DOI: http://dx.doi.org/10.18778/8142-988-7.10

Multi Facet Rasch Analysis in tests of oral production

Przemysław Krakowian

University of Łódź, Poland

Abstract
The difficulty in constructing tests of oral production lies with the rating schemes and assessment frameworks, their construction and validation and subsequently application. This paper advocates extensive use of numerical data and statistical procedures alongside qualitative and intuitive methodologies, where the common denominator lies in the fact that all of the empirical measures involved make use of some form of a performance model allowing to make predictions about the examinee behaviour in order to verify goodness of fit of the observable rating data.

Key words: testing, evaluation, validation, Multi Facet Rasch Analysis

1. Introduction

Weir (2005), in an attempt to look at language testing in an objective, empirical way, coined the term *evidence-based approach*. In a sense, this paper will follow a similar tenet of gathering hard evidence to make and substantiate claims about the merits of rating scales and the nature of oral examiner performance. While Weir's approach was based on evidence gathered at different stages of the process of test construction, administration and evaluation, and made use of an array of information collected at distinct points in test preparation and operation, the information only sometimes was of quantitative sort. In Weir's opinion relying on the evidence obtained at various stages of test construction, preparation, administration and analysis is crucial in determining the value of the test, its reliability, validity and effectiveness. Such evidence is also instrumental in performing the necessary adjustments to the procedure and operation; however, the evidence-based approach advocated by him sometimes relies on evidence that could be highly subjective and thus prone to errors of judgement.

There are several reasons why Weir's approach makes no consistent use of numerical data: firstly, because goodness of fit procedures require large scale

observations of performance; secondly, because constructs for such observations need to be defined as traits rather than competencies and skills; and thirdly, because unlike tests of writing or composition as well as tests of receptive skills, grammar or vocabulary, oral performance leaves no permanent record that can be re-examined at will. Finally, even if recordings of spoken production allowed researchers to take their leisure examining the data, it is rarely that oral performance is repeatedly evaluated by different examiners, thus producing a matrix that may be analysed with the statistical procedures that will be described later in this text.

The procedures presented on the subsequent pages of this paper may be treated as a set of guidelines for any exam construction and evaluation procedure. They involve a number of stages and require different procedures to be performed before a claim can be made that the rating scales; tasks and examiner behaviour have all been validated. The novelty of the approach assumed here lies in the fact that all of the procedures have been assembled together for the first time and constitute a convergence of three methodologies: intuitive, quantitative and qualitative, with the premise in mind that they all are limited with research shortcomings that can be reduced and perhaps even eliminated through the tri-partite arrangement assumed for this exposition.

2. Developing tests of oral expression

The process of constructing tests of oral expression, more commonly referred to as speaking tests, fits within a more general framework of test construction postulated as early as in the 40's of the previous century by Hughes (1946), later picked upon by various other practitioners (Rasch, 1960; Rasch 1980; Berk, 1984; Brennan, 1984; Cziko, 1984; Davies, 1977; Davies, 1990; Douglas & Selinker, 1985; Theunissen, 1987; Allen, Cummins, Mougeon, & Swain 1983; Alderson & Buck, 1993) to finally gain recognition in several texts after the seminal publication by Alderson et al. in 1995 (cf.: Luoma, 2004; Bond & Fox, 2007; Fox, Wesche & Bayliss, 2007; Fulcher & Davidson, 2007; Douglas, 2009; Taylor, 2011; Fulcher, 2013; Fulcher, 2014; Leong, Bartram & Iliescu, 2016).

In essence, this arrangement suggests that a number of distinct stages lead to the administration of the test, while information obtained at each of the stages, together with the information acquired from the administration itself provide the test constructor with valuable insights into the mechanics and functioning of the test allowing the administrators at the same time to implement any necessary changes, adjustments and modifications aiming at improving the procedure. This is very much consistent with Weir's evidence-based approach (2005). It is interesting to

note, however, that this idea was conceived much earlier, and coherently put forth by Alderson et al. (1995). Both Alderson's et al. (1995) and Weir's (2005) ideas, despite implicit notions to the contrary, explicitly suggest that this is a linear, one time process.

