JYX
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Zansen, Anna von; Huhta, Ari

Title: Developing Automated Feedback on Spoken Performance : Exploring the Functioning of Five Analytic Rating Scales Using Many-facet Rasch Measurement

Year: 2022

Version: Published version

Copyright: © 2022 Authors and University of Jyväskylä

Rights: CC BY 4.0

Rights url: https://creativecommons.org/licenses/by/4.0/

Please cite the original version:

Zansen, A. V., & Huhta, A. (2022). Developing Automated Feedback on Spoken Performance : Exploring the Functioning of Five Analytic Rating Scales Using Many-facet Rasch Measurement . In J. H. Jantunen, J. Kalja-Voima, M. Laukkarinen, A. Puupponen, M. Salonen, T. Saresma, J. Tarvainen, & S. Ylönen (Eds.), Diversity of Methods and Materials in Digital Human Sciences : Proceedings of the Digital Research Data and Human Sciences DRDHum Conference 2022, December 1-3, Jyväskylä, Finland (pp. 211-229). Jyväskylän yliopisto.
http://urn.fi/URN:ISBN:978-951-39-9450-1

Developing Automated Feedback on Spoken Performance: Exploring the Functioning of Five Analytic Rating Scales Using Many-facet Rasch Measurement

Anna von Zansen, Ari Huhta
University of Helsinki, University of Jyväskylä
E-mail: anna.vonzansen@helsinki.fi

Abstract

In this study, we used the Many-facet Rasch measurement (MFRM) to explore the quality of ratings as well as the functioning of five analytic rating scales developed for automated assessment of L2 speech. This study is part of a multidisciplinary research project that develops automatic speech recognition (ASR), automated scoring and automated feedback for L2 Finnish and Swedish. The data include the analytic ratings (task completion, fluency, pronunciation, range, accuracy) gathered from human raters (n=14) who assessed L2 Finnish learners' (n=64) speech samples using Moodle. The four-facet Rasch analysis showed that the raters performed and the rating scales functioned well, although task completion seems to be more challenging to apply consistently than the other criteria. Moreover, it proved to be more difficult to receive a certain score on some dimensions, namely fluency and range, than others. The study has implications for score reporting. We demonstrated that a) the different analytical rating scales have somewhat different structure, b) scores do not advance with equal intervals and c) a certain score on a certain dimension might require a bigger leap forward in ability than on other dimensions. The results will be used for designing encouraging and accurate automated feedback to L2 Finnish and Swedish learners.

Keywords: automated feedback, language assessment, rating scales, oral skills

1. Introduction

Automated speech processing technology has improved and become more popular in everyday life contexts. Also, the technologies for automated assessment of speaking skills have made considerable progress in recent years. Automated language assessment has many advantages: not only can it save time and money, but it can also standardize the scoring process. However, automated systems still have many limitations, for example, regarding construct coverage. Therefore, a hybrid approach that combines human and automated scoring is likely to be the most feasible solution (see Evanini & Zechner, 2020, 3–4; Xu et al. 2020).

As Gu & Davis (2020, 159) point out, automated speech processing technology can be used to provide immediate and individualized diagnostic feedback to L2 learners, regardless of time and place. Moreover, automated systems can give such feedback instantly (Zhang et al. 2020, 21). Automated feedback technologies are emerging also in language learning contexts, where tutoring systems can provide immediate and specific feedback or instruction to the learner (Golonka et al. 2014, 73). However, most of the automated language learning tools deal with written language and grammar (for a review of educational feedback systems see Deeva et al. 2021).

Turning to L2 speaking, most ASR-based software for training speaking are limited to computer-assisted pronunciation training (Golonka et al. 2014, 81), although de Vries et al. (2015) present an ASR-based system developed for practicing word order in Dutch. Many automated speech training systems such as EduSpeak, NativeAccent, English Discoveries and Duolingo provide feedback mainly on pronunciation (Gu & Davis 2020, 159–160). Nevertheless, some automated systems that provide feedback on spontaneous speech exist but the tools are often aimed only for L2 English learners. For example, in the context of TOEFL Practice Online Test, Gu & Davis (2020) describe the development of automated feedback on seven features related to speaking, whereas Xu et al. (2020) present validity argument for the Linguaskill Speaking Test which combines auto-scoring and human rating to produce a CEFR grade to the L2 English learner.

This study is part of the DigiTala research project (2019–2023) which develops an automated tool for assessing L2 Finnish and L2 Swedish learners' oral skills (see Kautonen & von Zansen 2020). In this multidisciplinary project, experts of pedagogy, technology and phonetics develop automatic speech recognition, automated scoring and automated feedback (see Evanini & Zechner 2020) for assessing L2 Finnish and Swedish learners' oral skills.

The research project has two goals: 1) to pave way for implementing a speaking section to the language tests of the Finnish Matriculation Examination (Vaarala et al. 2021) and 2) to develop an online tool for self-regulated learning purposes. The study reported here relates mostly to the second goal. The aim of the automated diagnostic feedback is to help both independent learners and learners with access to teacher support to develop their speaking skills by providing information about the strengths and weaknesses in their performance.

In this study, we use Many-facet Rasch Measurement (MFRM, see McNamara et al. 2019; Boone et al. 2014; for a review of Rasch measurement in language assessment see Aryadoust, Ng & Sayama 2021) to explore the functioning of the analytic rating scales which are used for designing diagnostic feedback on speech performances.

1.1. The Moodle plugin

The project designed a Moodle plugin (von Zansen et al. 2022) that records L2 learners' responses to a speaking task and displays automatically rated scores to the learner. Currently, the task types include read-aloud and spontaneous speech (up to 3 minutes). When the system receives a speech sample, it uses automatic speech recognition to produce a transcript of the sample. Then the system produces automatic scores on selected dimensions of speech and finally shows the results to the learner.

The Moodle plugin's (von Zansen et al. 2022) frontend is described in a user manual that presents the Moodle plugin in detail (Alanen et al. 2022). Moreover, a short video is available on the Github page (von Zansen et al. 2022) and a screenshot of the learner's report page is available in Appendix 1.

For the backend, we have trained automatic assessment systems using Finnish and Swedish learners' speech samples that were rated by human raters. We follow a feature-based approach, which enables the production of feedback on different dimensions of speech. However, we are also exploring whether better results could be achieved by applying deep learning methods, "the black box approach" (Al-Ghezi et al. forthcoming).

For read-aloud samples, the system produces scores for fluency and pronunciation while also showing the transcript of the sample to the learner and pointing pronunciation errors. For spontaneous samples, the system provides more detailed feedback: a transcript of the sample combined with analytic scores on fluency (e.g. breaks and repetitions on a 0–4 scale), pronunciation (control of sound and prosodic features on a 0–4 scale), task completion (does the speaker answer the question on a 0–3 scale) and range (extent of vocabulary, structures and expressions on a 0–3 scale) as well as an estimation of the proficiency level (from below A1 to C2 on the Common European Framework of Reference scale, see Finnish National Agency for Education 2003; Council of Europe 2001). The analytic scales include dimensions that human raters are familiar with and hat can be measured automatically (see Kautonen & von Zansen 2020 and section 2 for scale development).

In addition to the automated scoring, teachers have the possibility to comment on the scores produced by the machine. Finally, teachers or researchers can export the learners' speech samples and their scores. The rating data together with the speech samples are important for us in the future when we evaluate the reliability of the automated system (see also Evanini & Zechner 2020, 13).

1.2. Quality of ratings

In this study, we investigate the "quality of ratings", which refers to (a) raters' performance and (b) functioning of the rating scales.

Regarding rater performance, the Facets programme provides information on raters' relative severity, that is, how severely (or leniently) they rate compared to the other raters in the sample (McNamara et al. 2019, 108). Rater severity is not necessarily a significant concern in this project, since we use fair averages to train the automatic scoring system. Fair averages produced by Facets are scores adjusted for rater severity / leniency and they are, thus, more accurate indicators of learner ability than regular (raw) averages calculated across the ratings given to a particular speech sample. Second, Facets produces rater fit statistics, which are informative of rater consistency (McNamara et al. 2019, 109). In this study, we use the range 0.5–1.5 for acceptable fit statistics recommended by Linacre (2002a). High mean-square values (above 1.5) indicate misfit meaning that the rater performs inconsistently. Low mean-square values (below 0.5) indicate that a rater overfits the model which means that the rater shows less variation than was expected, possibly due to halo or central tendency effects (see McNamara et al. 2019, 109). In general, very inconsistent raters (as indicated by above 1.5 mean-square values) degrade the dependability of the rating data. However, also extremely severe or lenient raters are problematic since Facets can adjust the fair average score only up to a point – such extreme cases need to be spotted by visually inspecting Facets output and decisions need to be made whether to remove them from the data).

