INVESTIGATING RATER PERCEPTIONS IN THE ASSESSMENT OF SPEAKING

PRZEMYSŁAW KRAKOWIAN
University of Lodz, Poland
przemyslaw.krakowian@uni.lodz.pl

Abstract
In the assessment of spoken production, numerous reasons can be identified behind the decisions that raters make in evaluating samples of oral performance. Inter and intra rater factors are relatively well documented in various reliability and validity studies. Some that have been identified in literature involve the effects of examinee pairing or the familiarity with the examinees, others point in the direction of gender and gender role perceptions O'Sullivan (2008), others appear to be connected with body language and non-verbal cues that accompany oral production (cf.: Krahmer and Swerts 2004, Seiter, Weger, Jensen and Kinzer 2010). While some studies that address the assessment of speaking English in exam contexts suggest that raters may not feel as comfortable assessing pronunciation as they do other aspects of a speaker's performance (Orr 2002, Hubbard, Gilbert and Pidcock 2006, Brown 2006, De Velle 2008), more recent investigations of rater behaviour involving electronic evidence from training, maintenance and online examination programmes tentatively show that pronunciation, in fact, is the first category examiners attend to (Hubbard 2011, Chambers and Ingham 2011, Krakowian 2011, Seed 2012, Tynan 2015, Kang and Ginther 2019). This paper looks at large collection of assessments stored in an electronic system to investigate what raters really seem to pay attention to when allegedly following rating scales.

Keywords: Assessment, speaking, assessment of oral production, reliability, bias

1. Introduction

This paper looks at the use of a large batch of pre-tested and in some cases standardised samples of oral expression with different assessment schemas and raters, whose experience, exposure, and involvement in assessing spoken production varied considerably. The data for the claims in this paper comes from samples recorded during mock exams for students of the English Programme at the Philological Faculty at Lodz University. The mock exams were held as part of orientation and familiarisation procedure preparing students for the Speaking Paper of the Integrated Skills Exam. For the students, this offered an opportunity to familiarize themselves with exam tasks in a non-threatening environment, while from the organisational perspective, it presented opportunities to obtain recorded samples of spoken expression. The samples were then stored online for training and induction procedures as part of the examiner standardisation schema, where any examiner introduced, or re-introduced into the exam had to participate in an online training before rating students in live exams. The core component of this process involves the use of an Electronic Performance Support System (EPSS) where the samples are stored along with ratings, comments and in training can be browsed and compared, with a forum for exchange of opinions and discussion threads, constituting a meeting ground for assessors.

Most evaluation schemas involve provisions for handling assessment of pronunciation ranging from intelligibility and accurate production of individual sounds, through managing word and sentence stress and appropriate intonation, to such use of phonological features that they convey and enhance meaning. It is interesting, however; to look at what happens when examiners need to make ratings of oral expression in the absence of explicit scales to handle assessment of pronunciation. It is also interesting to see what happens when scales used to evaluate spoken performance focus on discrete elements of communicative performance and account for various aspects such as the quality of language expressed in range as well as accuracy, where range applies as separate categories to both vocabulary richness as well as variety of linguistic expression. Separate categories handle how discourse is developed and managed in interaction as well as in task completion and development. All of this leads to a conclusion expressed by Weir and Milanovic (2003) that with complicated scales and long performance descriptors, even when training is a mediating factor, assessors may turn to their beliefs and experiences and pay little or no attention to guidelines. Several such trends have been observed in the rater performance data collected in the EPSS. Fortunately, research methodology exists that allows to correct bias and improve reliability of the ratings.

2. Background information

Following Scollon and Scollon (1995), Nakane (2007), Lustig and Koester (1993) Hall and Hall (1990) and Barna (1994), an educational communication ecology may be defined as an arrangement characterised by the fact that the interaction of individual participants of different experience, background knowledge and familiarity with the subject at hand, sharing, for the purpose of work, pursuit of knowledge, collaboration or leisure activities, one or more languages for their communication, creates a situation where worldview, self-identification, behavioural paradigms, value orientations, ideosphere and memetic diversity are augmented to a larger or lesser extent and result in the emergence of fairly unique and individual cognitive and communication strategies, verbal and non-verbal means of information encoding, as well as the perception of reality in each of the cultures involved. In some respects, this leads to a greater efficiency of collaboration, but at the same time may lead to misunderstandings and ineffective communication (Bennet 1993, Lustig and Koester 1993, Hill, Wilson, and Watson 2004, NRC 2015).

