Nordic Journal of Modern Language Methodology

2020, 8 (1), p. 3-29

Effects of Rater Training on the Assessment of L2 English Oral Proficiency

Pia Sundqvist[1,2], Erica Sandlund[2], Gustaf B. Skar[3], Michael Tengberg[2]

ABSTRACT

The main objective of this study was to examine whether a Rater Identity Development (RID) program would increase interrater reliability and improve calibration of scores against benchmarks in the assessment of second/foreign language English oral proficiency. Eleven primary school teachers-as-raters participated. A pretest–intervention/RID–posttest design was employed and data included 220 assessments of student performances. Two types of rater-reliability analyses were conducted: first, estimates of the intraclass correlation coefficient two-way random effects model, in order to indicate the extent to which raters were consistent in their rankings, and second, a many-facet Rasch measurement analysis, extended through FACETS®, to explore variation regarding systematic differences of rater severity/leniency. Results showed improvement in terms of consistency, presumably as a result of training; simultaneously, the differences in severity became greater. Results suggest that future rater training may draw on central components of RID, such as core concepts in language assessment, individual feedback, and social moderation work.

KEYWORDS:

assessment, rater training, interrater reliability, L2 oral proficiency, MFRM

Acknowledgements: This research was supported by the Center for Language and Literature in Education and Research on Subject-specific Education (ROSE), Karlstad University, Sweden. We would like to thank our fellow project members for their respective contributions to the Equity in Assessment (EquA) project: Eric Borgström, Örebro University, Sweden, and Lena Lötmarker, Michael Walkert, and Kristina Wikberg, all Karlstad University, Sweden.

[1] Department of Teacher Education and School Research, University of Oslo, Norway
[2] Faculty of Arts and Social Sciences, Karlstad University, Sweden
[3] Department of Teacher Education, Norwegian University of Science and Technology, Norway

1. Introduction and Aims

In this exploratory study, we target equity in the assessment of second language English oral proficiency. In order to achieve equity in the assessment of a complex ability such as speaking in a second or foreign language (L2), it is crucial that raters are clear not only about the test construct and grade criteria, but also about the interpretation of scores. Further, subjective scoring of complex language abilities are bound to involve some disagreement between raters (Berge, 2005; Meadows & Billington, 2005; Stemler, 2004). However, if the disagreement is too large, the reliability of assessment is jeopardized and, consequently, also equity. In contexts like the one where this research has been conducted, Sweden, where oral assessments play a vital role in end of year report cards, low reliability can have direct and detrimental effects; the same level of oral proficiency displayed in a test may be scored differently by different scorers. Methods for raising interrater reliability are therefore needed, but the evidence for effective methods is still scarce (Jönsson & Thornberg, 2014; Weigle, 1998). In addition, if the rater is the test-takers' own teacher, which is the case in the test used for the present study, the situation possibly becomes even more challenging (McNamara, 2001; Sundqvist, Wikström, Sandlund, & Nyroos, 2018) and the evidence for effective methods even more scarce.

The teachers who were involved as raters in this study work in primary school and the test in focus is the national speaking test of English for students in the 6th grade in Sweden. Internationally, it is unusual that L2 oral proficiency assessment research focuses on tests taken by such young learners; it is much more common to examine large-scale speaking tests (e.g, TOEFL Speaking) and test formats (e.g., the Oral Proficiency Interview) taken by adults or young adults (Roca-Varela & Palacios, 2013). Yet, around the globe, the L2 speaking skills of young learners are assessed all the time. At best, these assessments are reliable and perceived as fair by the learners and helpful for individual learners' oral language development; at worst, the assessments are unreliable and they may also turn learners silent, as young learners tend to be particularly vulnerable to assessment (McKay, 2006). As such, research targeting the assessment of young learners is much needed.

This study addresses this under-researched area of L2 oral proficiency assessment with a particular focus on whether interrater reliability between teachers, as examiners of a high-stakes speaking assessment for young L2 English learners, can improve with a research-based training program. The overarching aim of this study was to explore certain training practices devised to support increased reliability in rating oral proficiency. Specific aims included to see whether a training program (Rater Identity Development, RID) would increase the interrater reliability in a group of raters and, additionally, improve the calibration of their assessments against benchmarks.

2. Literature Review

2.1 Rater agreement and rater training in L2 assessment

Broadly speaking, ‘L2 proficiency testing’ is about assessing a learner’s ability to use an L2. L2 *oral* proficiency, in particular, has been described as a complex, multifaceted construct for assessment (Davis, 2009; Galaczi, 2008; Gan, 2010; Iwashita, 2001; Lazaraton & Davis, 2008; Luoma, 2004; McNamara, 1996, 2001; Thomas, 1994), where interrater reliability is dependent on, among other things, raters’ understanding of different aspects for L2 oral proficiency, and their relative importance in assessing speaking (Ducasse & Brown, 2009; May, 2011).

With spoken proficiency, raters’ perceptions of a learner’s proficiency – described as capturing “a moving target” (Leclercq & Edmonds, 2014, p. 5) – are also linked to the tasks designed for obtaining assessable output, the interlocutor and/or examiner’s proficiency and conduct, and not least to individual preferences of raters (Sandlund, Sundqvist, & Nyroos, 2016). Further, it has been argued that in particular for the assessment of ‘complex’ performances, such as written discourse or spoken production, rater agreement is a challenge because of “the individualized uniqueness and complexity” (Wang, 2010, p. 108) of the tasks and performances to be assessed (see also Papajohn, 2002). Thus, in terms of rater agreement, as Papajohn (2002, p. 219) notes, a question arising from efforts to improve rating consistency is “whether different raters derive scores of the same response for the same reasons”, referring to the fact that raters make judgments on different grounds. In addition, previous research on rater severity reveals that whereas returning experienced raters tend to move towards much more consistency and severity, inexperienced raters show greater inconsistency (Bonk & Ockey, 2003).

To date, research on rater training is not conclusive as to the effects of training on rater performance. While Elder, Knoch, Barkhuizen, and von Randow (2005) report positive outcomes of rater training, in a later study, the same scholars conclude that there is "considerable individual variation in receptiveness to the training input" (Elder, Knoch, Barkhuizen, & von Randow, 2007, p. 37). Further, Weigle (1998) concludes that rater training is more successful in assisting raters to give more predictable scores (*intrarater* reliability) than in assisting them to assign identical scores (*interrater* reliability). Moreover, a recent study indicated that unreliability in part might be explained by teachers' different stances toward rating students' performances, with some teachers leaning toward focusing particularly on students' strengths and other teachers being more prone to match students' responses with assessment criteria (Jølle & Skar, 2018). To sum up, teachers' and raters' professional judgments "will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement" (McNamara, 1996, p. 117).

