ISSN 2074-9864 (online)

vienna english working papers

views

VIENNA ENGLISH WORKING PAPERS

HTTP://ANGLISTIK.UNIVIE.AC.AT/VIEWS/

VOLUME 25 2016

ARMIN BERGER

Rating scales for assessing academic speaking: A data-based definition of progression

Published online: 16 September 2016

Rating scales for assessing academic speaking: A data-based definition of progression

Armin Berger, Vienna[*]

This paper summarises a study examining the progression of speaking proficiency as defined by the Austrian English Language Teaching and Testing (ELTT) group in two analytic rating scales for the assessment of academic presentations and interactions. It outlines a multi-method approach to scaling the level descriptors, drawing on both classical test theory and item response theory. First, a *descriptor sorting* procedure involved university teachers ordering the scale descriptors into different levels of proficiency. Second, *descriptor calibration* subjected the sorting task data to multi-faceted Rasch analysis. Finally, in the *descriptor-performance matching* procedure, the teachers related the descriptors to student performances. The synthesis of the results suggested modifications to the original scales. The findings offer a specification of academic speaking, adding concrete details to the reference levels in the Common European Framework. At a more general level, the findings indicate that rating scale validation should be mapped onto theoretical models of performance assessment.

1. Introduction

Although analytic rating scales are widely used in oral performance assessment, empirical scale validation is often lacking. One of the main concerns is that intuitively developed rating scales created on the basis of expert judgement fail to represent the progression of increasing speaking proficiency adequately (Brindley 1998; Hulstijn 2007; Kramsch 1986; Lantolf & Frawley 1985). It is unclear to what extent such rating scales describe an implicational continuum of increasing language proficiency that corresponds to what speakers actually do in real speech, and, accordingly, there is a need for extensive research in educational language testing to demonstrate that the speaking constructs and the way they are operationalised in rating scales are related to the reality of language use (Kaftandjieva &

[*] The author's e-mail for correspondence: armin.berger@univie.ac.at

Takala 2003). The project reported here did exactly that. It examined the operationalisation of a speaking construct[1] in two analytic rating scales[2] for the assessment of academic speaking proficiency.

The research grew out of an ambitious inter-university construct definition and scale development project initiated by the Language Testing Centre at Klagenfurt University in close cooperation with the English Department at the University of Vienna. In an effort to harmonise assessment practices in high-stakes language examinations in Austrian English departments, the English Language Teaching and Testing (ELTT) initiative[3], a working group of applied linguists and language teaching experts from the Universities of Graz, Klagenfurt, Salzburg and Vienna, defined and operationalised a test construct for the certification of speaking proficiency at the end of their BA programmes. The outcome of this project was a set of analytic rating scales: one for the assessment of academic presentations and one for the assessment of discussion-based interactions. The scales encompass descriptors for lexico-grammatical resources and fluency, pronunciation and vocal impact, content and structure (presentation), genre-specific presentation skills, content and relevance (interaction), and interaction skills, covering levels C1 and C2 according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001).

The main purpose of the study was to investigate whether the level descriptors of the two ELTT rating scales actually represent an incremental pattern of increasing speaking proficiency. Although the ELTT group had used a methodologically triangulated approach to scale construction, the scales could not be assumed to form implicational scales of increasing speaking proficiency a priori. Instead, the hierarchy of the scale descriptors had to be established through investigation. To this end, a multi-method approach to scaling the level descriptors was developed. Drawing on both classical test theory and item response theory, the author developed a tripartite research design to relate the scale descriptors to

[1] In language assessment, simply put, the term *construct* refers to one or several traits a test is intended to measure. It is usually defined as an ability or set of abilities that are often not directly measureable but which can be inferred from test performance. Accordingly, a test or assessment instrument is an operationalisation of the construct. For an overview of constructs and construct issues in language assessment, see, for example, Jamieson (2014).

[2] Davies et al. (1999: 153) define a *rating scale* as "a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged". In performance assessment, the scale levels are typically defined by band descriptors representing an underlying scale of increasing language proficiency (Turner 2013: 2). While in a *systemati*c approach to formulating descriptors, *all* performance features are described at *every* level using qualifiers like 'some', 'many' or 'most', in a *salient features* approach concrete behavior characteristic and indicative of a particular level is described only at the level concerned (North 2014: 26). In the latter approach, higher-level descriptors imply lower-level descriptors, and learners at a higher level are considered to have the abilities described at the lower levels as well. In contrast to *holistic* scales, which assign a single rating for overall ability, *analytic* scales assign separate scores on various subscales representing different attributes of the performance.