In terms of the impact of diversity various participants of educational communication ecologies bring into the setting on the reliability of the process of language skills evaluation, especially in the area of speaking and writing assessment, it may result in a rather lengthy process of accommodating the beliefs and perceptions concerning language performance as well as the understanding of the descriptors of such performance (Fulcher 2003, Hawkey 2004, Taylor and Falvey 2007). It has even been claimed (O'Sullivan 2008) that despite the willingness to adjust and modify the understanding of rating scales in a language other than one's native and opportunities to undergo training, the characteristics of one's culture may consistently, though inconspicuously affect behaviour, including one's ability to adhere to marking criteria. Unfortunately, studies of educational ecologies in contexts of assessment are rare, and the ones performed very often concentrate on ascertaining a certain minimum reliability, rather than on the mechanics of the interaction (Weir and Milanovic 2003, Hawkey 2004, Taylor and Falvey 2007).

A Community of Practice (CoP), a term often associated with learning ecologies, is often defined as a network or a forum, both informal and with varying degrees of formal structuring and internal organisation, through which ideas are exchanged and solutions generated (Wenger 1998). It implies the existence of a group of professionals, associated with one another through similarity of interests and expertise, and working on a common set of problems, in common pursuit of solutions, and themselves constituting a store of knowledge (Wenger 1998, Wenger, McDermott and Snyder 2002). The community of practice additionally entails the process of social learning that takes place when individuals who have common interests in some field or problem collaborate and share ideas, come up with solutions and otherwise interact with each other to work (Saint-Onge and Wallace 2003, Hildreth and Kimble 2004).

It constitutes a very attractive arrangement for many undertakings as nothing binds people together faster than common interest and pursuit of solutions to common problems (Wenger 1998, Wenger, McDermott and Snyder 2002, Saint-Onge and Wallace 2003, Hildreth and Kimble 2004). One such scheme, in which communities of practice through the use of an online environment collaboratively evaluating spoken performance, allowed to register and discover that owing to a diversity of backgrounds, educational experience and a number of other persisting factors, it may be difficult under certain circumstances to reach a consensus and arrive at a satisfactory convergence in evaluating samples of oral production.

3. **Research objectives**

When inspected, the samples stored in the EPSS show that, apart from a set of benchmarked training materials where the samples were hand-picked from a larger set as typical for the various levels, representative and illustrative, and analysed using a variety of means to ensure reliable calibration, and where convergence in rating was very high, an inordinate number of assessment instances used for the training purposes differed considerably. Upon closer inspection, this divergence may be attributed to a number of factors: i) discrepancies in marking by different assessors; ii) individual differences in the perception of non-native speech related to phonetic accuracy, speed of talking, and the ensuing perceived fluency, range and accuracy of student vs. native speaker performance; iii) inter and intra rater variability affected by external factors as well as by the above.

Intra rater variability is something that may be partly attributed to fatigue and boredom related to the tediousness of the task, but it also relates to the learning curve and the perception of the rating scales overtime, as it can be concluded that with greater understanding of the mechanics of scoring, the accuracy and reliability of the scorer increases until it reaches a plateau, and the variability there is beyond an individual’s control and impervious to training effects.

Inter rater variability may be attributed to a number of factors, the most prominent of which lies in the fact that generally the more complicated the rating scale, the larger the opportunity to make errors of judgment. In relation to the various scales in the EPSS that reflect the changes in measures implemented in the Integrated Skills Exam over the years, the inescapable conclusions seem to be that: i) the overall scales for the tasks in the EPSS are complicated and sometimes excessively long inn detailed descriptions; ii) they are supplemented by additional scales pertaining to different aspects or facets of performance; iii) the additional scales are complicated. This is perhaps a common phenomenon in assessment, and manageable and modifiable by training, but it may be concluded that there is always some interference were rates will read in different ideas into the descriptors as assessment as an instance of communicating ideas in the form of evaluative statements *is a symbolic, interpretative, transactional, contextual process in which the degree of difference between people is large and important enough to create dissimilar interpretations and expectations about what are regarded as competent behaviours that should be used to create shared meanings* (Lustig and Koester 1993:51).