2.2 Moderation

In order to improve equity in the assessment of high-stakes tests, it is common that several raters assess the same performance, and compare and discuss their evaluations. Such *moderation* is designed to improve interrater reliability and to ensure that the individual assessment assigned to a learner performance is independent of the rater (Sadler, 2013, p. 5; see also Skar & Jølle, 2017). Research and intervention efforts to moderate raters' judgements toward greater consistency on what is to be assessed have brought forth the notion of *consensus moderation* (Sadler, 2013; see also Stanley, MacCann, Gardner, Reynolds, & Wild, 2009), meaning that raters/teachers work to reach consensus on how grades should be awarded. For such moderation to be effective, it is important with a non-threatening and non-judgemental environment in which teachers respect each other's opinions (Klenowski & Adie, 2009). Jönsson and Thornberg (2014) discuss two separate goals of teacher collaboration in assessment: collaborative assessment for increased rater consistency and agreement, and collaborative assessment as a means to reach a shared understanding of how assessment criteria are best interpreted and applied. By discussing and analyzing individual student performances, and managing disagreements about assessments, opportunities for reaching a shared understanding of steering documents and scoring rubrics emerge. As such, teacher/rater discussions on selected student performances may function as 'learning communities' (Wiliam, 2007) where extreme positions can be smoothened out as each teacher/rater will have to account clearly for their views underlying a particular assessment decision (Adie, Klenowski, & Wyatt-Smith, 2012; Klenowski & Adie, 2009).

In a study on teachers' views on moderation and judgement, it was shown that teachers had a positive attitude towards using standards in moderation (Connolly, Klenowski, & Wyatt-Smith, 2012). Furthermore, they perceived that the use of standards produced consistency in assessment/judgement. When teachers held varying opinions about the level of a specific student performance, the results showed that the actual moderation contributed to teachers' learning processes as raters. The researchers conclude that "teachers' assessment beliefs, attitudes and practices impact on their perceptions of the value of moderation practice and the extent to which consistency can be achieved" (Connolly et al., 2012, p. 593). In sum, in moderation sessions, raters' divergent views on assessment can be brought into the spotlight and be discussed (Adie et al., 2012; Connolly et al., 2012; Jönsson & Thornberg, 2014; Klenowski & Adie, 2009). All in all, this makes possible a shared understanding among teachers/raters.

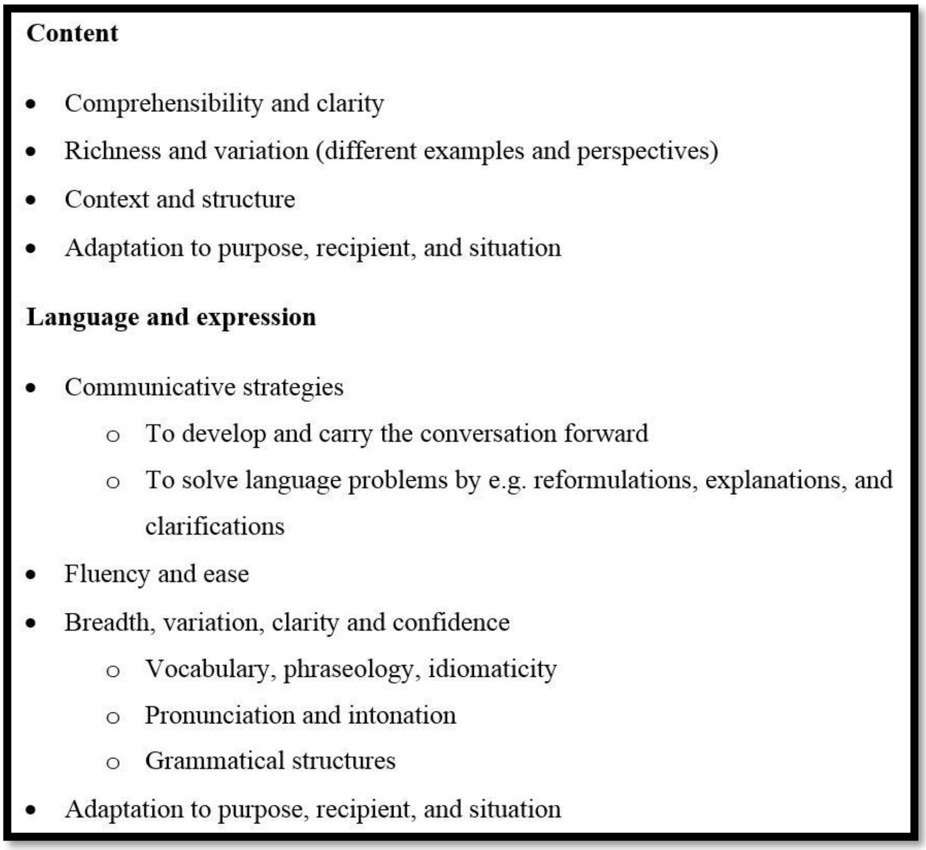
To date, few studies attempt to develop effective rater training specifically for L2 oral proficiency and interaction (but see research reports on the TOEFL test, Chalhoub-Deville & Wigglesworth, 2005; and Wang, 2010, in a Chinese context). In addressing this gap, the present study aims to test a program for rater development. Although test authorities and test constructors may provide raters/teachers with specific guidelines to be used in professional talks about L2 oral language assessment (East, 2016; Swedish National Agency for Education, 2013), such guidelines or materials seem to focus more on raising the general awareness about assessment practices as opposed to raising awareness of individual assessment practices specifically. In what follows, we turn to the speaking test used in this study.

3. The National Test in English in Sweden

World-wide it is common to apply systematic evaluations of student performances by use of high-stakes tests, including standardized testing of language skills (Brown & Abeywickrama, 2010). Sweden is no exception in this regard, and mandatory, summative national tests in core subjects, including English, have been used in secondary school since the 1990s. The purpose of these tests is twofold, namely to contribute to equity in assessment and to yield data for evaluation of goal-attainment (Swedish National Agency for Education, 2015). However, compared to many other countries, high-stakes testing in Sweden differs in that teachers (as opposed to external examiners) both serve as administrators and single raters of the externally produced tests, something that makes questions about the need for rater training (EACEA, 2009) and test reliability (Nusche, Halász, Looney, Santiago, & Shewbridge, 2011) critical.