[3] http://www.uni-klu.ac.at/ltc/inhalt/430.htm

actual speaking performances and to obtain an empirical hierarchy of the performance descriptions. The research can thus be classified as a construct validation study that is methodologically related to the scaling approach generally associated with the work of North (1995; 2000; 2002) and North and Schneider (1998) in connection with developing a common European framework for reporting language competency. The practical aim of the research was to revise the two ELTT scales in the light of the findings.[4]

2. Rating scales in performance assessment

2.1 Models of performance assessment

Assessing foreign language speaking proficiency is a great challenge, because it commonly involves human raters making judgements about the candidates' speaking ability on the basis of observed performances. Unlike receptive skills, which are usually assessed by means of dichotomously scored discrete-point items, speaking skills are normally assessed in a communicative context in which an extended sample of speech is produced by the candidate and judged by one or more raters. It is easy to see how construct-irrelevant factors can impact the judgements and affect the validity and reliability of the ratings. A number of variables, including test-taker, task and rater characteristics, as well as other facets of the test situation, may have an effect on the scores awarded.

A number of performance models have been suggested to investigate possible sources of such variability in speaking assessment (Fulcher 2003; McNamara 1996; Skehan 2001), and in all these performance models, the rating scale represents a crucial facet. Skehan (2001: 168) points out that the test score is directly influenced by the rating procedures and instruments. Not only will the performance be judged by human raters; it will be filtered through the rating instrument (Skehan 2001: 168). Due to these rater and rating scale factors, the score assigned to a particular performance is not a pure index of the candidate's speaking proficiency.

Although the major performance models acknowledge the crucial role that rating scales play in performance assessment, they differ in the degree of importance they place on the rating scale component. McNamara (1996), for example, acknowledges but does not further theorise about the potential impact of rating scales or scale criteria. Skehan (2001), by comparison, is more explicit about the rating scale component, pointing out that the test score is most immediately influenced by the rating procedures. Fulcher (2003) is one of the most comprehensive attempts to consider the influence of rating scales in speaking test performance to date. He places construct definition at the heart of rating scale design and emphasises that the nature of the rating scale, i.e. its orientation, scoring philosophy and

[4] A detailed account of the research has been published in the *Language Testing and Evaluation* series, edited by Rüdiger Grotjahn and Günther Sigott (Berger 2015). The current paper functions as a general outline of the study presented in the book.

focus, has an influence on the score and its meaning. However, what all these models have failed to do is to take adequate account of rating scale validation. While Knoch (2009), in the context of diagnostic writing assessment, included scale *development* methods as an additional variable, no one has modelled the impact of *validation* procedures in performance assessment yet.

In view of these models, one may call for a better understanding of how rating scales function in performance assessment. It is particularly the operationalisations of test constructs in rating scales that warrant investigation. At the same time, the validation procedures themselves and their potential impact on the rating instruments need to be considered more explicitly. The aim of this project was to analyse the operationalisation of the ELTT construct and to determine whether validation methods should be considered a key variable in performance assessment.

2.2 Rating scales as operationalisations of test constructs

The centre of each rating scale is the test construct. Rating scales are often regarded as the "de facto test construct" (Knoch 2011: 81) as they operationalise the construct that cannot be directly measured in itself. Since the rating scale is an operationalisation of the construct, it is important to acknowledge the fundamental role that the construct definition plays in rating scale design (Fulcher 2003: 115). The specific understanding of what is being targeted in the performance of a candidate and the type of inferences one wishes to draw from the scores will inform the content and format of a rating scale.

Operationalisations of test constructs in rating scales have been challenged, however, on several grounds. Most of these criticisms are related to the nature of the rating scale, the underlying concept of communicative competence or language use, and/or the hierarchy of language abilities as established by the descriptors at different levels. Intuitively developed rating scales in particular have been criticized for their a priori status (Chalhoub-Deville 1995; Fulcher 1996; North 1995; Upshur & Turner 1995), their validity being proclaimed by the scale developers or users rather than confirmed by empirical evidence. Furthermore, such scales are often said to be atheoretical in nature (Lantolf & Frawley 1985), to subsume performance features that do not necessarily co-occur in real speech under the same rating categories (Turner & Upshur 2002), to refer to features that bear little relation to the reality of language use (Brindley 1998; North 1995; Turner & Upshur 2002), or to neglect or even run counter to second language acquisition research (North 1995).