Sources of misunderstandings are numerous and have been pointed out by at least two separate theories. The psychological anthropology theory of intercultural interference (Barna 1994, Scollon and Scollon 1995) claims that interference emerges when: i) discourse participants, in our case evaluators who interact with the sample, assume that all humans are essentially similar and therefore behave and interpret behaviour in a similar way; ii) discourse participants incorrectly interpret non-verbal clues and non-verbal communication; iii) discourse participants read into their interpretation of discourse their pre-conceptions, stereotypes, and superstitions; iv) discourse participants assume a stance in which they judge and evaluate elements of the discourse according to their set of values and culture related norms of behaviour; v) additionally there is an element which is inherently connected with the language component, namely with the differences between any two or more languages and vi) ensuing tension, anxiety and other affective factors that appear when a language other than a native one is used by a speaker, a phenomenon also known as culture shock.

Alternatively, Hymes (1964, 1972), outlining his model of communication and the concept of communicative competence, and looking at where and how the act of communication is performed, what the purpose and aim of it is, how conventionally the communication is performed, claimed that the violation or misinterpretation of the culturally sanctioned norms of language behaviour leads to intercultural interference (cf. Krakowian 2011).

Some additional factors have been identified and postulated as the underlying causes for divergence following research outlined in O’Sullivan (2008). The first to be taken into consideration was gender in the perception of speech, where effeminate male language tends to be decidedly underscored by male raters and slightly, but statistically significantly overscored by female raters, but only of some nationalities. The next issue taken up was gender of the examiner and the alleged claim that female raters rate more leniently i.e.: overscore the subjects in general and those whom they are familiar with in particular. Notwithstanding, the familiarity with the subjects/examinees under evaluation with a tentative tendency to evaluate more favourably those who are known to us as opposed to those we are not familiar with. Some examining bodies (e.g.: ESOL formerly UCLES) require that their oral examiners familiarise themselves with the names of the examinees before the exam and identify those whom they have taught in the past several years.

Also, the difficulty of the task and the resource intensity of the task: the effect the task has on the complexity of the discourse and range of linguistic resources that have been implemented in an effective and efficient achievement of the task. A task which is not demanding enough may leave the evaluator with the impression that the performance was of a lesser value and that the level of competence is lower when actually the examinee had no opportunity to show his full potential. Likewise, the effect of the examinee pairing, obviously only in situations involving collaborative tasks, where male vs. male and female vs. female pairing received more favourable ratings than a setting involving female vs. male arrangement, where an additional claim was made that female speakers producing a comparable amount of discourse would tend to be perceived as overbearing and dominating the interaction (Seiter, Weger, Jensen and Kinzer 2010).

Apart form the above, the effect of the native background culture of the assessor might prove important when evaluating students speaking a language of a different culture and mediating the assessment in a language that is other than their native in the context of what is understood as in Hall (1959, 1966) and Hall and Hall (1990) as high and low context culture interference. Some of the characteristics of the high context cultures mean that a culture in which the individual has internalised meaning and information, little is explicitly stated in written or spoken messages. In conversation, the listener knows what is meant; because the speaker and listener share the same knowledge and assumptions, the listener can piece together the speaker's intentions. In a high context culture, the individual must know what is meant at the covert or unexpressed level and is supposed to know how to react appropriately. Discourse participants are expected to understand without explanation or specific details to the point that explanations may be considered insulting, as if the speaker regarded the listener as not informed or suave enough to understand.

