessed). Few large-scale language tests are rated entirely by a small group of raters who work out of a single location and who interact with each other on a daily basis. This rating context is interesting in itself and may well have contributed to the high quality of the ratings observed, and bears further investigation (cf. Fox, 2003). Given the nature of the data, the study was only able to address rater behavior by inference through rating outcomes. Thus, it is difficult to ascertain what might be the cause or causes behind, for example, the improvement of novice raters or why one experienced rater briefly became inconsistent. Future studies would benefit from collecting diverse forms of data alongside raters' ratings, which would allow them to make stronger claims about rater development. Also, in focusing on agreement among the raters, the study left untouched the question of what they are agreeing about, i.e. construct validity. Even so, in providing empirical evidence of novice raters' development over time, and of experienced raters' continuing rating behavior, the study has added to the field's knowledge about one key aspect of performance assessment. As raters' performance impacts on test validity, reliability, and fairness, a research priority should therefore be to examine in greater detail novice and experienced raters' rating quality and rationales at multiple junctures over an extended period of time in different testing contexts using various complementary forms of data. This could reveal how well the main findings of the study generalize beyond its context.

For the MELAB program, it has to be reassuring that the novice raters all learned to rate appropriately. On the other hand, had there been novice raters who did not learn to rate appropriately, additional questions could be asked such as why some succeeded and some failed, and what it is that makes people suitable and not suitable for the task of rating. The novice raters in this study were clearly a select group: they had backgrounds in linguistics, and worked for a testing organization. Future research could investigate the suitability of other individuals with other characteristics for the task of rating. For that matter, this study only included novice raters who all went through the training program. In the absence of novice raters who were not trained, it cannot be ascertained how much training contributed to these raters' eventual success, if at all. Thus, the necessity and efficacy of rater training cannot be addressed by this study.

Regarding when novice raters are ready to rate, individual differences may make it impossible to answer the question definitively; some appear to be ready immediately after training, while others are not. However, the data intimates that there is a relationship between rating volume and rating quality. This accords with theories of associative learning, which posits frequency effects (Ellis, 2002), and with the general observation in other domains of human activity that experience does lead to better performance (e.g. operating a motor vehicle). In this study, all novice raters had acceptable rating performance no later than the 130th composition rated. This suggests there is wisdom in the MELAB program requiring raters to rate at least 80 more compositions beyond training before they can be fully certified, as raters may require at least that much experience to develop their approach to rating and/or to settle into their rating behavior. It would appear to be prudent for other test providers to give new raters at least that measure of experience as well. The matter of when individuals are ready to rate clearly requires further investigation.

To the question of allowing raters to continue rating over time without additional intervention, as in the case of certification, the study suggests that that practice can be warranted to a certain extent. Raters do appear able to maintain their rating quality over time. However, the fact that raters who rated intermittently were excluded from the data needs to be borne in mind. The raters who were included were those who rated regularly, and in fact, among them, one rater did become inconsistent for a brief period, perhaps on account of reading very few compositions over that time. Thus, whether there is a minimum amount of continuing experience necessary to maintain one's rating quality, and what that minimum amount is, is certainly an interesting question and a proper subject for future research.

That raters can be shown to maintain their rating quality over time is important, because it means that the idea of an expert rater is potentially legitimate and does not need to be abandoned. (Granted, in this study, a particular definition of minimum quality was adopted, which may or may not be quality enough, though it was also seen in Time Period 2 that, in theory, a more stringent cutoff of +/−0.25 rather than +/−0.5 can be maintained.) The idea of an expert rater is valuable in that if the category exists and can be defined, it could then provide a goal and direction for rater selection, training, and development. To the end of making that category more concrete, a greater engagement with the study of expertise and expert performance (Ericsson, Charness, Feltovich, & Hoffman, 2006; Mislevy, 2008) would seem to have merit. However, a problem does come up immediately, that of definition. In a number of fields, identifying the expert is relatively easy (e.g. in athletics, the one who consistently crosses the finish line first). Such clear indicators are not available, though, in the case of rater performance, which has to do with the continuing quality of evaluative judgments. Who is to say whose judgments are better or worse? And does that not require judgment as well? For that reason several studies, including the present study, have identified the expert relative to the group norm (e.g. Wolfe et al., 1998). But then, one could be an expert or not an expert depending on the performance of the rest of the group; that cannot finally be the basis of an adequate definition. More theorizing in this regard is in order.

