Przemysław Krakowian
Państwowa Wyższa Szkoła Zawodowa w Płocku

RATER PERCEPTIONS IN TESTS OF ORAL EXPRESSION

Percepcje oceny w testowanych wypowiedziach ustnych

Abstract

Assessment of spoken performance is often viewed in terms of various different categories relating to the language aspect of performance including provisions for vocabulary use and linguistic resources, as well as accuracy of use, contrasted with fluency. Such categories are often coupled with several other categories relating to other aspects of spoken performance, and amongst those one almost always featured in rating schemes is that of pronunciation. It is interesting to observe what happens when this category is omitted in rating scales and how this affects the perceptions of oral examiners. In this investigation, data from an Electronic Performance Support System (EPSS) obtained in training and fine-tuning examiners of oral performance in tests of spoken performance in English from different educational contexts is used to draw some intriguing research conclusions.

Key words: Electronic Performance Support System (EPSS)

Streszczenie

Systemy oceniania języka mówionego posługują się rozmaitymi skalami zawierającymi różne kategorie odnoszące się do jakości języka zawierające podkategorie obejmujące słownictwo i środki językowe, jak również poprawność językową zwykle na tle biegłości i potoczystości w mówieniu. Te, z kolei, uzupełnione bywają rozlicznymi jeszcze kategoriami, wśród których właściwie zawsze poczesne miejsce zajmuje wymowa. Warto jednak przyjrzeć się sytuacji, gdzie tej kategorii brak w skalach oceniania i w jaki sposób wpływa to na percepcje egzaminatorów języka mówionego. W tym badaniu dane pochodzące z Elektroniczne Systemu Wspierania Oceniania (ang. EPSS) zgromadzone w trakcie treningu i doszkalania egzaminatorów języka angielskiego z różnych kontekstów edukacyjnych pozwalają na sformułowanie intrygujących wniosków badawczych

Słowa kluczowe: Elektroniczny system Wspierania Ocen (EPSS),

1. Introduction

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].

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.

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 from different educational contexts to make a claim that what *sounds nice* may sometimes obtain more merit than it actually deserves. The data for this claim comes from hard evidence registered in Electronic Performance Support System (EPSS) which is one of the deliverables in a European Commission - Education, Audiovisual and Culture Executive Agency Minerva Program project called WebCEF. Before this claim is verified and put under scrutiny, a general outline relating to different factors in the assessment of speaking is presented and other sources of discrepancies relating to the assessment of speaking samples are analysed alongside in terms of obtained meta information and ratings of samples. Findings relating to different kinds of bias are presented alongside the main focus of this paper because bias relating to the perception of pronunciation would not have been detected if the research had not postulated and investigated bias of other kind in the data matrix.

2. Background information

Following Scollon and Scollon [1995], Nakane [2007], Lustig and Koester [1993] Hall and Hall [1990] and Barna [1994], a multicultural or intercultural communication setting may be defined as an arrangement characterised by the fact that the interaction of individual participants from different cultures, speaking different native languages and 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].

In terms of the impact of an intercultural 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 assessment in intercultural contexts 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) 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 a web-based environment collaboratively evaluating spoken performance discovered that owing to a diversity of educational and cultural backgrounds 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.

A European Commission - Education, Audiovisual and Culture Executive Agency Minerva Program project called WebCEF, aimed at enabling collaborative assessment of oral language proficiency through a Web 2.0 environment, with the idea behind the project to allow language teachers and language learners to evaluate their own video and audio samples, working together with colleagues and peers across Europe as part of a community of practice. The EPSS (Electronic Performance Support System) used for the purpose of collecting, storing and annotating samples with comments additionally stored a wealth of information concerning the samples themselves, the raters, the raters' educational and language backgrounds and any other information the users of the system cared to log in as part of the assessment procedures.