One of the major concerns about intuitively developed rating scales is that the underlying assumptions about how proficiency develops could be flawed. Analytic rating scales are composed of a series of hierarchical level descriptions for separate criteria. The implication is that each subscale describes a dimension of growth along which language proficiency increases. In order to improve their proficiency, learners need to go through incremental stages in which they acquire or learn the features characteristic of more difficult speaking activities. However, it is often unclear to what extent rating scales in fact represent an implicational continuum of increasing proficiency.

2.3 Rating scale validation

Research needs to demonstrate that the stages of progression described in rating scales correspond to the reality of language use. The hierarchy of abilities characterising the various scale levels should not be assumed but established through investigation – a desideratum echoed by a number of writers (Kaftandjieva & Takala 2003; Knoch 2009; McKay 2000). Such validation research includes the generation of empirical evidence on how easy or difficult the activities described in the rating scales are. Difficulty values for these activities provide an empirical description of the dimension along which speaking proficiency is expected to grow. Providing empirical evidence that the hierarchy of proficiency descriptions in the ELTT scales corresponds to the reality of language use was indeed one of the key objectives of this study.

Methodologically, rating scale validation is commonly conducted in the context of item response theory, including the one-parameter Rasch model used in this study. One of the main validity concerns in scale development pertains to the question of dimensionality and the scalability of the scale components, i.e. the extent to which the descriptors represent an empirical continuum of increasing language ability rather than a hypothetical progression based on expert intuition (Milanovic et al. 1996; Stansfield & Kenyon 1996; Tyndall & Kenyon 1996). While classical test theory is unable to address this question adequately (Kaftandjieva & Takala 2003), a Rasch-based scaling approach has the capacity to do so (McNamara 1996; McNamara & Knoch 2012). However, many validation studies employ Rasch analysis merely to investigate the *use* of rating scales post hoc, failing to report on the *nature* of the dimension represented in the scale in the first place. The project reported here, in contrast, investigated the proficiency continuum underlying the ELTT scales prior to operational use.

Similar to the scaling project that generated the illustrative CEFR descriptors (North 2000), the current study calibrated descriptors of speaking proficiency onto a unidimensional measurement scale so as to ascertain whether they adequately describe a continuum of speaking proficiency. The present project is conceptually different, however, in that it analysed a speaking construct operationalised by a group of experts in a customised scale development project geared towards the specific needs of tertiary speaking assessment. Unlike North (2000), who provided a general description of the full proficiency range from the lowest level of generative language use to mastery, drawing on descriptors from various existing scales, this study validated a pre-defined continuum of academic speaking proficiency covering the highest levels only.

3. Research questions

The following research questions were addressed:

1. To what extent do the descriptors of the ELTT speaking scales (*presentation* and *interaction*) define a continuum of increasing speaking proficiency?

   a. To what extent does the ELTT speaking construct represent a single (psychometric) dimension?
   b. What does the empirical hierarchy of the ELTT scale descriptors look like?
   c. Are there any natural gaps and groupings on the vertical scale of descriptors that would suggest cut-off points to allow equidistant bands?
2. Does the validation methodology have an effect on the hierarchy of descriptor units?
3. Which rating scale descriptors are the most effective ones?

The basic hypotheses underlying the research were that (1) while most ELTT rating scale descriptors were expected to compose an adequate continuum of increasing speaking proficiency, some would prove dysfunctional, and (2) while the empirical order of most ELTT descriptor units was expected to correspond to the intended order as conceived of by the scale developers, some would turn up at different points along the continuum.