High context cultures, therefore, rely on indirect communication and use fewer words, tend to read between the lines and are highly tolerant of silences (Nakane 2007). A low context culture, on the other hand, is one in which information and meanings are overtly stated and where the individuals expect explanations when statements or situations are ambiguous. Information, context and meanings are not internalised by the individual but instead derived from the actual discourse. Hall (1959, 1966,) and Hall and Hall (1990) claim in their work that most of the information missing in the internal and external context must be included in the transmitted message or communication breakdown will ensue.

4. EPSS data and discussion

The following data and the samples were accumulated over the period of five academic years in nine separate events involving recorded mock exams for the students of the English Programme at the Philological Faculty at Lodz University. Samples of spoken productions are rare, and unless they are mandated by the institutional procedures, like it was the case during the COVID-19 pandemic, where the Speaking Paper and the ensuing discussion of the assessment were recorded in MS Teams to satisfy legal requirements for online exams regarding proctoring and record keeping (Journal of Law 2007), students need to agree to being recorded. The mock exams were held as part of orientation and familiarisation procedure preparing students for the Speaking Paper of the Integrated Skills Exam, with the requirement that the participants needed to agree to being recorded, but in the case of a successful outcome of the assessment that was performed post-hoc, the incentive was that such a recording would be used as an exemption from the actual exam. For the students, this offered an opportunity to familiarize themselves with exam tasks in a non-threatening environment, while from the organisational perspective, it presented opportunities to obtain recorded samples of spoken expression. Additionally, when students gradually got accustomed to being recorded, and opted for this type of exam in larger and larger numbers, this had positive impact on the actual Speaking Paper rosters, which was a logistical relief, especially in the winter exams, where the break between semesters allowed very little time between the exams and make-up exams. In total 219 samples were collected of which 180 were considered to be sufficiently and proportionately typical for the various levels, representative and illustrative of the scales used.

The analysis of the ratings was performed using two separate procedures and in the first it used the Chi-square statistic, where the model against which the residual is calculated was established as an average for all the ratings recorded for the samples, including the raters showing bias as sig.1., and excluding the raters showing bias as sig.2. The first batch of observations concerns the trend identified as very significant in the discussion of rater behaviour and bias discussed in O'Sullivan (2008).

The claim that effeminate male language tends to be decidedly underscored by male raters was confirmed on a population of 20 samples involving the assessments of 21 raters, 9 of whom were male raters (sig.1=0.90103, sig.2=0.02397). Effeminacy is, by no means a scientific category, nonetheless, for the purpose of this study the procedure involved judgements of a group of the raters not involved in the ratings, but who upon inspecting the samples classified speech samples as falling into this category based on tone of voice, gestures and overall impression of the sample as departing what was to be considered typically male behaviour on such occasions.

The tendency to evaluate more favourably those whom the raters are familiar with was confirmed at the statistically significant level for 45 raters (sig.1=0.33042, sig.2=0.04351). The familiarity with the subjects was determined on the premise of teaching history and class participation of the parties involved in a post-hoc metadata entered into the database.

Additionally, the effect of the examinee pairing was investigated in samples where both examinees were male or both female and where there were mixed genders. The effect of examinee pairing was visible for 37 raters; 20 female and 17 male in 97 samples where homogeneous pairing received statistically higher rating than a mixed gender pairing irrespective of gender of the examinees (sig.1=0.64331, sig.2=0.08515, sig.1=0.51825, sig.2=0.07813).

Some samples recorded in the EPSS were either shorter in duration or contained more pauses and as such were selected to test the claim concerning tolerance for silence and ambiguity (Nakane 2007) being greater for those whose first language was the same as the language of the evaluated discourse, i.e.: native speakers of English, though of divergent nationality – 9 in total of British, American and Canadian backgrounds. This group compared to 10 other raters involved in the rating of the same samples consistently and statistically significantly over-rated samples of considerably smaller discourse size, shorter or/and containing more pauses and hesitations (sig.1=0.40724, sig.2=0.08515) and did that irrespective of gender (sig.1=0.44221, sig.2=0.08235 for males, and sig.1=0.40739, sig.2=0.08728).