In total, 139 raters of different educational, professional and national background were registered in the WebCEF EPSS, of whom, following the available data, 72 were included in the analysis of the ratings in the English Showcase. The Annotation tool contains record concerning 267 samples for all partner languages, of which 109 were recorded in the English Showcase portion of the project, 79 of which contained sufficient ratings performed by different raters to permit analysis.

3. Research objectives

When inspected, the EPSS shows that apart from a showcase of benchmarked samples where convergence in rating is very high, an inordinate number of assessment instances differ considerably, where this divergence may be attributed to a number of factors: i) possible sources of discrepancies in marking by different assessors; ii) cultural 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.

Inter 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.

Intra 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 CEF scales in the project the inescapable conclusions seem to be that: i) the overall scales for either of the tasks in the Showcase are complicated; 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 relative phenomenon and it may be concluded that it is less so in the case of national languages and partner Showcases, but definitely in the case of the English Showcase, which was constructed based on efforts of individuals from different cultures, backgrounds and with different educational and evaluation histories, where the effects of the intercultural communication and intercultural communication interference were sometimes making the partners read in different ideas into the descriptors as [i]*ntercultural communication 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, s.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, Dell 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. Next, 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. Unfortunately, such claims remain largely unsubstantiated, as the number of samples with female vs. male pairing and the available metadata on the assessors participating in the evaluations is fragmentary and insufficient for performing the type of analysis outlined in the following sections.

Apart from the above, the effect of the native background culture of the assessor that might prove important when evaluating students from other countries and cultures in the sense as it is understood by 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 pertain to the portion of samples accumulated in the EPSS database relevant to the English speaking tasks. Similar material, albeit in smaller number, exists for other languages, but is not subject of investigation in the present study. The data were divided into several Research Groupings (RG's) and inspected for variability with the Chi-Square test as a goodness of fit test, where the standard from which departure was tallied was based on two separate models of performance. The first model (M1 and sig.1) was based on the average of all assessments including those made by the group under scrutiny. The second model (M2 and sig.2) was based on the average of the assessments remaining after the group under scrutiny was identified and other assessments have been eliminated. The result was considered significant if either (or both) Chi-Square tests performed for a particular Research Grouping (RG) were statistically significant (M1 and/or M2 p-values in excess of 0.5). The result was considered moderately significant if both p-values pertaining to respective models were in excess of 0.1, but not 0.5.

The first of the observations concerns the procedure performed prior to this analysis and relating to the alleged trend identified at the onset of discussion on convergence. The original claim that effeminate male language tends to be decidedly underscored by male raters was confirmed on a population of 11 samples involving the assessments of 13 raters, 7 of whom were male raters (RG7, M2, sig.1=0.90023, sig.2=0.02397). The second of the postulated trends of the early study, namely that effeminate male language tends to be statistically significantly overscored by female raters, but only of some nationalities, was only partly identified for 3 female raters whose native language was English (RG6, M2, sig.1=0.89031, sig.2=0.75512). The claim that female raters rate more leniently was confirmed at the statistically significant level for 19 female raters of different native languages (RG5, M4, sig.1=0.49047, sig.2=0.03267). The tendency to evaluate more favourably those whom the raters are familiar with was confirmed at the statistically significant level for 45 raters of all partner native languages (RG11, M7, sig.1=0.34021, sig.2=0.04175). The familiarity with the subjects was determined on the premise of the authorship of the speech sample if such information was present in the metadata and the sample was not submitted anonymously. The effect of the examinee pairing could not be investigated due to insufficient metadata and a relatively smaller number of samples for interactive tasks compared with the total number of samples. The effect of the native background culture of the assessor understood as high and low context culture interference was investigated in relation to the Finnish project partner (RG8 comprising 3 female raters forming RG9 and 4 male raters belonging to RG10), a decidedly high context culture notorious for exceptional tolerance for silence and ambiguity [Nakane, 2007]. RG8 consistently and statistically significantly overated samples of considerably smaller discourse size, shorter or/and containing more pauses and hesitations (RG8, M1, sig.1=0.40231, sig.2=0.08725) and did that irrespective of gender (RG9, M1, sig.1=0.44321, sig.2=0.08515, RG10, M1, sig.1=0.41845, sig.2=0.08817).