Rating scale functioning is the second focus area of this study, as we plan to use the rating scales as a starting point when providing automated feedback to the learners. Linacre (2002b) recommends following guidelines[79] for optimizing the functioning of a rating scale. For stable and precise estimates, each score category should have over ten observations (guideline 1). For optimal step calibration, the observations should be regularly distributed across the score categories (guideline 2). Furthermore, average measures should advance monotonically (guideline 3), in other words, higher score observations produce higher measures. In addition, outlier-sensitive MNSQs should be less than 2.0 (guideline 4) since score categories with larger Outfit MNSQs indicate too much randomness ("noise") and are therefore not useful for the measurement (see also Linacre, 2002a). According to the guideline 5, step calibrations must advance, that is, high measures are observed in the highest categories and vice versa. Disordering of step calibration may occur if the construct (speaking ability) is not well defined, or a score category reflects too narrow part of it. Finally, step difficulties (Rasch Andrich thresholds) should advance by at least 1.4 logits (guideline 6), yet less than by 5.0 logits (guideline 7). These guidelines are helpful when evaluating the functioning of rating scales. Sometimes scale revision such as combining neighbouring categories, might be needed, if raters cannot distinguish between such categories (see Linacre 2002b; McNamara et al. 2019, 70–78.)

1.3. Ongoing research and research questions of this study

To develop automated assessment of L2 learners' oral skills, we have followed the stages described in Figure 1. First, we analyzed human ratings after receiving and transcribing the speech samples from L2 Finnish learners. However, these analyses served a different purpose, that is, converting the ordinal rating scale data to linear measures (by using Facets analysis) in order to receive a fairer and more accurate score for each speech sample (see Boone et al. 2014). After these analyses, various machine learning methods are applied to the speech samples and their transcriptions in order to predict

[79] Guidelines renumbered 1–7; guideline 6 (Linacre 2002b) omitted from this study due to the complexity of the analysis

the human ratings. Emerging results (Al-Ghezi et al. forthcoming) suggest that the automated system could predict most of the analytical ratings statistically significantly. The prediction was best for the fluency ratings (Spearman correlation 0.47 for Finnish and 0.23 for Swedish) followed by range (0.28 for Finnish and 0.20 for Swedish) and accuracy (0.22 for Finnish and 0.18 for Swedish). For pronunciation, the correlation between human and machine ratings were significant for Swedish (0.17) but not for Finnish.

In addition to the rating data, we plan to take stakeholders' perceptions (see von Zansen et al. accepted; von Zansen, Sneck & Hilden accepted a; b) into account when designing automated feedback. Moreover, we are interviewing learners and teachers in order to investigate the usefulness and understandability of the automated feedback (von Zansen & Heijala forthcoming).

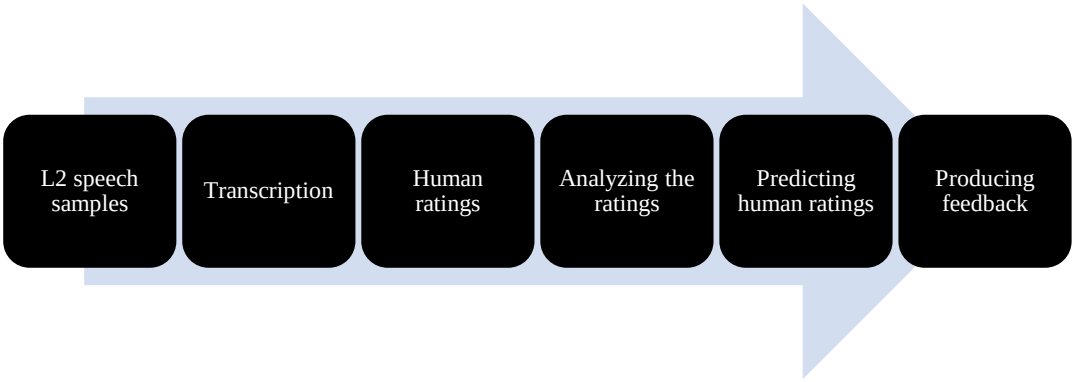


Figure 1: Stages of development

However, we have not yet investigated in detail how the raters performed nor how the scales functioned. Therefore, to address the research gap and to provide evidence for the validity of the human ratings that are important for the overall validity of the automated system, the study seeks answers to two main research questions (RQ): 1. What was the quality (i.e., consistency and agreement) of the ratings across the different analytic scales? 2. How did each analytic rating scale function as a scale?

In other words, we focus on the fourth and sixth stages presented in Figure 1. Results of the analyses support mainly carrying out the last stage (see Figure 1), since the analytic rating scales can be used as part of the automated feedback to learners and as a starting point for developing even more fine-grained feedback on a range of specific features of speech. The results of this explorative study serve proof-of-concept purposes.

2. Methods

The data of this study include ratings gathered from human raters (n=14) during the third rating round organized by the project in June 2021. Speech samples were rated by using a holistic (below A1–C2) and five analytic (task completion, fluency, pronunciation, range, accuracy) rating scales (see von Zansen 2022a) using Moodle. In Moodle, the raters listened to one sample at a time and provided both the holistic and analytic scores in the same window. We used a partially overlapping rating design where some of the samples (n=913) were systematically routed for two or multiple raters to rate while some were rated by a single rater. This way we saved human resources and were still able to investigate and compare all the ratings simultaneously with Facets (Linacre 2021). As a result, in addition to investigating the quality of ratings (RQ1) and scales (RQ2), by using Facets, we obtained fairer scores for training the automatic scoring system (see sections 1.2 and 1.3). During this rating

round we collected over 7500 ratings in total, of which 1030 were holistic scores. For details concerning rater training and instructions see von Zansen, Sneck and Hilden (accepted b).

In scale development, we used the level descriptors of the previous National Core Curriculum (Finnish National Agency for Education 2003) as a starting point for several reasons (see also Kautonen & von Zansen 2020). First, as they came from the National Curriculum and as they are local applications of the Common European Framework (Council of Europe 2001) descriptors, the scale is well-known both nationally and internationally. Secondly, this allowed us to address the first goal of the research project, that is, enabling implementing a speaking section to the language tests of the Finnish Matriculation Examination which aims to measure the outcomes of the level of education regulated by the above mentioned National Curriculum (see section 1). Third, the chosen descriptors suit assessment purposes in general as they describe learner skills in sufficient detail. Thus, their detailed, analytical nature makes them applicable to be used in automated scoring and feedback.

The rated Finnish language samples were collected during spring 2021 from upper secondary school students (n=64) using speaking tests (von Zansen 2022b, 2022c) targeting B1 and B2 levels of the Common European Framework of Reference (CEFR, Council of Europe 2001). Altogether, the tests consisted of eight different tasks and 26 subtasks (von Zansen 2022b, 2022c). Since the data collection took place during the COVID-19 pandemic, data from both raters and L2 Finnish speakers were collected using Moodle and Zoom (see Al-Ghezi et al. forthcoming; von Zansen, Sneck & Hilden accepted a).

To explore the functioning of the analytic rating scales, the ratings were analyzed using Many-facet Rasch Measurement (MFRM, see McNamara et al. 2019) using Facets version 3.83.5 (Linacre 2021). In this four-facet Rasch analysis we included 1) learner's speaking ability (64 speakers), 2) task difficulty (26 tasks), 3) rater severity (14 raters), and 4) difficulty of the criteria (five analytic criteria).

We used a partial credit model to the fourth facet (see McNamara et al. 2019, 115–116), which enables modelling each of the analytic criteria to have its own scale structure. This yields more information about the scales than the rating scale model that assumes all the scales to have the same structure (McNamara et al. 2019, 115–116).

3. Results

To give an overview of the data, we first present the calibration of the four facets as Figure 2. After that, we present results for RQ1 and RQ2. More comprehensive results of the Facets analysis can be found in Appendix 2.

Figure 2 (available also as Table 6.0 in Appendix 2) shows the Wright map, that is, the location of the elements in each facet in relation to the other facets. The measurement scale ("Measr") is an interval scale that ranges from -2 to +4 logits in this analysis and provides a common yardstick against which all the facets and all their elements can be compared. We allowed the first facet (learners' speaking ability) to float while other facets were anchored at zero.