Finally, a local twist in the data, which seems to be pointing in the direction of the 37 Polish raters scoring 52 samples with greater perceived phonetic accuracy more favourably than the rest of the raters (sig.1=0.82174, sig.2=0.03671). The identification of this bias was possible owing to the procedure itself, but in order to establish the parameters of the finding, the samples involved in the ratings of the identified sub-group were additionally submitted to a rating procedure from independent raters. The ratings, and their rank order in particular, were used as a baseline to confirm what earlier on was referred to as greater perceived phonetic accuracy.

The Chi-square statistic used in the first part of the study and presented above works on the premise that behaviour departing from the postulated model, whatever model that may be, is penalised by the statistic in the form of a residual. The residual, in turn, is squared to remove the negative sign and accumulated in order to be finally inspected for significance. The second part of the procedure involved Multi Facet Rasch Analysis, where for the logit-based statistic, the residual is essentially based on the same principle as in the Chi-square statistic, though the procedure is infinitely more complex and involves the application of the exponential function (for discussion on Rasch Analysis consult Wright and Stone 1979, Wright and Masters 1982, Wilson 2005, Bond and Fox 2007, Krakowian 2010). The research assumption beyond it was that there would be some overlap permitting to confirm already identified trends and to identify additional processes or additional samples that conformed to the patterns identified earlier.

The Multi Facet Rasch Analysis procedure, was performed using FACETS (Wright and Stone 1979, Wright and Masters 1982), a Many-Facet Rasch Analysis dichotomous and polytomous model program, which indeed did confirm the existence of the trends identified earlier. It additionally revealed more tendencies, which could not have been predicted in the Chi-square study.

Rasch Analysis identified a strong tendency for female raters in the data set overall to mark more leniently and more cautiously, which is reflected in poorer distribution of scores and smaller variance in their marking, a phenomenon mentioned earlier, namely, the investigation of marker statistics against examinee performance revealed elevated *t-fit* statistics pointing towards a tendency to overrate and mark more leniently and favourably those whom the raters are familiar; the analysis identified 2 subsequent raters, and 5 further samples, where no information was recorded regarding familiarity through teaching history, but when investigated further the familiarity was established through participation in other exams.

Rasch Analysis identified a group of raters who relatively consistently and in a statistically significant way rated more favourably a considerable group of samples determined later to be of smaller discourse size, shorter or/and characterised by a larger number of pauses and hesitations – the identified group constituted nearly a perfect match with the group identified in the Chi-square analysis,– the difference was in the number of raters, where of the original 9 raters only 4 were deemed as consistently biased and 5 had elevated t-fit indices.

As for greater phonetic accuracy, a term, or notion that may be disputed as a non-tangible, intuitive, but perceptually viable and noticeable, despite missing from the CEF scales, or determined by descriptors relating to the pronunciation of individual sounds, but not so much pitch, intonation or tempo, or just, as was the case in some scales used, as native, non-native, or foreign, Rasch Analysis allowed to identify 37 Polish raters in the Chi-square study, and additional 4 raters, who using the global scales, but not the performance specific scales such as range, accuracy, fluency and cohesion, marked samples more favourably than the rest of the raters.

Attempts were made to recognise moderate, lenient and severe markers, and their effectiveness – as well as bias they bring into the evaluation of oral performance – the attempts resulted in determining a general framework for recognising examiner severity/leniency – a term which is an umbrella expression conveniently embracing both the proclivity of a rater to award lower as well as higher ratings than those that could be regarded as objective, where objective for the sake of comparison have to be assumed normal or common for the examination setting with regard to the samples in question.

This systematic bias towards harshness/strictness or tolerance/leniency in rating can be attributed to a variety of factors (such as the examining setting, circumstances, expectations, attitudes and beliefs, examiner characteristics, as well as exam and examiner standards and preconceptions, not to mention ephemeral factors which need to be regarded as random and thus beyond systematic control) – the severity measure for examiners in the multi-facet Rasch Analysis is established using a summary of the ratings the rater allocated throughout the process of evaluating samples (O’Sullivan 2008, Martinez 2009, 2010) – rater uniformity and stability is measured by a mean-square fit statistic (*t-fit* statistic) – this measure is established on the proportion of empirical error variance to model postulated error variance - as it is based on a normally distributed Chi-square statistic its expected value equals 1 – the value of the mean square fit statistic for any of the examiners indicates the examiner’s uniformity and consistency in rating, or in other words how well their rating behaviour fits the postulated model.