With regard to *primary* school, in spring 2012, a 6th-grade English national test became mandatory and from spring 2013, learner results on this test should also inform the English grade awarded at the end of the 6th grade, which is the last year of primary school in Sweden (NAFS Project, 2012). The assessment data collected here stem from the speaking part of this test, henceforth referred to as the National English Speaking Test (NEST); other parts of the test include listening and reading comprehension, and writing (Swedish National Agency for Education, 2015). The purpose of the whole test is to measure students' global English proficiency.

The NEST aims to measure students' "oral production and interaction" (Swedish National Agency for Education, 2015, p. 30) and so-called topic cards (with statements or questions) are used to elicit talk amongst the students (for example, "Cats are better than dogs. Why?/Why not?"). Teachers should award each student performance a holistic score/grade based on a number of assessment factors and given grade criteria (A–E) in the curriculum, and if criteria are not met, the grade F is assigned (Swedish National Agency for Education, 2011a). These criteria are aligned with the descriptors for communicative abilities described in the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001). In grade 6, a passing grade (E) corresponds to CEFR level A2.1 (Swedish National Agency for Education, 2011b, p. 7). The assessment factors that relate to the NEST test construct – oral production and interaction – are listed under two headings, *Content* and *Language expression*, see Figure 1 (Swedish National Agency for Education, 2015, pp. 30, translated from Swedish):



Content

- Comprehensibility and clarity
- Richness and variation (different examples and perspectives)
- Context and structure
- Adaptation to purpose, recipient, and situation

Language and expression

- Communicative strategies
  - To develop and carry the conversation forward
  - To solve language problems by e.g. reformulations, explanations, and clarifications
- Fluency and ease
- Breadth, variation, clarity and confidence
  - Vocabulary, phraseology, idiomaticity
  - Pronunciation and intonation
  - Grammatical structures
- Adaptation to purpose, recipient, and situation

Figure 1. Assessment factors for NEST

Before NEST administration, teachers are under obligation to read a test booklet with instructions. This booklet includes a CD with sample NEST recordings that teachers should listen to, and it also provides written comments made on relevant criteria applicable to the specific assessments of the student performances on the CD (representative of different grade levels). Thus, these recordings with their accompanying grades function as benchmarks (National Assessment Project, 2015). Clearly, the faith put in teacher professionalism is strong, and for the sake of all stakeholders, students included, it is crucial that the system works so that the goal of equity in assessment can be reached (cf. Gustafsson & Erickson, 2013).

4. Research Questions

There is a need for methods of raising interrater reliability between teachers and as argued by Jönsson and Thornberg (2014) as well as by Weigle (1998), the evidence for effective methods is scarce. In this study, a training program called *Rater Identity Development* (*RID*, described below) was developed and tested with the purpose of raising interrater reliability in teachers' assessment of L2 English oral proficiency. An important aim with the program was to improve the extent to which teachers' assessments aligned with benchmarks provided by the test constructors. The research questions have to do with the RID training program, but each question serves the purpose of examining rater identity development from a slightly different perspective. The study poses the following two research questions (RQs):

RQ1. To what extent does the teachers' assessment of L2 English oral proficiency display variation?

RQ2. Upon rater training, to what extent do the teachers change their assessment practices with regard to calibrating assessments against benchmarks?

Quantitative assessment data collected at two points in time (pre- and posttest design) are used to answer both research questions.

5. Method

5.1 Purpose and design

The study is part of a project that combines (a) research and (b) professional development for teachers, Equity in Assessment (EquA). The overarching project aim was to contribute to equity in assessment in the national tests of two school subjects, Swedish (which is the majority and first language, L1) and English (L2). This study focuses on the English track only.

5.2 Outline of rater training program

The rater training program we developed included three integrated components, aiming at developing teachers' awareness of their own 'identities', or 'profiles', as raters. The three components were (i) detailed feedback on their individual assessments, (ii) theoretical input about language assessment, and (iii) repeated moderation sessions in small groups. In the program, the first day was a 'pretest day' (June), the second a 'rater training day' (that is, the actual intervention/treatment, August), and the third a 'posttest day' (September). In addition to collecting assessment and questionnaire data from the teachers on these occasions (see Material, section 5.4), members of the project team offered various lectures each day. These lectures included topics deemed relevant to the participating teachers for their professional development in general, and for their development as raters of various language abilities in particular. Some important concepts in the area of assessment covered in the lectures include *validity*, *reliability*, *benchmarks*, *standards*, *test construct*, *construct relevant/irrelevant criteria*, *formative* versus *summative assessment,* and *high-stakes* versus *low-stakes testing*. In the English track, special attention was paid to the assessment of L2 speaking.

Following the pretest assessments, each participant was provided (via email) with individual feedback in order to raise their awareness of their own rater profiles. For instance, they received information about their own and the group's assessments of student performances, and also about the benchmarks. One part of all this feedback was a figure that showed the mean assessment scores (based on ten assessed student performances) made by all participating teachers, where each teacher's own mean score was displayed in a bar.

As it turned out, the English group assessed relatively close to the benchmarks already on the pretest (see Table 4 in section 6.1), and this information was passed on explicitly in the emails. They also received information about the number of grades they had used on the six-graded scale in comparison with the group, and an individual table that summarized their own assessments on the student performances from the pretest. Based on all information, the participants were finally encouraged to start thinking about their rater profile, whether they were strict or lenient (or neither) in comparison with the group and the benchmarks, and, if the goal is equity in the assessment of L2 English oral proficiency, what (if anything) should they think about in the future?

These topics were addressed and additionally explained in a lecture by Sundqvist and Sandlund at the beginning of the rater training day, where the participants also had time to discuss the results from the pretest day and their individual feedback in small groups. Then the authors lectured more on the NEST assessment criteria and on the test construct *oral production and interaction*, including various aspects of oral proficiency. It was further emphasized that professionalism for teachers-as-raters in high-stakes assessment also involves the calibration of individual assessments against benchmarks. To facilitate the participants' talk about assessment and their understanding of what types of raters there may be, as well as how raters may be perceived by students, three 'bird metaphors' were introduced: *the rater as a hawk* (a severe rater, generally rating lower than the benchmark), *the rater as a dove* (a lenient rater, generally rating higher than the benchmark), and *the rater as a blackbird* (a ‘benchmark’ rater, generally rating on or very close to the benchmark).