Finally, a Polish twist in the data, the one which is the focus of this paper, which seems to be pointing in the direction of the 11 Polish raters (RG12) scoring samples with greater perceived phonetic accuracy more favourably than the rest of the raters (RG12, M1, sig.1=0.83954, sig.2=0.04235). 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* or what *souned nice*.

The data pertaining to the portion of samples accumulated in the EPSS relevant to the English speaking tasks was additionally subjected to analysis using the Multi Facet Rasch Analysis in order to confirm the postulated and identified trends and, if possible, to identify additional trends if any. 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.

In the case of the logit based statistic in Rasch Analysis, the residual is essentially based on the same principle, 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, 200; 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, performed using FACETS [Wright and Stone, 1979; Wright and Masters, 1982] a Many-Facet Rasch Analysis dichotomous and polytomous model program, and RaterGrinder PRO, a program developed by the author for the purpose of similar analyses, whose performance was validated in this procedure by comparing the performance of the two programs, did indeed confirm the existence of the groups postulated earlier in the *a priori* analysis, but only after the raters and samples were sorted by origin and working language and the analysis performed only for those raters who evaluated the same samples, or in groups which overlapped by at least two samples or by at least two raters [Wright and Stone, 1979; Wright and Masters, 1982]. It additionally revealed several interesting trends and tendencies, which could not have been predicted in the Chi-Square study.

Firstly, regarding the postulated observation concerning effeminate male language and the tendency for it to be underscored by male raters - this was confirmed on the original 11 samples involving 13 raters, with the Rasch Analysis additionally pointing in the direction of two more female raters (originally 7 male and 6 female, making now the total of 8 females, both whose nationality was German).

Additionally, the claim investigated in connection with gender perception concerning a tendency for effeminate male language to be statistically significantly overscored by female raters of some nationalities, was confirmed for the original 3 female raters whose declared working or native language was English, with the addition of one further British female for whom the outfit statistic was elevated but not critical – additionally two British females were identified as overly lenient, a tendency, which was later investigated for all raters, allowing to identify three essential patterns overall, that is for moderate, lenient and severe markers, with all three groups characterised by relatively smaller variance in their markings compared with other markers [cf. Martinez, 2009].

Rasch Analysis also 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 with – this was compared against the composition of marker groups investigated with the Chi-Square procedure and was confirmed for the same 45 raters belonging to a variety of rater native languages – additionally, the analysis identified 3 subsequent raters, and 11 further samples, whose origin and relation to markers cannot be confirmed through the analysis of the sample metadata, but which exhibited similar behaviour in Rasch Analysis, pointing perhaps to undisclosed familiarity with the samples and examinees.

What is more, samples in which the performance was recorded for two subjects i.e.: collaborative tasks, were analysed in search of the effect of examinee pairing, which in the Chi-Square procedure could not be investigated due to insufficient metadata and a relatively smaller number of samples for interactive tasks compared with the total number of samples – this identified 4 instances of performance by 2 paired teams, where elevated marks in comparison with their performance in monologue samples could lead to a tentative conclusion that perhaps the rapport between the examinees was a deciding factor which influenced the perception of performance of the subjects involved.

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, in which the effect of the native background culture of the assessor, namely tolerance for silence and ambiguity, was investigated – the difference was in the number of raters, where of the original 6 raters only 4 were deemed as consistently biased and two had elevated t-fit indices.

If greater phonetic accuracy is determined as that registering in the upper three CEF bands, Rasch Analysis allowed to identify 11 Polish raters, the original RG12 in the Chi-Square study, and additional 4 non-native English speaking 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 nationality or affiliation-wise – a closer inspection of the samples, however, revealed that while the students in the samples belonged to a variety of backgrounds, both educationally as well by nationality and native language, the feature they supposedly 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. It can be carefully assumed that some of them, such as the effects of examinee pairing or the familiarity with the examinees may be applied to all contexts. Some, however, can only be observed in a multicultural context, however that may be defined for the purpose of operalisation. 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 obtained from the project participants. Uncertainty as to the familiarity 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.