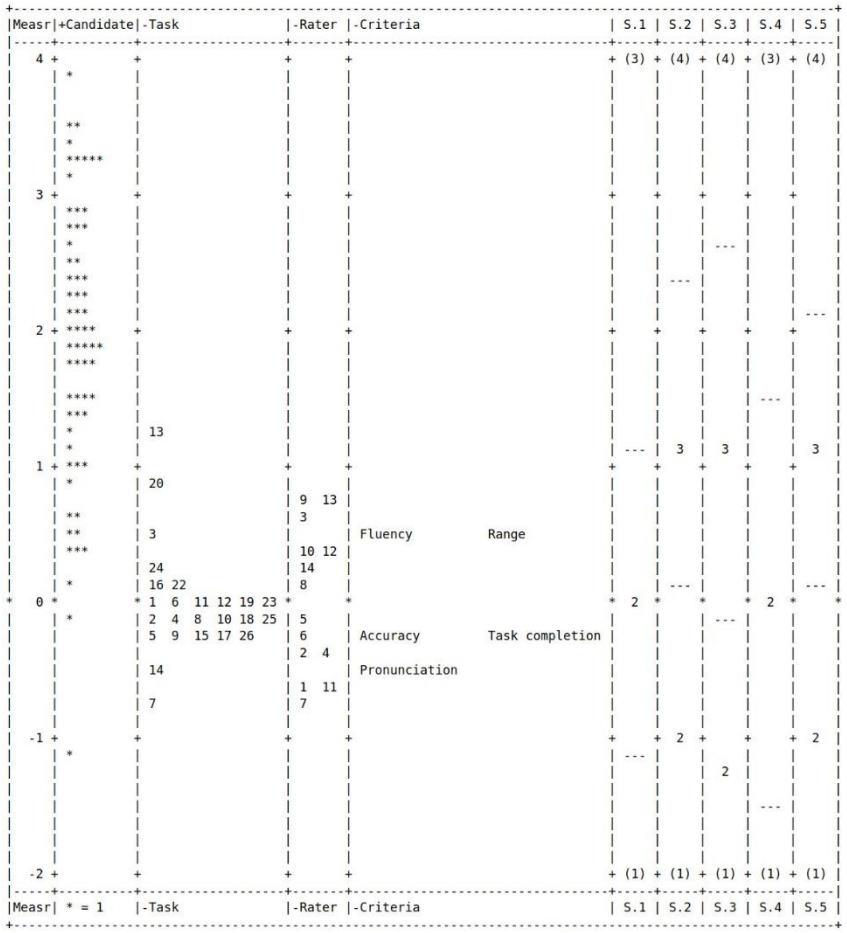


Figure 2: Wright map

In the second “Candidate” column, each star represents a test-taker showing that the learners are spread over five logits meaning that they differ considerably in their speaking ability as measured by the five analytical dimensions analysed here. Learners with higher speaking ability are at the higher end of the logit scale. The next “Task” column shows the tasks organized according to their difficulty. Based on the raters’ analytic ratings the test appears to have been quite easy as there are no tasks matching the best speakers. The tasks are spread over two logits; task 13 being the hardest and task 7 the easiest. The middle “Rater” column arranges the raters according to their severity raters 9 and 13 being the harshest and rater 7 the most lenient.

The “Criteria” column shows the relative difficulty of the five analytic criteria. We see that fluency and range are harder than accuracy and task completion and pronunciation, which is nearly one logit below the hardest criteria. In other words, it is harder for learners to receive a score on fluency or range compared to the other criteria. Furthermore, fluency and range as well as accuracy and task completion are similar in terms of difficulty.

On the right side of the Figure 2, we find each rating scale criterion (S.1 Task completion, S.2 Fluency, S.3 Pronunciation, S.4 Range, S.5 Accuracy) having a separate column. In the brackets, we see the highest and lowest scale levels, for example, the scale for task completion is 1–3 while fluency scale is 1–4. The horizontal lines in the scale columns show the points at which a learner at that logit level would score a half score (Rasch-half-point thresholds, see McNamara et al. 2019, 100). For example, speakers slightly above the logit value of 1 will likely receive a score 2.5 on task completion, 3 on fluency, 3 on pronunciation, 2 on range and 3 on accuracy.

After giving an overview of the data, we now present results regarding the quality of the ratings across the different analytic scales (RQ1). Table 7.3.1 in the Appendix 2 shows details on how the raters performed. Rater IDs are on the right column. "Total Count" shows the number of ratings performed by each rater. We notice that raters 1–4, who were researchers of the project, have provided fewer ratings than raters 5–14 recruited by the project for this rating round. The "Measure" shows raters' severity on the logit scale, which differs by 1.53 logits (rater 13 being the most severe while rater 7 being most lenient). Finally, the "Model S.E." column tells that the estimation of the rater measures is fairly precise especially for raters who provided more ratings while standard error for raters 1–4 is somewhat larger (.13–.15). The fit statistics (see columns "Infit MnSq" and "Outfit MnSq") indicate that all raters fit the model well (Infit MNSQs range .81–1.36, see Linacre 2002a). The reliability of the rater separation index (.96) indicates that the raters are a heterogeneous group. The last row of the Table 7.3.1 in the Appendix 2 shows that the inter-rater agreement was 57.8%, and that the raters agreed more than was expected by the model (51.6%).

Next, we look at findings relating to the RQ2, which deals with the functioning of the rating scales. Table 7.4.1 in the Appendix 2 shows details on the analytic scales used in this rating round. As mentioned earlier, it is more difficult to receive a score on Fluency (.51) and Range (.47) than it is on Accuracy (-.20), Task completion (-.28) or Pronunciation (-.51, see "Measure" column). Nevertheless, the fit statistics indicate that all analytic criteria are within the acceptable range (Infit MNSQs .91–1.20). However, Task completion has quite high MNSQs (Infit MNSQ 1.20, Outfit MNSQ 1.43) and seems to be more challenging to apply consistently than the other criteria. This finding is supported also by the list of unexpected responses (see Table 4.1 in the end of Appendix 2), where the majority of problematic ratings relate to Task completion. Finally, the separation ratio ("Separation" 5.57) and the separation index ("Strata" 7.76) indicate that the analytic dimensions differ in difficulty. The high criteria separation index ("Reliability" .98) shows that the test is measuring different dimensions of speaking rather than speaking as one unitary dimension.

We investigated the functioning of the rating scale following Linacre's guidelines (2002b). Results of the Facets analysis regarding this section can be found in Tables 8.1–8.5, see Appendix 2. Firstly, we noticed that the guideline 1 did not hold for two of the 4-pointed scales. Namely, for Pronunciation, score 1 was given only 6 times and for Accuracy, score 1 was given only 9 times. This might lead to unstable step calibration (Linacre 2002b) for those particular score levels. Second, with regard to the guideline 2, we noticed that the observations were not regularly distributed across the score categories (guideline 2). In general, lower scores (especially 1) were again given less frequently.

Third, we noticed that the average measures advanced (guideline 3). The average measures were also close to the expected values except of Task completion score category 1 (average .98, expected .60) and Pronunciation score category 1 (average 1.28, expected .45). Fourth, the guideline 4 did hold since all the outlier-sensitive MNSQs were less than 2.0. However, score 1 both in the Pronunciation (Outfit MNSQ 1.5) and Task completion (Outfit MNSQ 1.7) scales seems to have more noise that was expected. Fifth, we investigated the probability characteristic curves, and noticed that the score categories appear as a range of hills, indicating that guideline 5 holds. However, we observed one average measure being disordered (score category 2 for Pronunciation, average measure 1.05*, see Appendix 2, Table 8.3), presumably because there were only six observations for the lowest category one, even though it is also possible that the definitions of categories one and two are not clear enough. Sixth, investigation of the Rasch Andrich thresholds showed that guidelines 6 and 7 hold: the step difficulties advanced at least by 1.4 logits but less than 5.0 logits.

4. Discussion

Automated feedback systems are becoming common also in language assessment (Deeva et al. 2021), yet most of the tutoring systems (see Golonka et al. 2014) focus on written language and grammar or target a narrow aspect of speaking, such as pronunciation or word order (de Vries et al. 2015). This study extended previous research by exploring two aspects of automated feedback systems, namely rater performance (RQ1) and the functioning of several analytic rating scales (RQ2) in the context of developing an automated speech training system for L2 Swedish and Finnish learners. Unlike many existing systems, the Moodle-based tool (von Zansen et al. 2022) developed by the research project (Kautonen & von Zansen 2020) provides automated diagnostic feedback on learners' spontaneous speech performances (see section 1.1). To the best of our knowledge, this tool can be compared only with two systems targeted for L2 English learners. First, Gu & Davis (2020) have used a feature-based approach to develop automated feedback to L2 English speakers. Second, in addition to automated analytic feedback, we aim to provide an estimation of L2 speaker's CEFR level (Council of Europe 2001), which is in line with the work of Xu et al. (2020).

The experiment provided new insights into the five analytic rating scales that are, in addition to human scoring (hybrid approach, see Evanini & Zechner, 2020; Xu et al. 2020), also used for designing automated scoring and feedback. The methodological choices of the study proved to be useful for exploring the quality of ratings (RQ1) and the functioning of the five analytic scales developed by the project (RQ2). MFRM has many advantages compared to the more traditional approaches that focus on raw scores or compare pairs of raters. These include converting ordinal rating data into linear measures, creating rating designs that save resources, taking rater severity and task difficulty into account in order to compute fairer scores to the learner (see for example Boone et al. 2014; McNamara et al. 2019; Aryadoust, Ng & Sayama 2021 for a review of Rasch measurement in language assessment).