The participants were divided into groups of three or four for moderation sessions. They were reminded about the fact that the English group as a whole had assessed close to benchmarks (but slightly more severely) at the pretest. Then, for the remainder of the day, they assessed student performances in four test recordings, taking on one recording at a time, following these steps/instructions:

1) Individually: Listen to the test recording (links provided) and fill out the assessment template independently.
2) Group: Open the corresponding test recording envelope (containing the official comments and NEST benchmarks).
3) Group: Take turns ‘outing’ your own rater identity profile (from the pretest day feedback) to the other group members and tell them how you plan to use your newly gained knowledge about yourself as a rater in today’s assessment work.
4) Group: Discuss your own assessments of student performances in the group, make comparisons and identify differences. Discuss your own assessments in relation to the benchmarks.

After the posttest day, the second round of individual feedback was sent out, similar to the previous feedback but with new information about comparisons at group level between pre- and posttest assessments (identical to the results presented in Tables 5–6). The fourth day was set-up as a one-day conference during which results from the project and other studies were shared.

5.3 Participants

The English participants consisted of eleven primary school teachers, all women. They were from different schools and did not know each other prior to taking part in the training. The mean age was 43 (SD = 6.1; range 35–51). Ten had Swedish as a first language (L1) and one had English. The mean work experience was 9.2 years (SD = 5.7; range 1–17). All but one had a teacher’s degree. The amount of higher education in English varied, from nothing to two semesters at tertiary level.

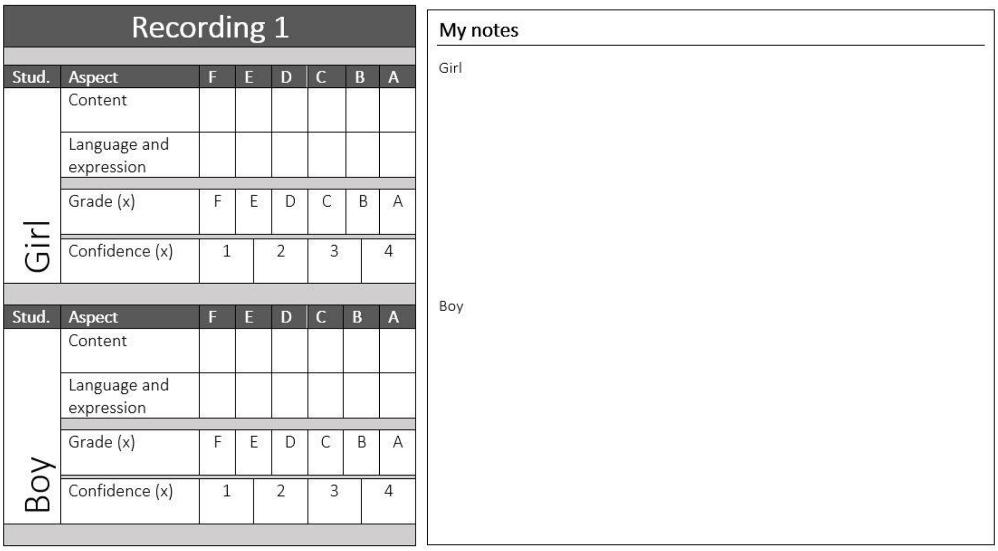
For the purpose of this study, the participants were asked to provide information about how many times they had assessed the English national test in grade 6 or grade 9 (questionnaire). On average, the participants had assessed the 6th-grade test almost five times (*M* = 4.8; SD = 1.8). In comparison, their experience of assessing the 9th-grade test was negligible (*M* = 1.2; SD = .4). They were also asked about their experience of assessing national tests in, for example, Swedish or Mathematics, with very similar findings. As for the experience of grading, the participants had assigned grades 4.7 times (SD = 2.0) in 6th-grade English.

5.4 Material

Ten student performances in five paired test recordings on each day were assessed by the 11 teachers, yielding a total of 220 assessments of student performances. Two student performances served as anchor performances and appeared both on the pre- and posttest for validation purposes (see, e.g., Kolen & Brennan, 2014), which means that in total, nine different tests were used in this study. These tests originate from four batches of NESTs (provided by the Swedish National Agency for Education) and in each test recording, there are two students (boy plus girl). Due to secrecy regulations (that is, restrictions on publicly sharing the tests, as they may be partly re-used during a stipulated number of years), exact test themes or topic card formulations cannot be revealed, but these tests tend to be about topics that most young people can relate to, such as spare time interests or technological developments.

It should be mentioned that the order of the test recordings on each day, as well as across the three days of the project, was carefully planned, taking the type of test and the quality of test-taker performances into account. The aim was to play a variety of tests each day of the project and to have a mix of performances over each day and across the three days, with the intention of minimizing any effects of the specific test or the test order, or rater fatigue for that matter.

Raters’ assessments were collected immediately after the participants had assessed a test. A pen-and-paper template for assessment was used (Figure 2). The template included six ‘boxes’ (A–F) for the holistic grade on the test, where teachers were supposed to tick one. Since these boxes were identical to the assessment material provided by the test constructors, the procedure was familiar. They were also instructed to provide information about how confident they felt about the holistic grade (four-graded scale, see Figure 2). Furthermore, they were asked to assess *Content* and *Language and expression*, that is, the two aspects emphasized in the criteria that teachers are explicitly instructed to consider when evaluating student performances. There was also space for taking notes.



Recording 1

Stud.	Aspect	F	E	D	C	B	A
Girl	Content						
	Language and expression						
	Grade (x)	F	E	D	C	B	A
	Confidence (x)	1	2	3	4		

Stud.	Aspect	F	E	D	C	B	A
Boy	Content						
	Language and expression						
	Grade (x)	F	E	D	C	B	A
	Confidence (x)	1	2	3	4		

My notes

Girl

Boy

Figure 2. Template used for raters' assessments (Confidence: 1 = not very confident at all, 2 = not confident, 3 = confident, 4 = very confident).