On top of that, a largely unexplored portion of data relating to spoken production in languages other than English still resides in the EPPS, as it was mentioned earlier. While the author has a working/survival command of French, and some proficiency in languages other than the ones registered in the EPSS, several European languages, numerous samples of performance and their perceptions in the form of the ratings remain unexplored. Further research on the data could yield very interesting results, especially as those would relate to languages less often discussed in professional literature of the subject.

REFERENCES

Bachman L.F., and Palmer, A. 2002. *Language Testing in Practice.* Oxford: University Press.

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

Bennet M.J. 1993. Towards Ethnorelativism: A Developmental Model of Intercultural Sensitivity. In *Education for the Intercultural Experience.* Yarmouth: Intercultural Press.

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

Brown A., 2006. "An examination of the rating process in the revised IELTS Speaking Test". *IELTS Research Reports* Volume 6, IELTS Australia, Canberra and British Council, London.

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

Chambers L., Ingham K. 2011. "The BULATS Online SpeakingTest". *ESOL Research Notes*, vol 34.

De Velle S. 2008. "The revised IELTS Pronunciation scale". *ESOL Research Notes*, vol 34.

Fulcher G. 2003. *Testing Second Language Speaking.* Harlow: 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*.

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

Hall E. 1959. *The silent language*. Garden City: Doubleday.

Hall E. 1966. *The hidden dimension.* Garden Day: 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. and Kimble Ch. 2004. *Knowledge networks: innovation through communities of practice.* Idea Group Inc.

Hubbard C., Gilbert S. and Pidcock J. 2006. "Assessment processes in speaking tests: a pilot verbal protocol study". *ESOL Research Notes*, vol 24.

Hubbard Ch. 2011. "Cambridge ESOL Professional Support Network Extranet: Development and impact". *ESOL Research Notes*, vol 49.

Hymes D. 1964. Introduction: "Toward Ethnographies of Communication". *American Anthropologist* 66, 6.

Hymes D. 1972. On communicative competence. In *Sociolinguistics*, eds J. B. Pride and J. Holmes. Harmondsworth: Penguin.

Krahmer E. and Swerts M. 2004. More about brows. In *From brows to trust: Evaluating embodied conversational agents*, eds. Ruttkay, Z. and Pelachaud, C., Norwell: Kluwer Academic Press.

Krakowian P. 2010. *Modern Test Theory Explained.* Warszawa: Scholar.

Krakowian P. 2011. *Investigating Rater Performance in Tests of Oral Expression*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego

Lustig M.W. and Koester J. 1993. *Intercultural Competence. Interpersonal Communication across Cultures.* New York: Harper Collins College Publishers.

Lustig M.W. and Koester J. 2009. *Intercultural Competence: Interpersonal Communication Across Cultures*. Boston: 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.* Amsterdam: John Benjamins.

O'Sullivan B. 2008. *Modelling Performance in Tests of Spoken Language*. Frankfurt: 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*, 26.

Orr M. 2002. "The FCE Speaking test: using rater reports to help interpret test scores". *System*, vol 30, no 2.

Saint-Onge H. and Wallace D. 2003. *Leveraging communities of practice for strategic advantage*. Oxford: Butterworth-Heinemann.

Scollon R. and Scollon S.W. 1995. *Intercultural Communication.* Oxford: Blackwell

Seed G. 2012. "Perceptions of authenticity in academic test tasks". *ESOL Research Notes*, vol 49.

Seiter J., Weger H., Jensen A. and 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* Vol. 150, No. 3.

Taylor L., and 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*, vol 61

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

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

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

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

Wright B.D. and Masters G. 1982. *Rating Scale Analysis.* San Diego: MESA Press

Wright B.D. and Stone M.H. 1979. *Best test design*. San Diego: MESA Press