4. Research design and methodology

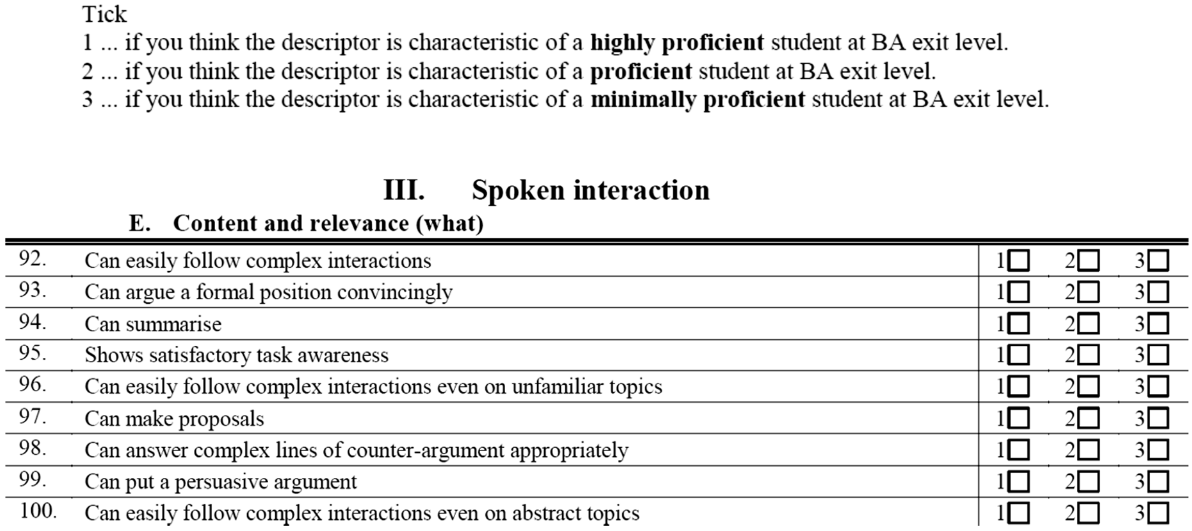
To answer the research questions, a multi-method design involving three phases was developed. While the first phase was part of a preliminary study (Berger 2012), the second and third phases were the focus of the study outlined here. The rationale for the multi-method approach was that a range of procedures can balance out method effects and thus produce more stable results. A synthesis of findings from different sources was considered to have the potential to provide more solid evidence.

4.1 Descriptor sorting

In the first phase, termed *descriptor sorting*, experienced language teachers unfamiliar with the ELTT scales carried out a descriptor sorting task with the main intention to order the descriptor components according to the perceived level of proficiency required to perform the activities described. The assumption was that if experts were able to categorise the descriptor units consistently, this could be interpreted as evidence of an underlying continuum of language proficiency.

A total of 21 language teachers from all five Austrian university English departments (Graz, Innsbruck, Klagenfurt, Salzburg and Vienna) took part in the study. They were all qualified teachers of English as a foreign language at different seniority levels, with at least three years' experience of teaching speaking classes at university. About half of them were native speakers of English and the others were native speakers of German. None of them had been involved in the scale construction process or seen the scales before.

The main instrument used in this part of the study was a sorting task questionnaire, illustrated in Figure 1.



Tick
1 ... if you think the descriptor is characteristic of a **highly proficient** student at BA exit level.
2 ... if you think the descriptor is characteristic of a **proficient** student at BA exit level.
3 ... if you think the descriptor is characteristic of a **minimally proficient** student at BA exit level.

III. Spoken interaction

E. Content and relevance (what)

92.	Can easily follow complex interactions	1□	2□	3□
93.	Can argue a formal position convincingly	1□	2□	3□
94.	Can summarise	1□	2□	3□
95.	Shows satisfactory task awareness	1□	2□	3□
96.	Can easily follow complex interactions even on unfamiliar topics	1□	2□	3□
97.	Can make proposals	1□	2□	3□
98.	Can answer complex lines of counter-argument appropriately	1□	2□	3□
99.	Can put a persuasive argument	1□	2□	3□
100.	Can easily follow complex interactions even on abstract topics	1□	2□	3□

Figure 1: Extract from the sorting task questionnaire

For this questionnaire, the level descriptions of the two ELTT scales were divided into a total of 174 minimally meaningful descriptor units. For example, the descriptor *can easily keep up with the discussion and argue a formal position convincingly, responding to questions and comments and answering complex lines of argument fluently, spontaneously and appropriately* was divided into seven independent descriptor units:

- Can easily keep up with the discussion
- Can argue a formal position convincingly
- Can respond to questions
- Can respond to comments
- Can answer complex lines of argument fluently
- Can answer complex lines of argument spontaneously
- Can answer complex lines of argument appropriately

These descriptor units were listed vertically in a random order on the questionnaire. To construct a grid for the teachers to indicate the perceived level of student proficiency for every descriptor unit, three categories of proficiency were added horizontally.[5]

In the analysis, both consistency and agreement indices were calculated. While the former provided information about the relative ranking or ordering of the descriptor units, the latter revealed the extent to which raters assigned exactly the same ratings.[6]

[5] In the original ELTT scales, only three band levels – one, three and five – had been defined by descriptors, while the two intermediate levels – two and four – had been left unworded. The purpose of the sorting task was to see if experts were able to reconstruct the original hierarchy of the descriptors, hence three proficiency categories in the questionnaire.