In sum, this study has provided evidence regarding initial and long-term development of raters and rating quality. Its finding implications for practical aspects of testing such as rater preparation and certification, as well as theoretical notions such as rater expertise. But it has at the same time raised more fundamental questions about raters and their ratings, which should rightly be the subject of more studies.

Acknowledgements

The author gratefully acknowledges the English Language Institute, University of Michigan, for the use of the data and for a Spaan Fellowship that in part supported this work. Thanks also to Talia Isaacs and Evelina Galaczi, and to the *Language Testing* editor and reviewers, whose comments and feedback have made this a much better paper. A version of this paper was presented at the 2010 meeting of the American Association of Applied Linguistics in Atlanta, Georgia.

References

Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review*, *64*(1), 99–134.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Breland, H., Lee, Y. W., Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups*. TOEFL Research Reports, RR-04-05. Princeton, NJ: Educational Testing Service.

Broer, M., Lee, Y. W., Rizavi, S., & Powers, D. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty*. ETS Research Report, RR 05-11. Princeton, NJ: Educational Testing Service.

Chalhoub-Deville, M. (2003). Fundamentals of ESL admissions tests: MELAB, IELTS, and TOEFL. In D. Douglas (Ed.), *English language testing in US colleges and universities* (2nd ed., pp. 11–35). Washington, DC: NAFSA.

Cho, D. W. (1999). A study of ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing*, *8*(1), 1–24.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163–178.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171–191.

English Language Institute, University of Michigan. (2005). *Michigan English language assessment battery: Technical manual 2003*. Ann Arbor, MI: English Language Institute, University of Michigan.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.) (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge: Cambridge University Press.

Fox, J. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing*, *3*(1), 21–48.

Hamp-Lyons, L. (1991). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5–15). Norwood, NJ: Ablex.

Henning, G. (1992). *Scalar analysis of the Test of Written English*. TOEFL Research Reports, RR-92-30. Princeton, NJ: Educational Testing Service.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206–232). Cresskill, NJ: Hampton Press.

Johnson, J. S. (2004). *Score gains for repeat MELAB administrations*. ELI Research Reports 2004-04. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2005). *MELAB 2004 descriptive statistics and reliability estimates*. ELI Research Reports 2005-03. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2006). *MELAB 2005 descriptive statistics and reliability estimates*. ELI Research Reports 2006-02. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2007). *MELAB 2006 descriptive statistics and reliability estimates*. ELI Research Reports 2007-03. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485–505.

Johnson, J. S., & Song, T. (2008). *MELAB 2007 descriptive statistics and reliability estimates*. ELI Research Reports 2008-03. Ann Arbor, MI: English Language Institute, University of Michigan.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3–31.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education.

Lee, Y. W., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT prompts for different native language groups*. TOEFL Research Reports, RR-04-24. Princeton, NJ: Educational Testing Service.

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment*. Retrieved from ProQuest Dissertations and Theses (Accession Order No. AAT 3392954).

Lim, G. S. (2010). Investigating prompt effects in writing performance assessment. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *8*, 95–116.

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

Linacre, J. M. (2002). What do infit, outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2006). *Facets Rasch measurement computer program*. Chicago: Winsteps.com.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Mislevy, R. J. (2008, September). *Some implications of expertise research for educational assessment*. Keynote address delivered at the 34th International Association for Educational Assessment Conference, Cambridge, UK.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371–389.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor, & P. Falvey (Eds.), *IELTS collected papers. Research in speaking and writing performance* (pp. 446–478). Cambridge: Cambridge University Press.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson, & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237–265). Cresskill, NJ: Hampton Press.

Purpura, J. E. (2005). Michigan English language assessment battery (MELAB). In S. Stoynoff, & C. A. Chapelle (Eds.), *ESOL tests and testing* (pp. 87–91). Alexandria, VA: TESOL.

Santos, T. (1988). Professors' reactions to the academic writing of non-native-speaking students. *TESOL Quarterly*, *22*(1), 69–90.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303–325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27–33.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

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

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465–492.

Wright, B. D., Linacre, M., Gustafsson, J. E., & Martin-Loff, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.