Luoma (2004, p. 4) makes a more convincing argument presenting test construction and development as an ongoing, never-ending and continual process, which she does by presenting it graphically in the form of a circle, which presents stages of test development. This circular exemplification is more convincing in the sense that it captures the repeated nature of efforts involved, but at the same time suggests that with subsequent administrations each of the stages is not merely overhauled but altogether re-invented, which, it is conceivable to assume, nearly never happens, if only for logistical reasons.

Alderson and Buck (1993) claim that a total overhaul usually results from a re-formulation of constructs underlying the testing framework, or technological advances allowing to analyse test performance and stake claims concerning constructs with greater accuracy, and estimate of a life cycle of a commercial test battery such as that of ESOL to be approximately ten years (Docherty & Corkill 2015), whereupon the test is re-evaluated, revamped and replaced by another, sometimes substantially different instance of operations.

Nonetheless, the stages that can be identified, though under different headings with various authors, come down to: i) the planning stage or the test specification stage where constructs are defined; ii) task development stage where, based on the constructs, tasks are written and moderated; iii) rating scale construction and verification; iv) training stage both for raters and administrators, and, finally, v) test evaluation and research stage. Once set in motion, the process continues till the test is decommissioned, usually being replaced by another test or examination (Luoma, 2004; Douglas, 2009; Taylor, 2011; Fulcher, 2013; Docherty & Corkill, 2015).

3. The planning and test specification stage

The planning and specification stage is crucial in a number of ways to the viability of the examination. The issues connected with the standarisation and verification pertaining to reliability and validity in connection with more general considerations in language testing, exam development and evaluation, construct development and operationalisation have been presented in numerous texts in the field (Luoma, 2004; Bond & Fox, 2007; Fox, Wesche & Bayliss, 2007; Fulcher & Davidson, 2007; Douglas, 2009; Krakowian, 2010, 2011; Taylor, 2011; Fulcher, 2013; Fulcher, 2014; Leong, Bartram & Iliescu, 2016).

A separate set of concerns is related to the design of the operationalisation of constructs and the design of constructs themselves, as rating scale analysis and the analyses of examiner performance can be accomplished under any set of circumstances, but binding conclusions can be drawn and problem areas can be identified and explained if constructs and their operationalisations are of a particular type (Taylor & Falvey, 2007). This is not so much a matter of the contents of the specifications as it is an issue of how constructs are formulated, as it has been claimed on numerous occasions by Fulcher and Davidson (2007), McNamara (2000), Harley, Allen, Cummins & Swain (1990).

In their collective opinions, constructs need to be defined as traits rather than competencies and skills, which is something that in turn reflects the intention to pursue such mode of design of ensuing rating scales that will be simple enough to adhere to, given the reservation already expressed earlier that any descriptor which is more elaborate than two clauses is largely disregarded as binding in assessment (Fulcher, 1994; Taylor & Falvey, 2007; Cambridge English: RN 2007 issue 30). Such design would need to be supported by an extensive, compound and robust framework drawing on the nature of latent traits. The idea that the complexity of human behaviour, including language behaviour, can be explained by a multitude of micro-traits which can later be reduced to a more general latent trait through trends analysis and correlational reduction is not new and has been advocated as early as Stevens (1946), Rasch (1960, 1980) Michell (1997), and pointed out more recently in educational applications and assessment by Embretson & Reise (2000) Kemp (2006) and Feary (2009).