5.5 Data analysis

To investigate rater variation, two types of rater reliability analyses were conducted. First the estimates of the intraclass correlation coefficient (ICC) two-way random effects model (McGraw & Wong, 1996) was computed. This statistic indicates to what extent raters are consistent in their ranking of student performances. Our interpretations of ICC agreement measures based on those data will be in line with Cicchetti's (1994) suggested guidelines: 'poor' (< .40), 'fair' (.40–.59), 'good' (.60–.74), and 'excellent' (.75–1.00). An ICC value of 0 means there is no agreement, whereas a value of 1 means total agreement.

To explore rater variation in terms of systematic differences of severity and leniency the data were also fitted to a many-facet Rasch measurement (MFRM) model. The MFRM builds on the basic Rasch model (Rasch, 1980), which states that the probability of a correct answer is given by the difference between person (student) ability and item difficulty. The basic Rasch model has been extended, for example through the computer program FACETS® (Linacre, 2017a), to include additional facets of measurement, such as raters and scales. In this instance, the following model was used:

$\log (Pnijk / Pnij(k-1)) = Bn - Di - Cj - Fk$, where
Pnijk is the probability of student *n* on item *i*, by rater *j* receiving a score of *k*, and
Pnmij(k-1) represents the probability of the same student under the same conditions receiving a score of *k*-1.
Bn is the ability of person n,
Di is the difficulty of item *i* (i.e., Content and Language and expression),
Cj is the severity of judge *j*, and
Fk is the barrier to being observed in category k relative to category k-1.

There are many advantages of fitting assessment data to a MFRM model in applied contexts (Eckes, 2015; McNamara, 1996), but in this particular case the purpose was to take advantage of how MFRM treats observed scores. These are transformed into 'logits' (log-odds units), and when data fit the model this transformation creates a linear scale (Engelhard, 2013). It should be noted that a large number of observations are desirable when fitting of data to a MFRM model. As in all statistical analysis, a small sample produces less precise estimates. Linacre (2020) states that for stable measures it is desirable with at least 30 observations per "element" (i.e., students, raters, and assessment scales) and at least 10 observations for each rating scale category. In our case there were 22 observations per student, 20 per teacher, and on average 36.6 observations per scale score category (one category [F] was below desired level with 8 observations). The study did have fewer observations than wished for, but, as Linacre (2020) notices, it is possible to obtain useful measures with much less than the minimum requirement.

The analysis was done using FACETS® (Linacre, 2017a), which provides a number of useful outputs. First, to assess the 'global fit,' the researcher can compute the proportion of standardized residuals with values of ±2 or ±3. According to Linacre (2017b, p. 170), "[w]hen the data fit the model, about 5% of standardized residuals are outside ±2, and about 1% are outside ±3." Second, FACETS® produces two reliability indices of particular interest, ‘R’ which is analogous to Cronbach’s alpha (and ranges from 0 to 1) and ‘G’ which reports number of statistically distinct classes of rater severity, and ranges from 0 to infinity (cf. Eckes, 2015; Schumacker & Smith, 2007). Somewhat counter-intuitively, in a situation where it is desirable for raters to be interchangeable, these indices should be low as they both provide estimates of the extent to which an analysist can be sure that the suggested differences between raters are true. A low measure indicates non-significant or non-reliable differences between raters. A third measure is ‘infit,’ which can be used to estimate inter-rater reliability (Weigle, 1998). The latter has a predicted value of 1.0 and values exceeding this indicate behavior that is unpredictable to the Rasch model (values below indicate behavior that is ‘too’ predictable, for example, when raters limit their use of the scale). In line with Bond and Fox (2015, p. 273), we treat values between 0.4 and 1.2 as acceptable. Third, FACETS® reports logit values for each element, making it possible to track the relative distance between raters as well as the order of raters from pretest to posttest.

6. Results and Discussion

Data analysis revealed findings about changes in the assessments of L2 oral English proficiency. Results about variation in raters’ assessments (pre- and posttest) are presented first, followed by what was found with regard to the calibration of assessments against benchmarks.

6.1 Displayed variation in assessments (RQ1)

The results (Table 1) show the change in raters’ agreement on grades from the pretest to the posttest. Agreement is expressed as correlations and the coefficients indicate a much stronger agreement at the posttest compared with the pretest; according to Cicchetti’s (1994) guidelines, all three ICC measurements at the posttest are excellent.

Table 1. Intraclass correlation coefficients (ICC) for aspects and the holistic test grade at the pre- and posttest

Assessment	**Pretest**			**Posttest**		
	Student performance (N)#	ICC	CI 95%	Student performance (N)#	ICC	CI 95%
Content	10	.59	.37, .84	10	.82	.67, .94
Language and expression	10	.54	.32, .81	10	.85	.71, .95
Test grade	10	.58	.35, .85	10	.83	.68, .95

#*N* = Number of student performances that was rated.

The results of the MFRM analysis showed acceptable overall fit. At the pretest, 5.5 % and 0.5 % standardized residuals exceeded 2.0 and 3.0, respectively. At the posttest, 5 % and 0 % standardized residuals exceeded 2.0 and 3.0, respectively.

The results of the MFRM at group level shown in Table 2 comprise three reliability measures: reliability index (*R*), separation index (*G*), and exact agreement (%). Table 2 also contains the expected agreement as modelled by the MFRM analysis.

Table 2. Results of Multi-Facet Rasch Analysis at group level

Measure	**Pretest**		**Posttest**	
	Students	Raters	Students	Raters
R (reliability index)	.97	.86	.99	.89
G (separation index)	5.8	2.4	9.1	3.5
% (exact agreement)	n/a	36.0	n/a	39.5
% (expected agreement)	n/a	37.5	n/a	40.7

As shown in Table 2, the reliability index increases for both students and raters from the pretest to the posttest. For the students this means that the raters as a group could distinguish between them with even higher consistency at the posttest. The increase in reliability for the raters, however, must be interpreted as a decrease in rater consistency: the severity differences between the raters were more distinct at the posttest than at the pretest. Further, the spread of assessments between the raters increases from the pretest to the posttest, as indicated by the higher *G* value (3.5 compared with 2.4). This finding indicates that the difference between the raters in terms of severity was more profound at the posttest. However, while raters moved further apart, there was an increase in consistency. Exact agreement increased by almost four percentage points, but remained below expected agreement.