In this study, MFRM enabled investigating how the raters performed (RQ1) and the step structure of each rating scale criterion (RQ2). Regarding the quality of the ratings (RQ1), results of this study indicate that the raters performed well. The inter-rater agreement (57.8%) exceeded what was expected by the model. Furthermore, we did not observe raters misfitting. The results of the RQ1 indicate that the overall reliability of the raters was good. Recruitment of the raters had been successful, and we had provided sufficient training and instructions to the raters. Turning to the functioning of five analytic ratings scales (RQ2), we noticed that the rating scales functioned reasonably well, although task completion seems to be more challenging to apply consistently for the raters. It appears to measure a somewhat different aspect of speaking than the other scales. However, this is not surprising when we think about the content-relatedness of the concept "Task completion" compared to the more linguistic concepts "Fluency", "Pronunciation", "Range" and "Accuracy". A reasonable conclusion from this finding is that if Task completion is to be used as part of the automated feedback, it needs to be defined more clearly. It is possible, for example, that the meaning of task completion may differ somewhat depending on the particular task or task type, and, thus, ideally, different tasks may require slightly different scales for Task completion.

Ultimately, possible threats regarding the functioning of the rating scale might produce imprecise estimates, which in turn can lead to unfair decisions and conclusions. The results of the RQ2 suggest that some scale revisions might be needed. However, in the case of this study, we think that some of the observed problems result from the small sample of speakers (n=64), which is one limitation of this study. Moreover, the tasks targeted only B1 and B2 level speakers. Due to the lack of A-level speakers, the results cannot confirm whether raters would use the lower end of the scales

reliably enough. More research is needed to verify whether score 1 is likely to be used when raters assess A-level speakers' samples. Therefore, we plan to do a follow-up study with a larger sample: same criteria but 255 speakers representing different proficiency levels, responding to 13 tasks, rated by 20 human raters in spring 2022. Further research is also needed to investigate whether the human raters display bias when using the criteria developed by the research project. Otherwise, the bias of the human ratings may threaten the validity of the automated scoring (see for example Zhang et al. 2020).

In line with our assumption, it proved to be more difficult to receive a certain score on some dimensions, namely Fluency and Range, than others. As McNamara et al. (2019, 75) state, in language assessment contexts, rating scales are usually assumed to have equal steps, although certain scale steps may in fact require more, or less, progress to achieve than others. Moreover, a certain score might reflect a narrower range of abilities than others. In other words, scales do not often advance with equal intervals (McNamara et al. 2019, 75). As shown by this study, a certain score on a certain dimension might require a bigger leap forward in ability than on other dimensions. The results should be taken into account when designing initial report pages that will be shown to the learners. Basically, the goal of this explorative study was to pave way for providing encouraging and accurate automated feedback to learners on their speaking performance.

5. Acknowledgements

We would like to thank the following researchers for their help in collecting human ratings: Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik from the Aalto University and Heini Kallio from the University of Jyväskylä.

We are grateful for the software engineering students from the University of Helsinki who developed the Moodle plugin for us during spring 2022: Tuomas Alanen, Joona Erkkilä, Topi Harjunpää and Maikki Heijala.

The DigiTala project is funded by the Academy of Finland 2019–2023, and combines expertise in speech and language processing, language education and phonetics at the University of Helsinki (grant number 322619), Aalto University (grant number 322625) and the University of Jyväskylä (grant number 322965).

References

Alanen, T., Erkkilä, J., Harjunpää, T., & Heijala, M. (2022). *Digitala Moodle plugin user manual* (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.6535377

Al-Ghezi, R., Vosboinik, K., Getman, Y., von Zansen, A., Kallio, H., Akiki, C., Kuronen, M., Huhta, A. & Hilden, R. (forthcoming). *Automatic speaking assessment of Spontaneous L2 Finnish and Swedish*.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing* 38(1), 6–40. https://doi.org/10.1177/0265532220927487

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Science & Business Media.

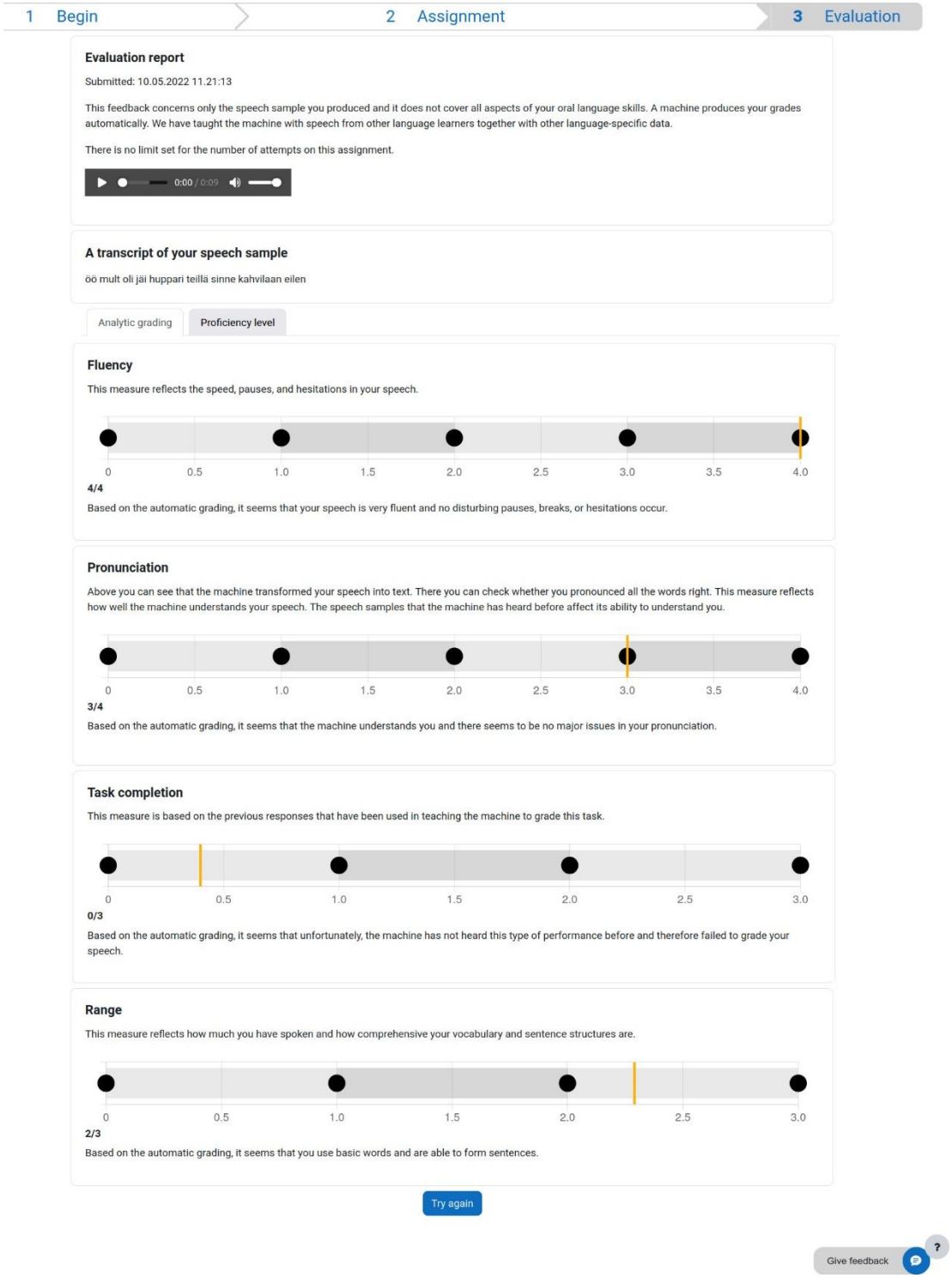
Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education* 162, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York:

Routledge, 3–20.
Finnish National Agency for Education. (2003). *Lukion opetussuunnitelman perusteet 2003 [National core curriculum for general upper secondary schools 2003]*. https://www.oph.fi/sites/default/files/documents/47345_lukion_opetussuunnitelman_perusteet_2003.pdf
Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer assisted language learning* 27(1), 70–105. https://doi.org/10.1080/09588221.2012.700315
Gu, L., & Davis, L. (2020). Providing SpeechRater Feature Performance as Feedback on Spoken Responses. In K. Zechner & K. Evanini (eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York: Routledge, 159–175.
Kautonen, M. & von Zansen, A. (2020). DigiTala research project: Automatic speech recognition in assessing L2 speaking. *Kieli, koulutus ja yhteiskunta [Language, Education and Society]* 11 (4). https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-kesakuu-2020/digitala-research-project-automatic-speech-recognition-in-assessing-l2-speaking
Linacre J. M. (2002a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 16(2), 878.
Linacre J. M. (2002b). Optimizing rating scale category effectiveness. *Journal of applied measurement* 3(1), 85–106.
Linacre, J. M. (2021). *Facets Rasch measurement* [computer program]. Chicago, IL: Winsteps.com.
McNamara, T., Knoch, U. & Fan, J. (2019). *Fairness, Justice & Language Assessment*. Oxford: Oxford University Press.
Vaarala, H., Riuttanen, S., Kyckling, E., & Karppinen, S. (2021). *Language Reserve. Now! Follow-up on Pyykkö's Report Multilingualism into a strength (2017) : Summary in English*. Jyväskylä: Centre for Applied Language Studies. https://www.jyu.fi/hytk/fi/laitokset/solki/tutkimus/julkaisut/pdf-julkaisut/summary_languagereservenow.pdf
de Vries, B. P., Cucchiarini, C., Bodnar, S., Strik, H., & van Hout, R. (2015). Spoken grammar practice and feedback in an ASR-based CALL system. *Computer Assisted Language Learning* 28(6), 550–576. https://doi.org/10.1080/09588221.2014.889713
Xu, J., Brenchley, M., Jones, E., Pinnington, A., Benjamin, T., Knill, K., Seal-Coon, G. & Geranpayeh, A. (2020). *Linguaskill Building a validity argument for the Speaking test*. Cambridge: Cambridge Assessment English. https://www.cambridgeenglish.org/Images/589637-linguaskill-building-a-validity-argument-for-the-speaking-test.pdf
von Zansen, A., Alanen, T., Al-Ghezi, R., Erkkilä, J., Harjunpää, T., Heijala, M., Kallio, H. (2022). *DigiTala Moodle plugin*. https://github.com/aalto-speech/moodle-puheentunnistus
von Zansen, A. (2022a). *DigiTala's rating criteria: Holistic and analytic scales for assessing L2 speaking. Zenodo*. https://doi.org/10.5281/zenodo.6477089
von Zansen, Anna. (2022b). *DigiTala's speaking tasks for L2 Finnish learners (proficiency level B1)*. Zenodo. https://doi.org/10.5281/zenodo.6562855
von Zansen, Anna. (2022c). *DigiTala's speaking tasks for L2 Finnish learners (proficiency level B2).* Zenodo. https://doi.org/10.5281/zenodo.6562865
von Zansen, A., Kallio, H., Sneck, M., Kuronen, M., Huhta, A., Hilden, R. (accepted). Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista [Human raters' perceptions of the automated assessment of oral language skills, the digital assessment process and the dimensions to be assessed from speaking performances]. In T. Seppälä, S. Lesonen, P. Iikkanen & S. D'hondt (eds.) *AFinLA yearbook*.
von Zansen, A., Sneck, M., & Hilden, R. (accepted a). Lukiolaisten käsitykset ja heidän antamansa palaute suullisen kielitaidon arvioinnista. [Upper secondary school students' perceptions and feedback on automated speaking assessment.] In R. Kantelinen, M. Kautonen, & Z. Elgundi (eds.). *LINGUAPEDA 2021*. Conference Proceedings. Suomen ainedidaktisen tutkimusseuran julkaisuja. Ainedidaktisia tutkimuksia 21.
von Zansen, A., Sneck, M., Hilden, R. (accepted b). *"It was cool and comfortable!"Akateemisten alkeistason S2-opiskelijoiden kokemuksia tietokoneella suoritettavasta puhumisen kokeesta ["It was cool and comfortable!" Academic L2 Finnish learners' perceptions of a computer-based speaking test]*. Ainedidaktisia tutkimuksia.
von Zansen, A. & Heijala, M. (forthcoming). *Kielten opettajien ensivaikutelmia suomen ja ruotsin oppijoiden puheen automaattiseen arviointiin kehitetystä työkalusta [Language teachers' first impressions of an automated tool developed for assessing Finnish and Swedish learners' speech]*.
Zhang, M., Bridgeman, B., & Davis, L. (2020). Validity Considerations for Using Automated Scoring in Speaking Assessment. In K. Zechner & K. Evanini (eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York: Routledge, 21–31.

Appendix 1 Screenshot of the learner's report page



Appendix 2 Results of the MFRM analysis

Table 6.0 All Facet Vertical "Rulers".

Vertical = (1*,2N,3N,4A,S) Yardstick (columns lines low high extreme)= 0,8,-2,4,End

Measr	+Candidate	-Task	-Rater	-Criteria	S.1	S.2	S.3	S.4	S.5
4 +	+	+	+	+	(3)	(4)	(4)	(3)	(4)
	*								
	**								
	*								
	*****								
	*								
3 +	+	+	+	+	+	+	+	+	+
	***								
	***								
	*						---		
	**								
	***					---			
	***								
	***								---
2 +	****	+	+	+	+	+	+	+	+
	*****								
	****								
	****							---	
	***								
	*	13							
	*				---	3	3		3
1 +	***	+	+	+	+	+	+	+	+
	*	20							
			9 13						
	**		3						
	**	3		Fluency Range					
	***		10 12						
		24	14						
	*	16 22	8			---			---
0 *	*	1 6 11 12 19 23 *	*	*	* 2	*	*	* 2	*	*
	*	2 4 8 10 18 25	5				---		
		5 9 15 17 26	6	Accuracy Task completion					
			2 4						
		14		Pronunciation					
			1 11						
		7	7						
-1 +	+	+	+	+	+	+ 2	+	+	+ 2
	*				---				
							2		

-2 +	+	+	+	+	(1)	(1)	(1)	(1)	(1)
Measr	* = 1	-Task	-Rater	-Criteria	S.1	S.2	S.3	S.4	S.5

S.1: Model = ?,?,?,1,R4 ; Criteria: Task completion
S.2: Model = ?,?,?,2,R4 ; Criteria: Fluency
S.3: Model = ?,?,?,3,R4 ; Criteria: Pronunciation
S.4: Model = ?,?,?,4,R4 ; Criteria: Range
S.5: Model = ?,?,?,5,R4 ; Criteria: Accuracy

Table 7.3.1 Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Rater
1644	582	2.82	2.83	.75	.07	1.05	.8	1.08	1.3	.88	.48	.55	53.5	47.1	13	13
1619	569	2.85	2.84	.73	.07	.90	-1.7	.91	-1.5	1.07	.55	.56	44.0	47.5	9	9
424	149	2.85	2.87	.62	.14	1.10	.8	1.41	3.0	.76	.37	.57	51.8	48.6	3	3
1721	590	2.92	2.97	.40	.07	1.11	1.9	1.06	.9	.96	.48	.54	56.3	50.2	10	10
1638	554	2.96	2.99	.36	.07	1.02	.4	1.06	.9	1.01	.60	.57	61.1	50.5	12	12
1764	591	2.98	3.03	.24	.07	.82	-3.4	.82	-2.9	1.21	.56	.53	61.4	51.0	14	14
1678	547	3.07	3.10	.06	.08	1.04	.7	1.00	.0	1.00	.50	.52	58.5	52.0	8	8
1758	571	3.08	3.15	-.10	.07	.81	-3.3	.92	-1.1	1.12	.57	.53	61.8	52.5	5	5
1867	582	3.21	3.22	-.30	.08	1.14	2.1	1.24	2.8	.88	.54	.51	58.8	53.3	6	6
470	149	3.15	3.21	-.33	.15	.85	-1.2	1.00	.0	1.14	.52	.52	61.9	53.7	4	4
472	149	3.17	3.23	-.38	.15	.84	-1.3	.79	-1.3	1.15	.59	.51	59.8	53.8	2	2
638	201	3.17	3.30	-.63	.13	.82	-1.8	.74	-1.9	1.25	.59	.48	61.4	54.0	1	1
1836	566	3.24	3.31	-.64	.08	.93	-1.0	1.01	.1	1.06	.49	.50	62.0	54.0	11	11
1807	550	3.29	3.35	-.78	.08	1.36	4.8	1.39	3.7	.71	.47	.46	57.7	54.1	7	7
1381.1	453.6	3.05	3.10	.00	.09	.99	-.1	1.03	.3		.52				Mean (Count: 14)	
562.7	185.2	.15	.17	.51	.03	.15	2.2	.20	1.9		.06				S.D. (Population)	
584.0	192.2	.16	.18	.53	.03	.16	2.3	.20	2.0		.06				S.D. (Sample)	

Model, Populn: RMSE .10 Adj (True) S.D. .50 Separation 4.98 Strata 6.97 Reliability (not inter-rater) .96
Model, Sample: RMSE .10 Adj (True) S.D. .52 Separation 5.17 Strata 7.23 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-squared: 492.9 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-squared: 12.6 d.f.: 12 significance (probability): .40
Inter-Rater agreement opportunities: 14606 Exact agreements: 8449 = 57.8% Expected: 7532.1 = 51.6%

Table 7.4.1 Criteria Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Criteria
5159	1678	3.07	3.11	.51	.04	.96	-1.4	.95	-1.6	1.07	.62	.59	2	Fluency
2430	999	2.43	2.49	.47	.06	1.01	.1	.99	-.1	.99	.52	.52	4	Range
3354	996	3.37	3.47	-.20	.05	1.08	1.7	1.06	1.0	.92	.53	.57	5	Accuracy
2694	998	2.70	2.79	-.28	.07	1.20	3.3	1.43	4.4	.81	.34	.47	1	Task completion
5699	1679	3.39	3.44	-.51	.04	.91	-2.6	.92	-2.3	1.11	.59	.53	3	Pronunciation
3867.2	1270.0	2.99	3.06	.00	.05	1.03	.2	1.07	.3		.52			Mean (Count: 5)
1321.3	333.5	.38	.38	.42	.01	.10	2.1	.19	2.4		.10			S.D. (Population)
1477.3	372.9	.42	.42	.47	.01	.11	2.4	.21	2.7		.11			S.D. (Sample)

Model, Populn: RMSE .05 Adj (True) S.D. .41 Separation 7.67 Strata 10.56 Reliability .98
Model, Sample: RMSE .05 Adj (True) S.D. .46 Separation 8.59 Strata 11.79 Reliability .99
Model, Fixed (all same) chi-squared: 398.6 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-squared: 4.0 d.f.: 3 significance (probability): .27

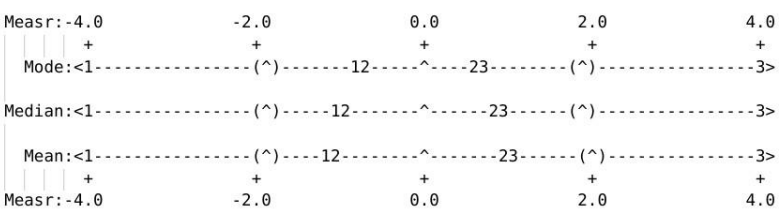
Table 8.1 Category Statistics.