A number of underlying latent traits have been identified in connection with second language oral proficiency in several prominent studies as far back as in the late 50's, early and mid 80's and early 90's of the previous century (Campbell & Fiske, 1959; Harley, 1987; Kavanagh, MacKinney & Wolins, 1971; Anderson, Bachman, Perkins & Cohen, 1991; Griffin, 1985; Adams, Griffin, & Martin 1987; Henning, 1992; Oltman & Stricker, 1990; Boldt, 1989). Confirmed accounts of analyses leading to the identification of a number of interrelated micro-traits in three main latent traits accounting for oral language ability and oral communication resulted in a tentative conclusion that oral production can be described and analysed in terms of: i) grammar trait understood in terms of range and accuracy of morphology, lexis and syntax; ii) discourse trait perceived as the abilities comprising the capacity to understand and produce coherent and cohesive text, including specific linguistic realisations of coherence; iii) sociolinguistic trait envisaged as the ability to produce and recognise language that is socially acceptable within a particular set of contexts, including the ability to execute a planning and strategic component that is in accordance with social conventions (Bailey, 1998; McNamara, 2000; Fulcher & Davidson, 2007; Fulcher, 2014; Leong, Bartram and Iliescu, 2016).

4. Task development stage and task moderation

The next logical step in the development of oral language tests, indeed in the development of any test and examination, is to determine the operationalisation of the constructs envisaged by traits, by designing the set of procedures and operations through which they will manifest themselves in the form which allows rating (Milanovic, Saville, Pollitt & Cook 1996; Skehan, 1998; Bygate, Skehan & Swain 2001; Hughes, 2002; Galaczi & Ffrench, 2007). This usually involves outlining the general procedure of the oral examination or interview, drafting tasks and moderating in order to determine construct validity of the individual tasks and the whole procedure. Galaczi & Ffrench (2007) point out that it is common practice to seek out and pursue existing rating scales to determine their suitability for measuring the postulated constructs, as designing tasks with a rating scale in mind is easier and more expedient than designing tasks from scratch and building rating scales to reflect construct related operations that are supposed to be obligatory in the process of task completion.

Nonetheless, it is conceivable to imagine that in some educational and evaluative contexts, no suitable scales exist that could be used as a starting point for adapting and developing performance descriptors and ensuing appropriate tasks. Such situations require foresight and perhaps a certain amount of experience, as a badly designed task, for instance in terms of scope and range (Upshur & Turner, 1995) may be so undemanding for the test taker that it may result in underscoring of ability, as the examinees are not offered the opportunity to exhibit their full spectrum of language potential. Another danger lies in the fact that envisaging such undemanding tasks as sufficient to elicit the desired types of language behaviour leads to under-representing constructs (Skehan, 1998; Bygate, 1987; O'Loughlin, 2001). Weir & Milanovic (2003), Hawkey (2009) and Martyniuk (2010) all additionally point out the necessity for diversity in exams of oral proficiency, which in their opinion is the only guarantee of representativeness and consequently also construct validity.

In his analysis of UCLES now ESOL, Hawkey (2009) identifies how the constructs recognized in the design stage for the spoken competence result in the composition of the oral interview tasks and how they are reflected in the rating scales. Hawkey (2009) expounds how the FCE speaking tasks are designed in order to account for postulated constructs such as coherence and cohesion in a range of contexts conceivable for a SL/FL learner of English. In an interview consisting of four parts, the interviewer assumes two distinct roles in interacting with the examinee, that of an organiser of events and participant of discussion (Hawkey, 2009). Despite being heavily scripted on the part of the interviewer,

the interaction is designed to look and feel spontaneous and natural and prompts the examinee to use appropriate register for each type of interaction, though both of them preclude becoming too friendly with the interviewer. In addition to that, the informal register is required when the examinee interacts with a peer, and Hawkey (2009) claims that efforts are made to pair FCE candidates according to age based on the available applicant information, though the present author's experience as an ESOL examiner point to the contrary in numerous observations. The candidate's language production involves providing factual information on request, presenting an opinion, negotiating a point of view, explicating, describing and narrating content (Suto, Greatorex & Nadas 2009; Hawkey, 2009).