The results of a simple Rasch analysis for the assessments of all participants are shown in Table 3. The infit statistic indicated few overall discrepancies between pre- and posttest. The statistic ranged from 0.30 to 1.69 on the pretest and from 0.48 to 1.77 on the posttest. There were three raters who had high infit values at the pretest and two who had it at the posttest. However, there were some noticeable individual changes. Rater 206 assessed more consistently at the posttest, decreasing infit from 1.44 to 0.96. Rater 208 showed a similar decrease. Rater 210, however, showed an increase, thus rating more inconsistently at the posttest. The rating of Rater 205 was inconsistent at the pretest, and remained inconsistent.

What is perhaps most interesting to compare is the rank of severity of each participant (Rater ID) at the pretest and at the posttest. As Table 3 shows, two participants remained in the exact same rank position (Raters 202 and 212), whereas others moved several positions. Rater 204 was ranked 2 at the pretest but 9 at the posttest, which indicates that in comparison with the group, her assessments were more generous at the posttest (cf. Rater 208, from 3 to 8, and Rater 210, from 2 to 6). Others moved in the other direction of the severity scale, for instance, Rater 201 (from rank 9 to rank 3), Rater 205 (from 6 to 1) and Rater 211 (from 10 to 5), while the remaining raters were ranked similarly on both occasions (Raters 203, 206, and 209).

When comparing all raters' assessments from the pretest with the posttest in terms of making use of the full range of assessment possibilities 'offered' by the 6-graded scale, a noticeable change appeared (see Table 4). For example, at the posttest, proportionally more A grades (25 %) were awarded by the raters as compared with at the pretest (7 %). In general, the group was leaning towards employing and assigning only a few grades to begin with (i.e., at the pretest), rather than making use of the full range of grades. A close look at the individual raters (not reported in Table 4) revealed a particularly interesting case; while Rater 208 did not award a single grade A at the pretest (0 %), she awarded 35 % As at the posttest. Although speculative, it is possible that the training emboldened this rater to use a wider range of the scale at the posttest.

Table 3. Results of simple Rasch analysis, pre- and posttest

Rater ID	**Observed Average**		**Fair Average**		**Logit**		**Stand. Error**		**Infit**		**Rank**	
	Pre	Pos	Pre	Pos	Pre	Pos	Pre	Pos	Pre	Pos	Pre	Pos
201	3.75	3.40	3.62	3.12	-.59	1.50	.29	.35	.71	1.00	9	3
202	4.15	4.60	4.11	4.92	-1.26	-1.81	.29	.40	1.06	.63	11	11
203	3.55	4.60	3.36	4.92	-.26	-1.81	.29	.40	.68	.82	8	10
204	3.00	4.45	2.68	4.67	.69	-1.33	.30	.40	1.17	.55	2	9
205	3.45	3.25	3.22	2.91	-.09	1.88	.29	.35	1.69	1.77	6	1
206	3.15	4.15	2.85	4.20	.43	-.44	.30	.38	1.44	.96	5	7
208	3.00	4.25	2.68	4.35	.69	-.73	.30	.38	1.42	.57	3	8
209	2.80	3.35	2.49	3.05	1.06	1.63	.31	.35	1.11	.94	1	2
210	3.35	3.95	3.05	3.91	1.63	.11	.35	.37	.54	1.51	2	6
211	4.10	3.85	4.05	3.77	-1.18	.37	.29	.36	1.00	.48	10	5
212	3.05	3.75	2.74	3.63	.60	.63	.30	.36	.30	.88	4	4
M	3.40	3.96	3.18	3.95	.00	.00	.30	.37	1.01	.92	n/a	n/a
SD	.45	.49	.56	.73	.78	1.34	.01	.02	.42	.40	n/a	n/a

Note. Observed average = observed raw score average. Fair average = score on original scale based on logit value. Logit = value on logit scale after Rasch analysis. Standard error = individual standard error of logit value. Infit = infit statistic indicating predictability by MFRM model.

Table 4. Scale use: percentage of awarded grades, all raters (N = 11).

Grade	Pretest	Posttest
A	7%	25%
B	14%	18%
C	27%	16%
D	20%	13%
E	31%	24%
F	1%	4%
Sum	100%	100%

It ought to be repeated that the test recordings used during the posttest day were not identical to the ones used on the pretest day (except for the anchor student performances). Thus, one might suspect that the new set of test recordings would be another possible explanation for our findings. However, as mentioned, the order of the test recordings on each day was carefully considered in the design, including taking the quality of the test-taker performances into account. Therefore, the fact that the raters assessed different tests/student performances does not constitute a likely alternative explanation for our results.

6.2 Changed assessment practices and calibrating assessments against benchmarks (RQ2)

To find out to what extent the participants improved in terms of calibrating their assessments closer to benchmarks upon having taken part in the rater training program, we compared the results from the pretest (Table 5) with the posttest (Table 6).

Table 5. Rater results from the pretest day in comparison with benchmarks

Recording ID	Test order	Student	Benchmark#	Mean grade (11 raters)	Difference
2	1	Girl	4	4	0
		Boy	5	4	1
11	2	Girl	3	3	0
		Boy	2	2	0
43*	3	Girl	5	4	1
		Boy	6	5	1
31	4	Girl	2	2	0
		Boy	3	3	0
14	5	Girl	5	4	1
		Boy	4	4	0
Mean					**.40**

*Anchor test
#A = 6, B = 5, C = 4, D = 3, E = 2, F = 1

Table 6. Rater results from the posttest day in comparison with benchmarks

Recording ID	Test order	Student	Benchmark#	Mean grade (11 raters)	Difference
43*	1	Girl	5	5	0
		Boy	6	6	0
32	2	Girl	4	5	-1
		Boy	3	4	-1
13	3	Girl	2	2	0
		Boy	2	2	0
33	4	Girl	6	6	0
		Boy	5	5	0
21	5	Girl	3	3	0
		Boy	2	2	0
Mean					**-.20**

*Anchor test
#A = 6, B = 5, C = 4, D = 3, E = 2, F = 1

Whereas our sample was rather close to the benchmark already at the pretest (mean difference: .40) (Table 5), it was even closer at the posttest (mean difference: -.20) (Table 6). This means that, based on the ten student performances assessed at the pre- and posttest respectively, the mean grade from our 11 raters differed on four occasions on the pretest (more strict) and on two occasions on the posttest (more lenient). It can be mentioned that the assessments for the anchor test were on benchmark at the posttest.