Model = ?,?,?,1,R4 ; Criteria: Task completion

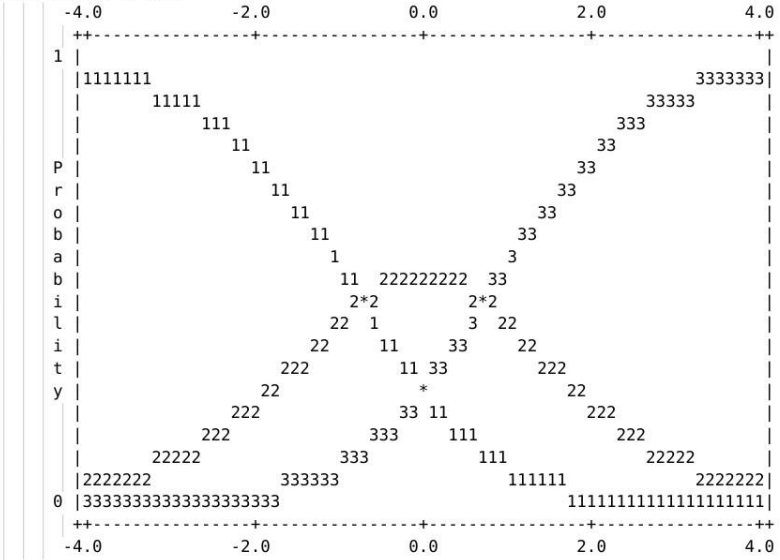
DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	40	40	4%	4%	.98	.60	1.7			(-1.93)		low	low	100%
2	220	220	22%	26%	1.64	1.42	1.4	-.72	.17	.00	-1.10	-.72	-.90	51%
3	738	738	74%	100%	2.35	2.44	1.1	.72	.08	(1.95)	1.10	.72	.89	100%

(Mean) (Modal) (Median)

Scale structure



Probability Curves



Expected Score Ogive (Model ICC)

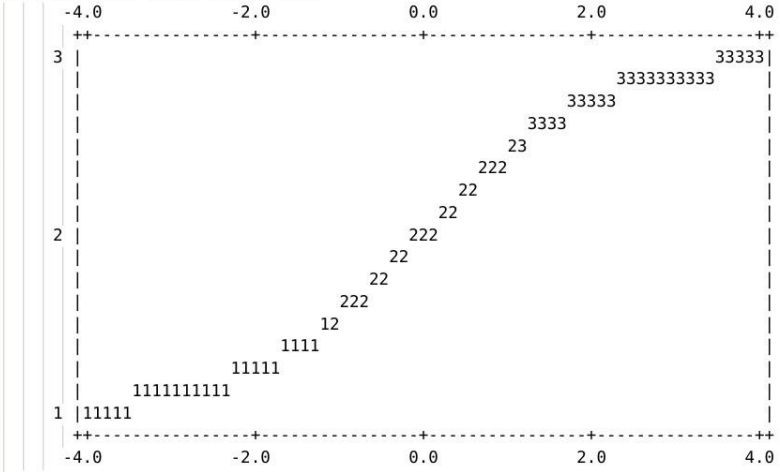


Table 8.2 Category Statistics.

Model = ?,?,?,2,R4 ; Criteria: Fluency

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	35	35	2%	2%	-.46	-.27	.9			(-3.38)		low	low	100%
2	366	366	22%	24%	.50	.48	1.0	-2.25	.18	-1.04	-2.43	-2.25	-2.33	63%
3	716	716	43%	67%	1.29	1.35	.9	.24	.07	1.16	.11	.24	.16	55%
4	561	561	33%	100%	2.25	2.19	.9	2.02	.06	(3.20)	2.33	2.02	2.15	100%
									(Mean)		(Modal)	(Median)		

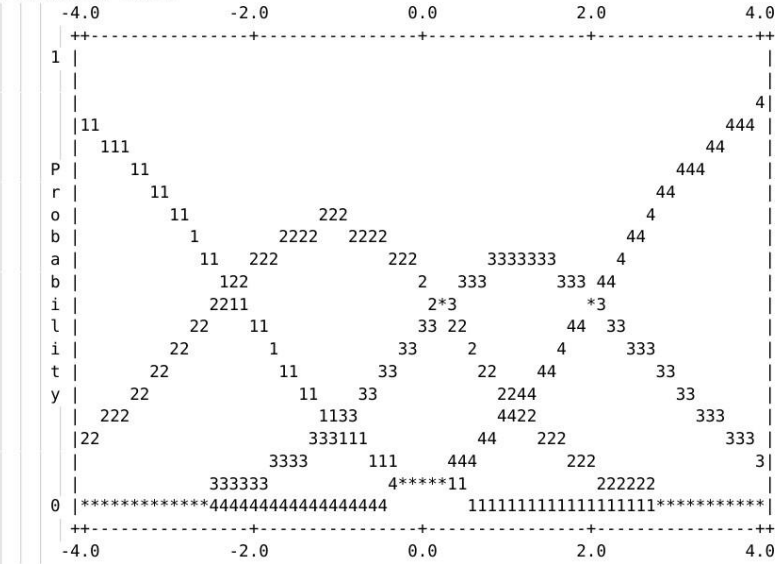
Scale structure

```
Measr:-4.0             -2.0             0.0              2.0              4.0
      +                 +                +                +                +
  Mode:<1---(^)-------12---------^---------23-------^-----34--------(^)-----4>

Median:<1---(^)------12----------^--------23--------^------34-------(^)-----4>

  Mean:<1---(^)-----12-----------^--------23--------^--------34-----(^)-----4>
      +                 +                +                +                +
Measr:-4.0             -2.0             0.0              2.0              4.0
```

Probability Curves



Expected Score Ogive (Model ICC)

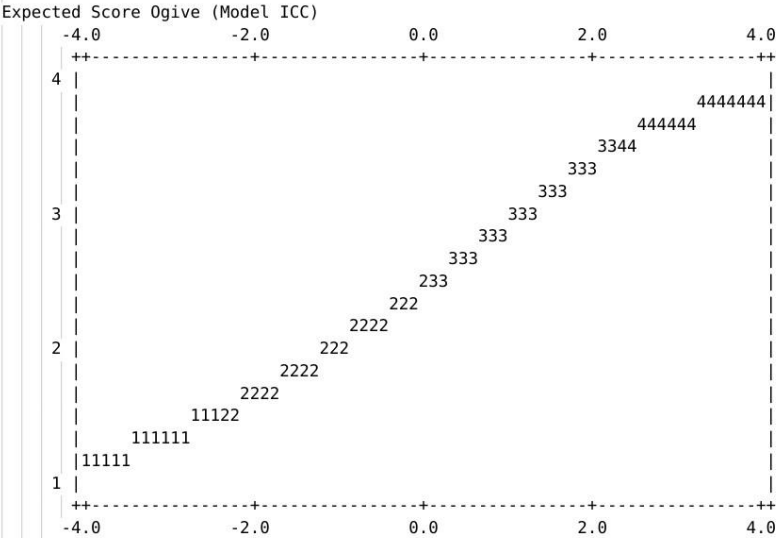


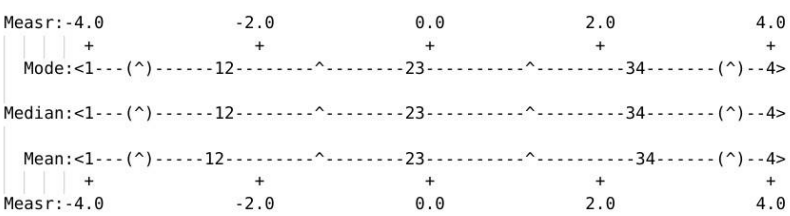
Table 8.3 Category Statistics.