The variety of contexts in which the language is produced is designed to reflect differences in how coherence and cohesion are achieved in different types of discourse and is meant to be an adequate reflection of the underlying construct. Evidence, however, exists (Suto, Greatorex & Nadas, 2009; Taylor & Falvey, 2007; Laming, 2004; Weir & Milanovic, 2003) that despite the efforts to design tasks to be instrumental in eliciting construct postulated behaviours, with reference to cohesion and coherence, rating scales seem to be vague enough to encourage raters to formulate judgements reflecting cohesion rather than coherence or both. This clearly points to the need for careful, thorough and informed rating scale construction and verification, using a variety of sources of insight.

5. Rating scale construction and verification

Rating scale construction involves a number of operations that need to be performed in order to ensure that the scales constitute a fair reflection of underlying constructs, provide a convincing translation of the operations involved in successful task realisation into postulated behaviours and grant the examiner ease in making decisions as to which of the descriptors of performance is applicable when looking at individual aspects of performance (Milanovic, Saville, Pollitt & Cook, 1996; Galaczi & Ffrench 2007; Fulcher & Davidson, 2007; Douglas, 2009; Taylor, 2011). Once the constructs have undergone operationalisation through task design, descriptors are formulated and *an priori* validation is performed. This involves intuitive, qualitative and quantitative methods of analysis applied at different stages of rating scale construction. The *an priori* validation is followed by a small scale pilot involving actual examinee performance on tasks, to finally conclude in *a posteriori* in-depth analyses of rater performance in a regular administration (Milanovic, Saville, Pollitt & Cook, 1996; Galaczi & Ffrench, 2007; Taylor 2000; Weir & Milanovic, 2003; Fulcher, 1996; Hawkey, 2004; Laming, 2004).

While *an priori* validation and a dry, test run belong to the stage of rating scale construction and validation, a full scale validation is part of the exam performance investigation and provides a much more thorough and exhaustive picture of the process owing to the fact that much more comprehensive data is available offering a much wider-range account of the examinee behaviour and examiner rating scale interactions. Naturally, since the data obtained in the pilot is smaller and less exhaustive it requires other sources of insight to engender confidence of the test constructors in the performance of scales. Nonetheless, empirical validation of scales prior to the scales being used in an actual, live examination has since the 90's of the previous century become a staple practice in numerous educational and examination contexts. Laming (2004) and Milanovic, Saville, Pollitt & Cook (1996) claim that on one hand this has been so owing to availability of expertise concerning the application Multi Facet Rasch Analysis (MFRA), but increasingly so owing to the MFRA procedures becoming requisite and obligatory in various testing and assessment communities, and becoming more and more the norm in appraising the viability and validity of rating scales.

6. The Principles of Rash Analysis and Testing Relevance

The class of models referred to on subsequent pages is named after Georg Rasch, a Danish mathematician and statistician who postulated them in the late 1950's and early 1960's (Rasch 1960, 1980), and which were later elaborated on by Wright and Stone (1979) and Wright & Masters (1982). It was Wright & Stone and the MESA Psychometric Laboratory in Chicago who publicised Rasch's theories and who created computer models for their implementation in the form of BIGSTEPS, a computer program for two facet Rasch analysis and FACETS, a Multi Facet Rasch Analysis program. The sections below outline the major tenets of the theory.

7. The Concept of Latent Traits (LT's)

Before Rasch analysis, extended Rasch analysis models or the Multi Facet Rasch Analysis can be delineated, a central concept necessary for the understanding of the Rasch rationale needs to be introduced. The term *latent trait* in psychometrics is derived from psychology, and it refers to a psychological dimension necessary for the description of an individual and is assumed to underlie and explain observed behaviour of that individual (Bond & Fox, 2007; Salkind, 2007; Kaplan

& Saccuzzo, 2009). In relation to language testing, latent traits are those relatively stable characteristics, attributes or capacities which account for the consistencies in the behaviour of the individual or a group of individuals.

Latent traits have been postulated both to be fixed, unchanging and stable entities (Anastasi 1988), but also as phenomena characterised by change, adaptation and augmentation. In short, though being stable, at the same time they are inherently dynamic and interactive (Bond & Fox, 2007). Lord & Novick (1968) and Kaplan & Saccuzzo (2009), however, logically point out that the actual account of the nature of the latent trait has no implications for the mathematical models of mental performance; it matters, though, at the level where assumptions are made about the content of the language test.