As for making a contribution to the field, on the one hand, this study corroborates findings from similar investigations; raters do differ. On the other hand, it is also a contribution to a growing body of research on how rater behavior can be identified and subject to change. This is indeed interesting, as it suggests that it may be worthwhile to continue to pursue rater training as a means for increased equity in assessment, rather than opting for other alternatives such as selected response formats or superfluous ratings by computers.

7. Implications and Limitations

The study has some important practical implications. First, teachers in primary school who teach English ought to be offered continuous professional development that focuses on the assessment of complex language abilities, including L2 oral proficiency.

Second, an assessment module in primary school teacher education programs should be linked to the major/specialization (e.g., English) and be compulsory. In the Swedish teacher education context, there is a compulsory but modest 5-week assessment module offered as part of core (but not subject-specific) teacher education courses. In other words, future primary school teachers are not guaranteed an opportunity to develop subject-specific assessment skills and knowledge, such as the ability to assess L2 English oral proficiency (unless there is a local university decision to include such content within the major/specialization). In addition, there is an apparent risk in the present system of higher education that future primary school teachers mainly get to study and learn about assessment at a theoretical and general level – never at a specialized level, which would encompass learning about assessment specifically in relation to English, both as regards theoretical and practical aspects of assessment. In relation to this topic, it can be mentioned that, for the school year 2019/2020, official statistics from the Swedish National Agency for Education shows that 82.1 percent of all primary school teachers (grades 4-6) were certified teachers. Thus, the remaining teachers are not certified, yet they may still be teaching English and they may still be administering and assessing the NEST. It is a well-known problem in Sweden that quite a large proportion of primary school teachers can be assigned to teach (and assess) English even though they do not have the adequate qualifications. Although the primary school teacher education program in Sweden is used as an example here, the situation is similar in other national contexts (for Norway, see Charboneau Stuvland, 2019; see also Nikolov, 2009). For the group of primary school teachers who lack subject-specific training and qualifications, we recommend that responsible authorities provide adequate support, if these teachers are expected to be involved in formal assessment, especially in high-stakes assessment, such as the NEST.

Third, it should be pointed out that L2 English teachers are regularly involved in the assessment of their students' oral proficiency, as this is part of any language teacher's daily practice (McKay, 2006). Further, it is not unusual that L2 English teachers, or other language teachers for that matter, have to take on a role as gatekeepers. For example, they may be responsible for deciding whether a student can be deemed eligible for taking a certain oral proficiency exam (Swedish National Agency for Education, 2019).

Fourth, the individual feedback was used as a central part of the rater training program and our results indicate that it had an effect, because there was indeed change in the assessment practices. The relative standings between the raters changed, and the infit statistic revealed changed rating patterns. This implies that teachers are sensitive to this type of precise and individual feedback. However, considering the fact that we conducted a 'snapshot' study, it is of course possible that the participating teachers changed 'too much', which is something the design did not allow us to control for. This is an apparent limitation of the study. Nevertheless, the feedback was designed to help raters become aware of their own profiles as raters (that is, their rater identity), which we believe it did, but *to what extent* and *how raters more exactly were helped* by the feedback need to be examined more closely in future research, preferably by adopting a qualitative or mixed-methods design.

8. Conclusions and Future Research

In this study, we have reported on the effects of a training program intended to increase interrater reliability in the assessment of L2 English oral proficiency. All participants were primary school teachers of English with some, but not extensive, experience of assessing students' oral production and interaction, that is, the test construct in focus. Assessment of complex language abilities, such as speaking, is without a doubt a demanding and difficult task but one the participants in this study are obliged to perform as part of their profession. Taken together, the results of this study indicate that the rater training program was beneficial, not only from the perspective of equity in assessment, but also from the perspective of providing professional development for in-service teachers.

Overall, the results showed improvement in terms of consistency as well as in terms of awarding assessments more aligned with benchmarks. The intraclass correlation coefficients increased from the pre- to the posttest, presumably as a result of rater training. However, at the same time, admittedly, the differences in severity increased. In order to make substantial claims, a larger sample of raters and of ratings and a longitudinal design would be necessary, which we recommend for future research. The present study represents a small, yet promising, explorative study on rater training related to the assessment of L2 oral language, a specific area with few previous empirical studies.

It is worth emphasizing that this study contributes with new knowledge based on data from an educational context that is clearly under-represented in the literature: primary school. Two recent overviews of L2 oral proficiency assessment research both reveal that most studies are carried out at tertiary level and some at secondary level, whereas studies from the primary level are extremely scarce (Roca-Varela & Palacios, 2013; Sandlund et al., 2016). These overviews additionally show an abundance of studies targeting internationally well-established tests of L2 oral proficiency, such as TOEFL Speaking and the IELTS Speaking Test, while 'speaking tests' used in different countries are less commonly reported on, even though they have huge impact in their respective educational contexts (see, e.g., East, 2015; Hasselgren, 2000). In light of the fact that L2 teachers world-wide are expected to assess their students' oral skills as part of their regular work duties, it seems important to conduct studies that tap into such core assessment practices.

Based on the results, we are confident enough to encourage scholars interested in the assessment of L2 oral proficiency and in offering training for professionals and/or student teachers to design test-specific training programs that include the components presented here, such as core concepts in language assessment, individual feedback, and collaborative work. Further, we suggest that our findings are relevant not only to the research community, but also to important stakeholders, such as teachers, students, and teacher educators.

References

Adie, L. E., Klenowski, V., & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgment in the context of social moderation. *Educational Review*, *64*(2), 223–240. doi:10.1080/00131911.2011.598919

Berge, K. L. (2005). Tekstkulturer og tekstkvaliteter. In K. L. Berge, L. S. Evensen, F. Hertzberg, & W. Vagle (Eds.), *Ungdommers skrivekompetanse, Bind II, Norskeksamen som tekst* (pp. 11–190). Oslo: Universitetsforlaget.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3 ed.). New York, NY: Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110. Retrieved from http://ltj.sagepub.com/content/20/1/89.abstract. doi:10.1191/0265532203lt245oa

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment. Principles and classroom practices* (2 ed.). White Plains, NY: Pearson Education.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383–391. Retrieved from http://dx.doi.org/10.1111/j.0083-2919.2005.00419.x. doi:10.1111/j.0083-2919.2005.00419.x

Charboneau Stuvland, R. A. (2019). PhD revisited: Approaches to English as a foreign language (EFL) reading instruction in Norwegian primary schools. In U. Rindal & L. M. Brevik (Eds.), *English Didactics in Norway - 30 years of doctoral research* (pp. 229-251). Oslo.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Connolly, S., Klenowski, V., & Wyatt-Smith, C. M. (2012). Moderation and consistency of teacher judgement: Teachers' views. *British Educational Research Journal*, *38*(4), 593–614. Retrieved from http://dx.doi.org/10.1080/01411926.2011.569006. doi:10.1080/01411926.2011.569006

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396. doi:10.1177/0265532209104667

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=42527179&site=ehost-live.