Model = ?,?,?,3,R4 ; Criteria: Pronunciation

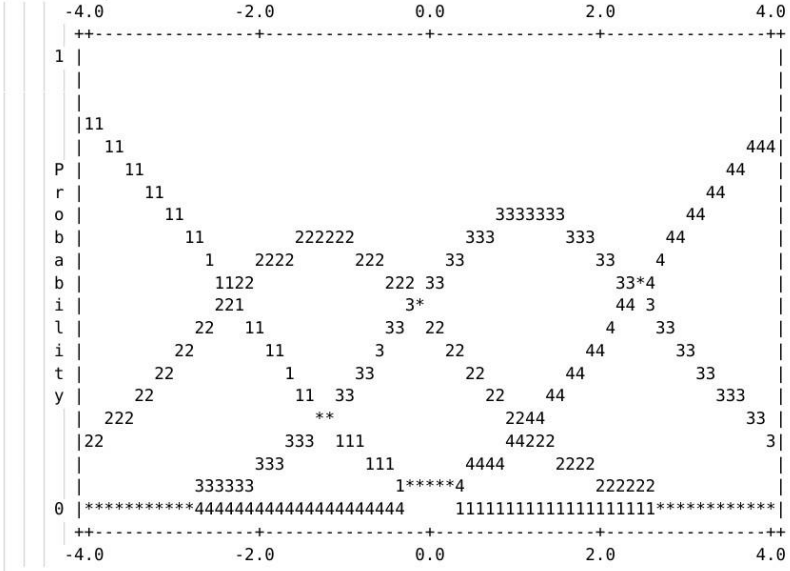
DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	6	6	0%	0%	1.28	.45	1.5			(-3.45)		low	low	100%
2	131	131	8%	8%	1.05*	1.15	.9	-2.30	.42	-1.24	-2.53	-2.30	-2.39	59%
3	737	737	44%	52%	1.97	2.05	.9	-.14	.10	1.17	-.07	-.14	-.11	64%
4	805	805	48%	100%	3.08	3.00	.9	2.44	.06	(3.57)	2.61	2.44	2.50	100%

(Mean) (Modal) (Median)

Scale structure



Probability Curves



Expected Score Ogive (Model ICC)

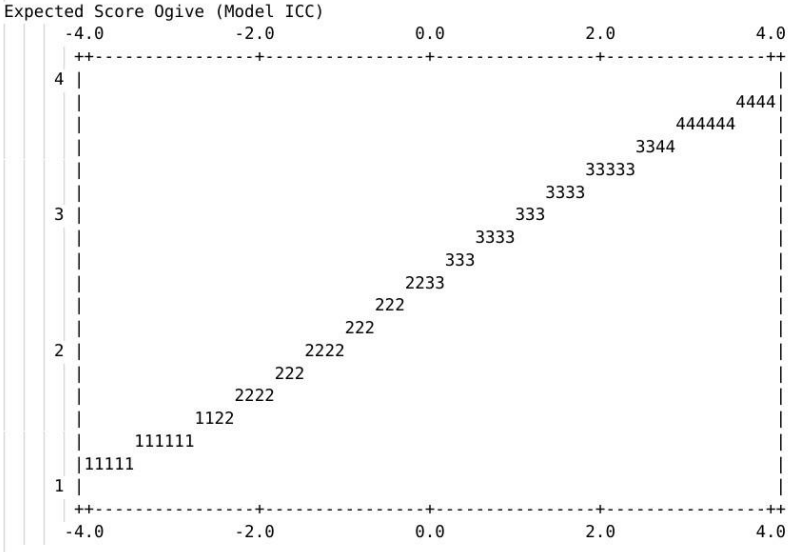
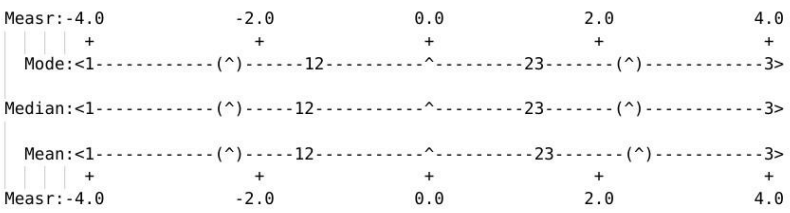


Table 8.4 Category Statistics.

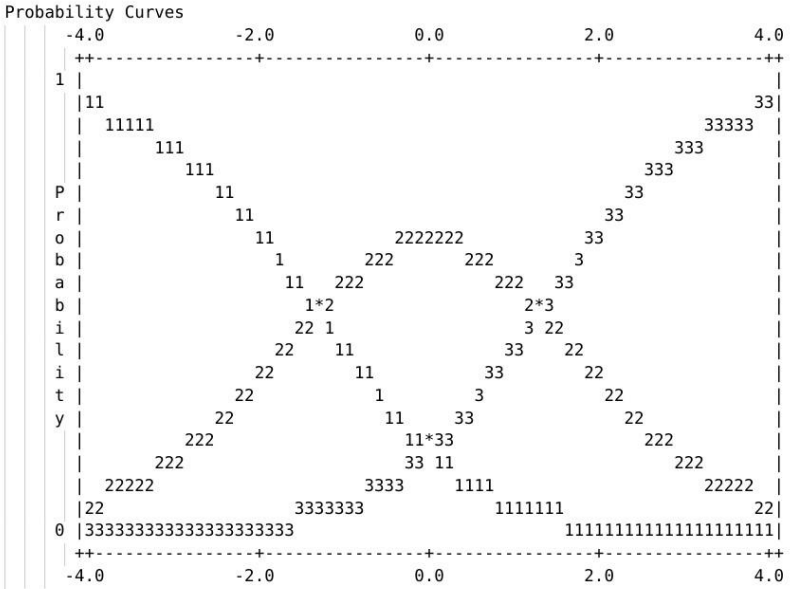
Model = ?,?,?,4,R4 ; Criteria: Range

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	70	70	7%	7%	.00	.09	.9			(-2.41)		low	low	100%
2	427	427	43%	50%	1.01	.97	1.0	-1.29	.13	.00	-1.45	-1.29	-1.36	65%
3	502	502	50%	100%	1.91	1.93	1.0	1.29	.07	(2.42)	1.46	1.29	1.35	100%
(Mean)										(Modal)		(Median)		

Scale structure



Probability Curves



Expected Score Ogive (Model ICC)

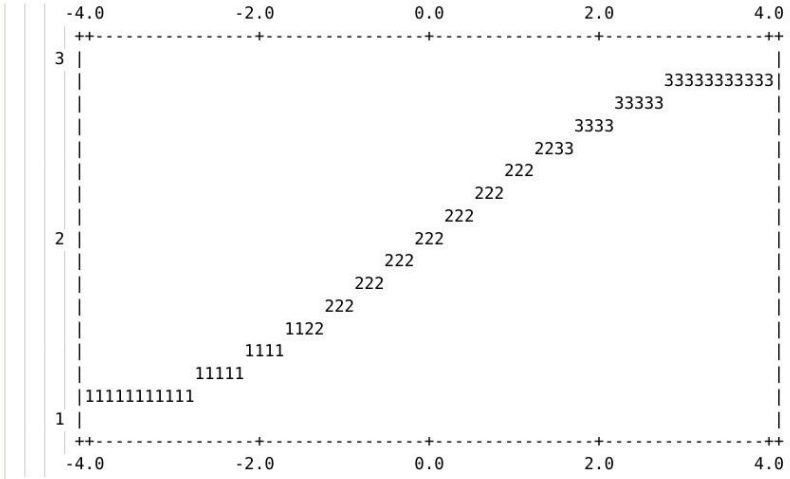


Table 8.5 Category Statistics.

Model = ?,?,?,5,R4 ; Criteria: Accuracy

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts		Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	9	9	1%	1%	.46	.28	1.2			(-3.22)		low	low	100%
2	130	130	13%	14%	.99	.90	1.1	-2.09	.34	-.94	-2.28	-2.09	-2.17	62%
3	343	343	34%	48%	1.72	1.72	.9	.32	.10	1.09	.14	.32	.22	50%
4	514	514	52%	100%	2.59	2.61	1.1	1.77	.07	(3.00)	2.16	1.77	1.94	100%
									(Mean)		(Modal)	(Median)		