8. The Notion of the Item Characteristic Curve (ICC)

An essential feature of the Rasch model is that of a relationship between the observable performance of individuals in an assessment situation and the unobservable underlining characteristics or abilities responsible for that performance (Bond & Fox, 2007). That relationship is described by the *Response Function* or the *Item Characteristic Curve* (ICC), which is a curve relating the probability of a desired behaviour of a person in a task or a set of tasks or items to such parameters as ability and difficulty. Various ways of formulating this curve have been proposed by numerous existing Rasch models, all of which have made the assumption that the rate of success depends on the information about the person's ability and the difficulty of the task, with some models incorporating additional variables. When the probability of a correct answer is expressed as a function of ability, such an expression is referred to as the Test Characteristic Curve (TCC) or when tasks are composed of items as the case is with pen and paper test Item Characteristic Curve in short ICC (Bond & Fox, 2007).

A distinction is often made between theoretical ICCs and empirical ones, i.e.: ones obtained from a set of response data (Bond & Fox, 2007) Historically speaking ICCs have often been formulated by observing empirical data, as a starting point for the development of response models (Bond & Fox, 2007; Salkind, 2007; Kaplan & Saccuzzo, 2009).

9. The Measurement Models

A family of possible curves exists for such a relationship and accounts of various types of mathematical models are given by e.g. Hersen (2003), Bond & Fox (2007) and Salkind (2007). The major difference between those various models lies in how the responses are evaluated and scored, and how that scoring reflects the relationship between the dimensionality of the response data and the number of traits assumed to underlie that data. For clarity of the argument and for practical reasons connected with the interpretation of data, the following considerations account for a single underlying trait and assume that only two factors will come into play raters and samples. Such models representing task-person or rater-sample interaction are referred to as dichotomous or two facet uni-dimensional probabilistic response models, which rely on the exponential and logarithmic (logistic) functions, and therefore are also known as logistic response models (Lord & Novick, 1968; Bond & Fox, 2007; Salkind, 2007; Kaplan & Saccuzzo, 2009). And while principally, this logic is applied to investigating person and item interaction in paper and pen based tests, more and more often the principles of Rasch Response Models are applied in attempts to assure quality and control the process of evaluating subjectively rated tests. Such procedures look predominantly at the information relating to the goodness of fit of the observed performance data to the data postulated by the model.

Numerous programs and procedures exist that are capable of performing such analyses (cf. Krakowian, 2010, 2011). BIGSTEPS and FACETS, are by far the best known, and what is more, now feely available in the original DOS based versions, following the development of newer more user-friendly, and commercially available versions. They both provide the goodness of fit information as unweighted, or infit, and information weighted, or outfit, indices which provide some measures for dealing with aberrant rater behaviour patterns. While the *t-fit* goodness of fit index is indicative of how well or how poorly the rating pattern adheres to the model, outfit and infit may be somewhat instrumental in detecting patterns of ratings that are overly lenient or overly strict or severe. Neither of them separately or together, however, can be exhaustively indicative of raters rating carelessly without paying attention to the true merit of the samples (Wright & Masters, 1982; Bond & Fox, 2007; Salkind, 2007; Kaplan & Saccuzzo, 2009). A freeware program for Multi Facet Rasch Analysis, RarterGrinder, is available at the Institute of English Studies at Łódź University, and its operations are somewhat documented in two studies by Krakowian (2011). The main focus of the program, apart from providing the usual indices of goodness of fit is to explicitly identify the aberrant patterns of behaviour in raters and to indicate, leniency and severity as well as careless ratings.

From the point of view of reliability of the rating in a test of speaking, raters who are too lenient, too severe, just as the ones who play it safe and tend to assign the same or largely similar grades for performance of different quality, or who assign grades that bear little or no relation to the quality of the performance, should be identified and dealt with at the stage of rater training and verification, before the live roll out of the test (Bond & Fox, 2007; Salkind 2007, Kaplan & Saccuzzo, 2009). The stage of rater training is perhaps the only time in test construction and administration where all or most of the raters have to deal with the same samples of oral production and when their performance can be collectively evaluated for the purpose of providing feedback and retraining.