EACEA. (2009). National testing of pupils in Europe: Objectives, organisation and use of results. Retrieved from https://publications.europa.eu/s/lJkp

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*(1), 101–120. Retrieved from http://ltj.sagepub.com/content/early/2014/08/12/0265532214544393.abstract. doi:10.1177/0265532214544393

East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation*. Singapore: Springer.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2 ed.). Frankfurt am Main: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*(3), 175–196.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64.

Engelhard, G. (2013). *Invariant measurement*. New York, NY: Routledge.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119. doi:10.1080/15434300801934702

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585–602. doi:10.1177/0265532210364049

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust?—teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*(1), 69–87. doi: 10.1007/s11092-013-9158-x

Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: an innovative approach. *Language Testing, 17*(2), 261–277. doi:10.1177/026553220001700209

Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative-nonnative interaction in Japanese as a foreign language. *System, 29*(2), 267–287. doi:10.1016/S0346-251X(01)00015-X

Jølle, L. J., & Skar, G. B. (2018). “Digging for Gold” or “Sticking to the Criteria”: Teachers’ rationales when serving as professional raters. *Scandinavian Journal of Educational Research.* doi:https://doi.org/10.1080/00313831.2018.1541821

Jönsson, A., & Thornberg, P. (2014). Samsyn eller samstämmighet? En diskussion om sambedömning som redskap för likvärdig bedömning i skolan. *Pedagogisk forskning i Sverige, 19*(4-5), 386–402.

Klenowski, V., & Adie, L. E. (2009). Moderation as judgement practice: Reconciling system level accountability and local level practice. *Curriculum Perspectives, 29*(1), 10–28.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3 ed.). New York, NY: Springer.

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly, 4*(4), 313–335. doi:10.1080/15434300802457513

Leclercq, P., & Edmonds, A. (2014). How to assess L2 proficiency? An overview of proficiency assessment research. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 3–23). Bristol: Multilingual Matters.

Linacre, J. M. (2017a). Facets® (version 3.80.0) [Computer Software]. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2017b). A user's guide to FACETS. Rasch-model computer programs. Program manual 3.80.0. Retrieved from http://www.winsteps.com/a/Facets-ManualPDF.zip

Linacre, J. M. (2020). Facets computer program for many-facet Rasch measurement, version 3.83.2. Beaverton, OR: Winsteps.com.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127–145. doi:10.1080/154303.2011.565845

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. New York, NY: Longman.

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing, 18*(4), 333-349. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=7393260&site=ehost-live.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. London: National Assessment Agency.

NAFS Project. (2012). Ämnesprov i engelska för årskurs 6 [English National Test Year 6]. Retrieved from http://nafs.gu.se/prov_engelska/grundskolan/ap6

National Assessment Project. (2015). National Assessment Project. Retrieved from http://nafs.gu.se/english/?languageId=100001&disableRedirect=true&returnUrl=http%3A%2F%2Fnafs.gu.se%2F

Nikolov, M. (Ed.) (2009). *Early learning of modern foreign languages: Processes and outcomes*. Bristol: Multilingual Matters.

Nusche, D., Halász, G., Looney, J., Santiago, P., & Shewbridge, C. (2011). OECD reviews of evaluation and assessment in education. Sweden. Retrieved from https://www.oecd.org/sweden/47169533.pdf

Papajohn, D. (2002). Concept Mapping for Rater Training. *TESOL Quarterly, 36*(2), 219–233. Retrieved from http://dx.doi.org/10.2307/3588333. doi:10.2307/3588333

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Roca-Varela, M. L., & Palacios, I. M. (2013). How are spoken skills assessed in proficiency tests of general English as a Foreign Language? A preliminary survey. *International Journal of English Studies, 13*(2), 53–68. doi:10.6018/ijes.13.2.185901

Sadler, D. R. (2013). Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5–19. Retrieved from http://dx.doi.org/10.1080/0969594X.2012.714742. doi:10.1080/0969594X.2012.714742

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second language oral proficiency testing. *Language and Linguistics Compass, 10*(1), 14–29. doi:10.1111/lnc3.12174/epdf

Schumacker, R. E., & Smith, E. V. (2007). Reliability. A Rasch Perspective. *Educational and Psychological Measurement, 67*(3), 394–409. doi:10.1177/0013164406294776

Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: Investigation of a long-term writing assessment program. *L1 Educational Studies in Language and Literature, 17*(Open Issue), 1–30. doi:http://doi.org/10.17239/L1ESLL-2017.17.01.06

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development*. Oxford: University of Oxford Centre for Educational Assessment.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing, 35*(2), 217–238. doi:10.1177/0265532217690782

Swedish National Agency for Education. (2011a). *Curriculum for the compulsory school, preschool class and leisure-time centre 2011*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2011b). *Kommentarmaterial till kursplanen i engelska*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2013). *Bedömarträning - engelska årskurs 6. Handledning*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2015). *English. Ämnesprov, läsår 2014/2015. Lärarinformation inklusive bedömningsanvisningar till Delprov A. Årskurs 6 [English. National Test, 2014/2015. Teacher information including assessment instructions for Part A. Year 6]*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2019). Bedömning i svenska för invandrare [Assessment in Swedish for Immigrants]. Retrieved from https://www.skolverket.se/a-o/landningssidor-a-o/bedomning

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44*(2), 307–336. Retrieved from http://dx.doi.org/10.1111/j.1467-1770.1994.tb01104.x. doi:10.1111/j.1467-1770.1994.tb01104.x

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching, 3*(1), 108–112

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. doi:10.1177/026553229801500205

Wiliam, D. (2007). Content then process: Teacher learning communities in the service of formative assessment. In D. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 182–204). Bloomington, IN: Solution Tree.