The model here being the prevailing or normal/common behaviour of the rating population, or any other if *a priori* postulated models exist e.g.: as a result of prior study and investigation - in any examining setting neither too high nor too low fit statistics are desirable, in essence several models in fact establish control lines (O’Sullivan 2008, Martinez 2009, 2010) – e.g.: when the examiner fit statistic falls below than 0.5, it is indicative of more than fifty per cent lower variance in ratings than ensues from the model – what can be deduced from this is that the examiner is for one reason or another, or through a combination of factors inclined to award the same rating to numerous candidates, and does so without regard to their real abilities – to the Multi Facet Rasch Analysis such examiner behaviour is not only easily identifiable, but it additionally carries the danger of under-distinguishing examinee characteristics – however, if the fit statistic exceeds 1.5 it becomes indicative of over fifty percent greater variance than could be deduced from the model (O'Sullivan 2008, Martinez 2009, 2010) – this in turn translates into unexpectedly high or low ratings without regard to examinee real abilities – Rasch Analysis identified several instances of both: low variance or safe middle raters numbered 17 in total and tended to be female, who totalled 12 – on the other hand high variance or careless extreme markers tended to be male, who totalled 9 out of 13 identified.

Despite the variability introduced into the ratings by careless and overly safe markers, it can, following Martinez (2010), however, be concluded that overall, neither of the processes disturbs the rank order of ratings affected by either of them – by and large, differences in rating severity (or leniency, whichever term we choose to use) have been shown for subjective performance ratings i.e.: those that require the intervention of human examiners, to reflect the assumptions of the Multi Facet model that more able candidates will still obtain better scores regardless of the severity/leniency of the examiners – despite unusually high or low variances exhibited in their ratings, the examiners manage to achieve the same ranking of the examinees as the models, both the overall performance model which includes all ratings as well as the model consisting of all the ratings, but those of the raters in question – the data sets used in this investigation confirm the above, albeit clearly indicate departure from the model, and in essence show rater disregard for examinee real abilities.

Apart from the above, a number of samples and ratings were identified with the use of the Multi Facet Rasch Analysis whose provenience could not be attributed to any of the trends postulated and could not be associated with any of the groups of raters, neither gender nor familiarity or experience-wise – a closer inspection of the samples, however, revealed that the feature they shared in the samples in question related to their body language/posture and/or gestures and head movements and/or facial expressions - and while the relationship between those factors and the evaluation by raters cannot be determined owing to a limited number of samples and raters involved, one observation can be made, namely that they provoked divergent reactions in raters – this phenomenon could be consistent with an explanation provided by Guaïtella, Santi, Lagrue and Cavé (2009), Krahmer and Swerts (2004), Cavé, Guaïtella and Santi (2002) that facial expressions, both those involving any number of facial muscles expressing feelings, reactions and attitudes of the interlocutors, as well as those muscular movements involved in speech such as eyebrow movements and muscular contractions related to lip position, are all linked to discourse production and are instrumental in reading non-verbal turn-taking indicators and discourse markers –

additionally, nonverbal behavior such as body posture and gestures is postulated by Seiter, Weger, Jensen and Kinzer (2010), following their research on political discourse, to influence audience perceptions of televised debaters' credibility, appropriateness, objectivity, rhetoric skill, and the degree to which the audiences considered their debate to be won.

Notoriously, exams, and especially oral exams and in particular those that are recorded for posterity, are stress and anxiety inducing events – consequently discourse participants assume a very characteristic, withdrawn stance and use very few gestures; body movements and non-verbal clues do not abound - on those few occasions identified by Rasch Analysis, and consistently with research by Seiter, Weger, Jensen and Kinzer (2010), some of the raters could have been reacting to such clues, though the exact relationship remains yet to be determined – an alternative, or perhaps complementary explanation can be found in Ockey (2009), who links personality to Bachman's and Palmer's (2002) notion of the role of context in communication with the test takers ability to skilfully react to contextual clues, including attitudinal, non-verbal as well as personality clues, which may be reflected in the discourse through a variety of forms, some being those discussed earlier.