The departure from a model of performance in Multi Facet Rasch Analysis (MFRA) in the case of both BIGSTEPS and FACETS is measured using a goodness of fit test, essentially a test that is indicative of how well a set of empirical data, such as rater performance data in a test of oral production, fits the postulated model (Bond & Fox, 2007). There are numerous tests of fit available, but both programs use a residual based goodness of fit statistic estimated in an iterative procedure called UCON (Wright and Stone 1979). This procedure, however, is capable of accounting for undesired rater behaviour only on the premise that since some samples are rated correctly by a smaller number of raters they should be considered more difficult. The degree of departure from the model is estimated based on the implausibility of the response and not on the actual difference between the rater rating and the model. In practical terms, this means that in analysing response patters a number of different indices need to be taken into consideration at the same time, and the observations sometimes may be considered as guesswork rather than binding conclusions, especially so in situations in which numerous raters exhibit different rating patterns rather than consistently underrate or overate certain samples or groups of samples, as can be seen below in Figure 1:

Obsvd Score	Obsvd Count	Obsvd Average	Fair-Z Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	N. Rater
61	17	3.6	3.41	1.36	.39	.09	7	**2.0**	5	.00	1
55	17	3.2	3.35	.79	.49	0.3	-2	0.3	-2	.98	2
5	17	3.3	3.43	-1.36	.49	**4.5**	-1	**8.4**	-1	.98	3
25	16	1.6	1.21	0.71	.70	1.3	4	.6	4	.11	4
56	17	3.3	3.43	-.25	.49	**1.7**	-2	**2.5**	-1	.98	5
63	18	3.5	3.51	1.34	.46	**1.9**	8	**1.6**	6	.15	6
58	17	3.4	3.61	1.35	.49	**1.7**	-3	1.2	-2	.99	7
93	17	3.5	5.89	1.35	.66	.3	4	1.0	9	.11	8
Obsvd Score	Obsvd Count	Obsvd Average	Fair-Z Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	N. Rater

Fig. 1. FACETS printout listing infit and outfit statistics for raters (developer data)

A different approach is assumed in the case of RaterGrinder, where a dedicated set of indices is used in the analysis of rater behaviour. Apart from the *t-fit* goodness of fit statistic and rater measure, the program provides a summary severity/leniency measure, additionally broken down into separate severity and leniency indices. All three measures are logit-based, which makes them suitable for relating to the rater measure and fit indices, as they are represented in the same scale and change with the same magnitude order.

No	Measure	MN.SQ	t-fit stat.	Overall S/L	Leniency	Severity	No
1	1.375633	0.7142857	-0.1951748	-2.197225	0	2.197225	1
2	0.7611303	1.785714	0.4415075	-1.386294	0	1.098612	2
3	-1.375633	3.928571	1.600518	-2.079442	0	2.079442	3
4	-0.7611303	3.214286	1.261665	1.94591	1.94591	0	4
5	-0.2450292	2.5	0.8771017	-2.484907	1.609438	1.94591	5
6	1.375633	3.571429	1.623343	-2.197225	1.386294	1.609438	6
7	1.37563	3.571443	1.643233	-1.791759	0	1.609438	7
8	1.375633	1.428571	0.2130144	-1.098612	0	1.098612	8

Fig. 2. RaterGrinder indices of severity and leniency (developer data)

Even superficial analysis of tables in Figures 1 and 2 shows that FACETS provides less information on patterns of rater behaviour. While raters 3, 5, 6 and 7 can be identified as suspicious, until respective variances in their response patters are analysed (Bond and Fox 2007) or patterns themselves are investigated, it is difficult to arrive at binding conclusions. RG, on the other hand, helps to determine that raters 1, 3 and 7 are overly severe, rater 4 is overly lenient, raters 5 and 6 are indiscriminate in their ratings. In the process of rating scale construction and verification, as well as at later stages in the process of rater training and exam review and rater performance review, information of this kind offers invaluable insights into the mechanics of the examination.