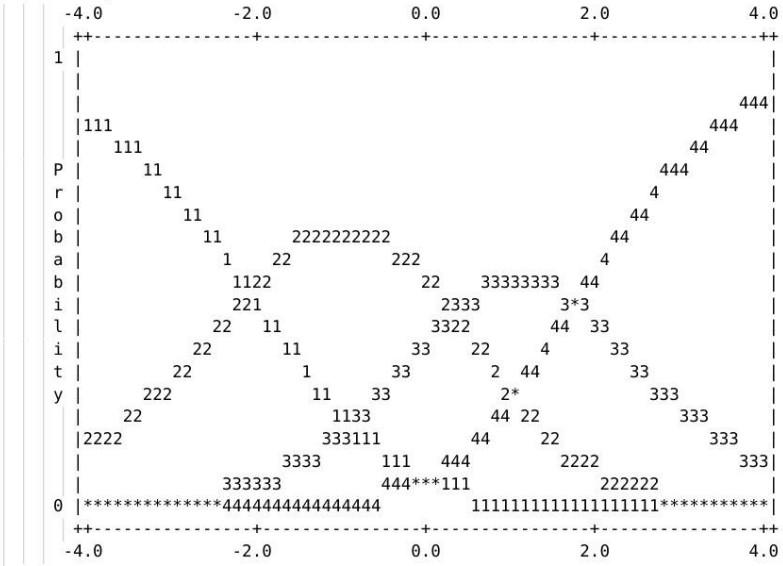
Scale structure

```
Measr:-4.0             -2.0             0.0              2.0              4.0
       +                +                +                +                +
  Mode:<1-----(^)------12---------^---------23-----^----34--------(^)-------4>

Median:<1-----(^)------12---------^--------23------^------34------(^)-------4>

  Mean:<1-----(^)-----12----------^-------23-------^-------34-----(^)-------4>
       +                +                +                +                +
Measr:-4.0             -2.0             0.0              2.0              4.0
```

Probability Curves



Expected Score Ogive (Model ICC)

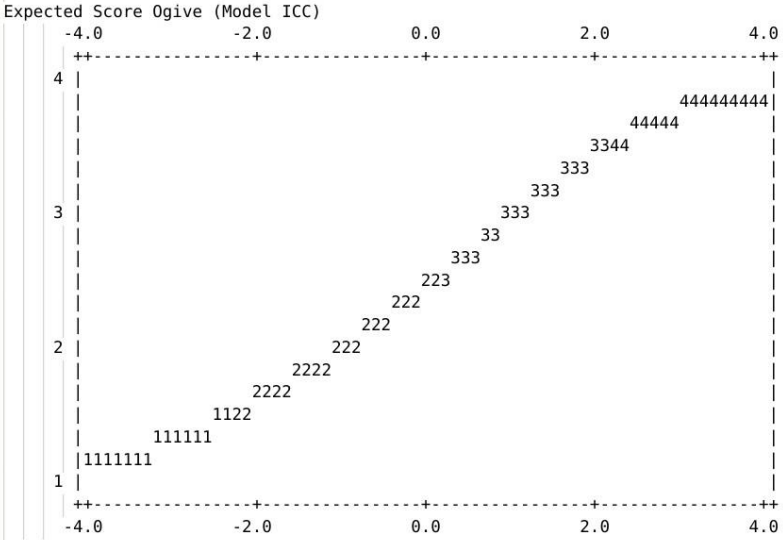


Table 4.1 Unexpected Responses (61 residuals sorted by u).

Cat	Score	Exp.	Resd	StRes	Nu	Ca	Nu	Ta	Nu	Ra	N	Criteria	Sequence
2	2	3.0	-1.0	-7.9	36	36	25	8c	11	11	1	Task completion	6057
2	2	3.0	-1.0	-6.4	43	43	2	2	7	7	1	Task completion	1180
2	2	3.0	-1.0	-6.2	68	68	2	2	7	7	1	Task completion	1195
1	1	2.9	-1.9	-5.7	6	6	2	2	4	4	1	Task completion	745
1	1	2.9	-1.9	-5.6	37	37	26	8d	6	6	1	Task completion	6177
2	2	3.0	-1.0	-5.3	69	69	25	8c	6	6	1	Task completion	5932
1	1	2.9	-1.9	-5.1	39	39	11	3b	11	11	1	Task completion	4056
2	2	3.0	-1.0	-5.0	49	49	26	8d	6	6	1	Task completion	6182
1	1	2.8	-1.8	-4.9	24	24	26	8d	12	12	1	Task completion	6277
1	1	2.8	-1.8	-4.8	18	18	2	2	3	3	1	Task completion	700
2	2	3.0	-1.0	-4.6	12	12	23	8a	5	5	1	Task completion	5458
3	3	3.9	-.9	-4.2	36	36	9	1f	11	11	3	Pronunciation	3689
1	1	2.8	-1.8	-4.2	59	59	10	3a	7	7	1	Task completion	3796
2	2	2.9	-.9	-4.1	16	16	2	2	6	6	1	Task completion	995
2	2	3.8	-1.8	-4.1	43	43	4	6b	12	12	5	Accuracy	2779
1	1	2.8	-1.8	-4.0	3	3	10	3a	6	6	1	Task completion	3761
2	2	3.8	-1.8	-4.0	68	68	4	6b	12	12	5	Accuracy	2784
1	1	3.3	-2.3	-3.9	10	10	2	2	7	7	3	Pronunciation	1097
2	2	2.9	-.9	-3.9	19	19	25	8c	11	11	1	Task completion	6042
1	1	3.3	-2.3	-3.9	29	29	7	1c	13	13	3	Pronunciation	3384
1	1	2.8	-1.8	-3.9	34	34	11	3b	5	5	1	Task completion	3949
1	1	2.8	-1.8	-3.8	17	17	2	2	3	3	1	Task completion	695
2	2	3.8	-1.8	-3.8	68	68	23	8a	12	12	5	Accuracy	5642
2	2	3.7	-1.7	-3.7	5	5	15	3f	7	7	3	Pronunciation	4374
2	2	3.7	-1.7	-3.7	12	12	2	2	3	3	3	Pronunciation	677
1	1	2.7	-1.7	-3.7	14	14	12	3c	1	1	1	Task completion	4107
1	1	2.7	-1.7	-3.7	25	25	2	2	3	3	1	Task completion	735
2	2	3.7	-1.7	-3.7	52	52	24	8b	7	7	3	Pronunciation	5760
2	2	2.9	-.9	-3.7	54	54	2	2	7	7	1	Task completion	1190
2	2	2.9	-.9	-3.6	18	18	2	2	6	6	1	Task completion	1005
2	2	3.7	-1.7	-3.5	1	1	2	2	1	1	5	Accuracy	474
2	2	3.7	-1.7	-3.5	23	23	4	6b	7	7	3	Pronunciation	2527
1	1	3.1	-2.1	-3.5	39	39	11	3b	13	13	3	Pronunciation	4074
1	1	2.7	-1.7	-3.5	53	53	16	4a	14	14	4	Range	4527
2	2	3.7	-1.7	-3.4	1	1	6	1b	7	7	3	Pronunciation	3108
2	2	2.9	-.9	-3.4	6	6	2	2	7	7	1	Task completion	1085
1	1	2.7	-1.7	-3.4	10	10	2	2	7	7	1	Task completion	1095
1	1	3.3	-2.3	-3.4	10	10	2	2	7	7	5	Accuracy	1099
2	2	2.9	-.9	-3.4	12	12	2	2	3	3	1	Task completion	675
2	2	2.9	-.9	-3.4	16	16	19	4d	5	5	1	Task completion	4904
2	2	2.9	-.9	-3.4	31	31	25	8c	10	10	1	Task completion	6037
1	1	2.7	-1.7	-3.4	32	32	3	5	8	8	1	Task completion	2131
2	2	3.7	-1.7	-3.4	41	41	19	4d	6	6	5	Accuracy	4943
1	1	3.0	-2.0	-3.3	29	29	1	1d	13	13	3	Pronunciation	433
1	1	2.6	-1.6	-3.3	29	29	17	4b	7	7	4	Range	4584
1	1	2.7	-1.7	-3.3	64	64	23	8a	6	6	1	Task completion	5493
2	2	2.9	-.9	-3.2	24	24	24	8b	7	7	1	Task completion	5748
2	2	2.9	-.9	-3.2	27	27	2	2	7	7	1	Task completion	1170
2	2	3.7	-1.7	-3.2	32	32	10	3a	6	6	5	Accuracy	3775
2	2	3.7	-1.7	-3.2	55	55	2	2	10	10	5	Accuracy	1554
1	1	2.7	-1.7	-3.2	64	64	25	8c	5	5	1	Task completion	5927
2	2	3.7	-1.7	-3.2	66	66	23	8a	6	6	5	Accuracy	5502
2	2	3.6	-1.6	-3.1	8	8	7	1c	8	8	3	Pronunciation	3290
1	1	3.2	-2.2	-3.1	15	15	10	3a	7	7	5	Accuracy	3785
1	1	2.6	-1.6	-3.1	16	16	2	2	9	9	4	Range	1358
2	2	2.9	-.9	-3.1	18	18	14	3e	12	12	1	Task completion	4298
3	3	3.9	-.9	-3.1	21	21	7	1c	5	5	3	Pronunciation	3250
1	1	2.6	-1.6	-3.0	7	7	4	6b	4	4	1	Task completion	2380
3	3	3.9	-.9	-3.0	36	36	9	1f	11	11	2	Fluency	3688
2	2	2.9	-.9	-3.0	41	41	2	2	6	6	1	Task completion	1060
2	2	2.9	-.9	-3.0	53	53	18	4c	12	12	1	Task completion	4862
Cat	Score	Exp.	Resd	StRes	Nu	Ca	Nu	Ta	Nu	Ra	N	Criteria	Sequence