5. Conclusions and suggestions for further research

As can be seen from the discussion of identified trends, there exists a multitude of reasons behind the decisions that raters make in evaluating samples of oral performance. Fortunately, research methodology exists to identify, and if need be, correct their influence on the reliability of the process. The data analysed in the course of this investigation leaves room for further research, provided more metadata information can be stored regarding factors in question. Uncertainty as to the familiarity of some of the raters with the subjects, non-verbal cues and their influence on the raters' perceptions of performance could be further explained in the light of the information, which at the moment of writing is not available to the author.

Definitely, more research is needed to determine the relationship between perceptions, ratings, and rating consistency and non-verbal components of communication. It is difficult to estimate how much influence on the raters is exerted by what is not spoken, but from the analysis it can be seen that this is a factor that is not to be ignored as it registers in the Multi Facet Rasch Analysis. This fact could have pedagogical implications for teaching speaking skills, as this aspect of communication seems to be neglected in classroom practice.

Other avenues of research that seem promising relate to the fact that speech recognition technology in the form of affordable online tools that have become increasingly available in recent years. Speech turned into text can be analytically vetted against human ratings with the aid of type-to-token ratios (TTR) and subordination to coordination indexing, perhaps helping to establish more objective models and benchmarks for analysis.

References

Bachman, L.F., & Palmer, A. (2002). *Language Testing in Practice.* Oxford University Press

Barna, M.L. (1994). Stumbling Blocks in Intercultural Communication. In Samovar L.A. & R.E. Porter, (eds.), *Intercultural Communication.* Wadsworth.

Bennet M.J. (1993). Towards Ethnorelativism: A Developmental Model of Intercultural Sensitivity. In Paige, R.M, (ed.), *Education for the Intercultural Experience,* pp. 21-71. Intercultural Press.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences.* University of Toledo Press

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test., *IELTS Research Reports, 6,* 41-65. IELTS Australia, Canberra and British Council, London.

Cavé, C., Guaïtella, I., & Santi, S. (2002). Eyebrow movements and voice variations in dialogue situations. In Hansen, J.H.L & Pellom, B., (eds.), *Proceedings of the 7th International Conference on Spoken Language Processing,* 2353–2356.

Chambers, L., & Ingham, K. (2011). The BULATS Online Speaking Test., *ESOL Research Notes, 34*, 21-25.

De Velle, S. (2008). The revised IELTS Pronunciation scale., *ESOL Research Notes, 34*, 36-38.

Fulcher, G. (2003). *Testing Second Language Speaking.* Pearson Longman

Guaïtella, I., Santi, S., Lagrue, B., & Cavé, Ch. (2009). Are Eyebrow Movements Linked to Voice Variations and Turn-taking in Dialogue? An Experimental Investigation. *Language and Speech, 57*, 207-222.

Hall, E., & Hall, M. (1990). *Understanding cultural differences: Germans, French and Americans.* Intercultural Press

Hall, E. (1959). *The silent language*. Doubleday

Hall, E. (1966). *The hidden dimension.* Doubleday Anchor Books

Hawkey, R. (2004). A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills, CELS, examinations. *Cambridge ESOL Research Notes* Volume 16

Hildreth, P.M., & Kimble, Ch. (2004). *Knowledge networks: innovation through communities of practice.* Idea Group Inc., IGI.

Hill, S.B., Wilson, S., Watson, K. (2004). Learning Ecology. A New Approach to Learning and Transforming Ecological Consciousness. In: O'Sullivan, E.V., Taylor, M.M., (eds.) *Learning Toward an Ecological Consciousness: Selected Transformative Practices*. Palgrave Macmillan, New York.