It is now considered more or less a norm in educational assessment and the testing industry to perform empirical validation in relatively extended trials to confirm the soundness of the descriptors, especially in procedures of mapping the descriptors under development with descriptors of already established status, reputation and recognition in order to concurrently validate the scales and provide a point of reference to potential users of the exam scores (Taylor & Falvey, 2007). This is becoming especially common in connection with CEF rather than with any other examination (Martyniuk, 2010). While substantial and oftentimes satisfactory validation of the scales can be performed prior to a live administration, a full quantitative analysis of the scale and rater performance can only be performed *post factum*, once the data has been collected from the actual administration.

10. Final notes

This paper looked at ideas in relation with validating oral performance assessment frameworks and procedures involved in investigating marker performance and detecting marker bias. The approach assumed here advocates extensive use of numerical data and statistical procedures alongside qualitative and intuitive methodologies, where the common denominator lies in the fact that all of the empirical measures involved make use of some form of a performance model allowing to make predictions about the directly observable behaviour in order to verify goodness of fit of the observable data with the postulated model of performance, while the intuitive and qualitative procedures prepare ground for analysis of hard facts.

In the course of this paper the notion of Multi Facet Rasch Analysis (MFRA) was introduced to show how the goodness of fit indices may be applied to identify unusual rater behaviour in tests of oral expression. However, in order to arrive at a protocol to ensure satisfactory control of the process of implementation and maintenance in the assessment of oral expression, the paper postulated convergence of three methodologies: intuitive, quantitative and qualitative, with the premise in mind that they all are limited with design shortcomings that can be reduced and eliminated through such a tri-partite arrangement.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

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

Docherty, C. & Corkill, D. (2015). Test construction: The Cambridge English approach. *Cambridge English: Research* Notes, 59, 10–14.

Douglas, D. (2009). *Understanding Language Testing.* London: Hodder Education.

Fox, J., Wesche, M., & Bayliss, D. (2007). *Language Testing Reconsidered*. University of Ottawa Press/Les Presses de l'Université d'Ottawa.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: an advanced Resource Book*. London, New York: Routledge.

Fulcher, G. (2013). *Practical Language Testing*. New York: Taylor & Francis.

Fulcher, G. (2014). *Testing second language speaking*. London: Routledge.

Fulcher, G., & Davidson, F. (2013). *The Routledge Handbook of Language Testing.* Hoboken: Taylor and Francis.

Galaczi, E. & Ffrench, A. (2007). Developing revised assessment scales for Main Suite and BEC Speaking tests. *Cambridge ESOL Research Notes Issue 30*, 28–30.

Krakowian, P. (2010). *Modern Test Theory Explained.* Warszawa: Scholar.
Krakowian, P. (2011). *Investig.ating Rater Behaviour in Tests of Oral Expression*. Łódź: WUŁ.
Leong, F. T., Bartram, D., & Iliescu, D. (2016). *The ITC International Handbook of Testing and Assessment*. Oxford University Press.
Luoma, S. (2004). *Assessing Speaking.* Cambridge: Cambridge University Press
Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainnment Tests. Vol.1 in Studies in Mathematical Psychology*. Danmarks Paedagogiske Institut.
Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: The University of Chicago Press.
Salkind, N. S. (2007). *Encyclopedia of measurement and statistics.* New York: Sage.
Taylor, L. B. (2011). *Examining Speaking : Research and Practice in Assessing Second Language Speaking*. Cambridge, New York: Cambridge University Press.
Weir, C. (2005). *Language Testing and Validation: An Evidence-based Approach.* Basingstoke: Palgrave-Macmillan.
Wright, B. D. & Masters, G. (1982). *Rating Scale Analysis.* San Diego: MESA Press.
Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. San Diego: MESA Press.