Hubbard, C., Gilbert, S., & Pidcock, J. (2006). Assessment processes in speaking tests: a pilot verbal protocol study., ESOL *Research Notes, 24,* 14-19

Hubbard, Ch. (2011). Cambridge ESOL Professional Support Network Extranet: Development and impact., *ESOL Research Notes, 49,* 17-26.

Hymes, D. (1964). Introduction: Toward Ethnographies of Communication *American Anthropologist, 66*(6), pp. 1–34

Hymes, D. (1972). On communicative competence. In J. B. Pride & Holmes, J., (eds.), *Sociolinguistics,* pp. 269–285. Penguin.

Journal of Law 2007 no. 188, Section 1374: Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego z Dnia 25 Września 2007 r. w sprawie warunków jakie muszą być Spełnione, aby zajęcia dydaktyczne na studiach mogły być prowadzone z wykorzystaniem metod i technik kształcenia na odległość, dz. u. 2007 nr. 188, poz. 1374 z Późn. Zm.

Kang, O., & Ginther, A. (2019). *Assessment in second language pronunciation*. Taylor and Francis.

Krahmer, E., & Swerts, M. (2004). More about brows. In Ruttkay, Z. & Pelachaud, C., (eds.), *From brows to trust: Evaluating embodied conversational agents,* pp.194–216. Kluwer Academic Press.

Krakowian, P. (2010). *Modern Test Theory Explained.* Scholar

Krakowian, P. (2011). *Investigating Rater Performance in Tests of Oral Expression*., Wydawnictwo Uniwersytetu Łódzkiego

Lustig, M. W., & Koester, J. (1993). *Intercultural Competence. Interpersonal Communication across Cultures.* Harper Collins College Publishers.

Lustig, M. W., & Koester, J. (2009). *Intercultural Competence: Interpersonal Communication Across Cultures,* 6th Edition,. Allyn and Bacon

Martinez, L. (2009). How Examiners of Different Severity Grade Candidates of Different Ability. *Test Insights 2009.* Measurement Research Associates, Inc.

Martinez, L. (2010). The Relationship between Examiner Severity and Consistency. *Test Insights 2010.* Measurement Research Associates, Inc.

Nakane, I. (2007). *Silence in Intercultural Communication: perceptions and performance.* John Benjamins.

National Research Council (NRC). (2015). *Identifying and supporting productive programs in out-of-school settings*. Washington, DC: National Academies Press.

O'Sullivan, B. (2008). *Modelling Performance in Tests of Spoken Language*. Peter Lang

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing 2009, 26*, 161-179.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores., *System*, vol 30, no 2, pp 143-154

Saint-Onge, H., & Wallace, D. (2003). *Leveraging communities of practice for strategic advantage*. Butterworth-Heinemann

Scollon R., & Scollon, S.W. (1995). *Intercultural Communication.* Blackwell

Seed, G. (2012). Perceptions of authenticity in academic test tasks., *ESOL Research Notes*, vol 49, pp. 17-26

Seiter, J., Weger, H., Jensen, A., & Kinzer, H. (2010). The Role of Background Behavior in Televised Debates: Does Displaying Nonverbal Agreement and/or Disagreement Benefit Either Debater? *Journal of Social Psychology*, *150*(3), 278–300.

Taylor, L., & Falvey, P. (2007). IELTS Collected Papers: Research in speaking and writing assessment. *Cambridge ESOL Research Notes* Volume 19

Tynan, R. (2015). Creating ePortfolios to facilitate and evidence progress using learning technologies., *Cambridge Exams Research Notes*, *61*, 55-68

Weir, C., & Milanovic, M. (2003). Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913-2002 *Cambridge ESOL Research Notes, 15*.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*.Cambridge University Press.

Wenger, E., McDermott, R.A., & Snyder, W. (2002). *Cultivating communities of practice: a guide to managing knowledge*. Harvard Business Press

Wilson, M. (2005). *Constructing Measures: An Item Response Model* Lawrence Erlbaum Associates

Wright, B.D., & Masters, G. (1982). *Rating Scale Analysis.* MESA Press

Wright, B.D., & Stone, M.H. (1979). *Best test design*. MESA Press