[6] Considering that there is no one best statistical index for the purpose of assessing inter-rater reliability, it is recommended to compare different indices and interpret them in the light of their limitations (Abedi et al. 1995; Kaftandjieva & Takala 2002). The consistency indices calculated for this study were Cronbach's alpha, the

The problem with this standard approach to rater reliability was that the consistency and agreement indices within classical test theory were unable to deal with the issue of rater variability adequately. The fundamental problem, sometimes referred to as the "agreement-accuracy paradox" (Eckes 2011: 29), is the possibility that a high degree of consistency or agreement does not automatically mean a high degree of accuracy of the ratings. For example, if raters display the same type of bias consistently, such as an unusual degree of severity or leniency, the ratings are obviously not accurate, but the reliability indices will still be high. Similarly, raters may award exactly the same scores, i.e. show perfect agreement, but may be using the scales incorrectly. Therefore, high reliability indices may lead to the erroneous conclusion that the ratings are accurate when in actual fact they are not. This serious limitation of the sorting task methodology required a follow-up investigation employing a measurement approach.

4.2 Descriptor calibration

In phase two, termed *descriptor calibration*, a multi-faceted Rasch analysis of the data from the sorting task was carried out to complement the standard approach adopted in the previous phase. This measurement approach had two advantages over the standard approach. Firstly, it was able to provide difficulty estimates for the descriptor units and thus display more accurately the hierarchical relationships between the units. Secondly, it was able to take adequate account of different facets of the test situation, including, most notably, rater variability. Unlike the standard approach, the multi-faceted Rasch analysis generated an in-depth account of the differences and similarities between the raters' judgements, and the impact of rater characteristics or other facets of the test situation on estimates of descriptor difficulty.

The multi-faceted Rasch model (Linacre 1989) is a significant development in the family of probability-based item response theory models, working on the assumption that ordinal observations are the outcome of an interaction between different elements, such as a candidate, an item and a rater. For the validation of the ELTT scales, this model was particularly useful in that it enabled the mapping of rating scale descriptors onto a mathematical scale. That is, the analysis yielded difficulty values for individual descriptor units, which is similar to an item-banking methodology, except that items are descriptor units judged by expert raters rather than test questions answered by candidates. The software used to compute multi-faceted Rasch statistics was FACETS (Linacre 2013a).[7]

While the descriptor calibration methodology reduced the impact of some of the limitations of the descriptor sorting approach, other limitations remained. In particular, the data still reflected what university teachers *believed* about speaking proficiency rather than

Pearson product moment correlation coefficient, Spearman's rank coefficient, Kendall's tau coefficient, and Kendall's coefficient of concordance. The agreement indices included the Exact Observed Agreement statistic (Linacre 2013b), the Rater Agreement Index (Burry-Stock et al. 1996), and Fleiss' kappa.

[7] For an accessible introduction to multi-faceted Rasch measurement, see, for example, Eckes (2015).

what speakers actually do in real speech. In the worst case, this would have meant that the results of the two phases described so far revealed a picture of erroneous rater beliefs divorced from real speaking proficiency. This important limitation warranted another research phase, designed to establish a link between expert knowledge and actual language use.

4.3 Descriptor-performance matching

In the third and most comprehensive phase, a smaller group of teachers were asked to match descriptor units with actual speaking performances, hence the term *descriptor-performance matching*. While the first two phases were based on decontextualised, abstract conceptualisations of student proficiency dissociated from real performances, the third one connected the construct descriptions to samples of real speech. The main aim of this phase was to elicit the teachers' perceptions of the extent to which the rating scale descriptors represented a particular student performance. It was reasoned that if a descriptor-performance matching task of this kind were able to generate another set of meaningfully calibrated descriptors reflecting the intended hierarchy, this could be interpreted as further evidence of an underlying continuum of language proficiency.

Three groups of participants were involved in this part of the study. Firstly, 16 undergraduate students at the English Department at the University of Vienna conducted a trial run of a role play task to check whether it elicited oral interaction effectively. Secondly, 78 undergraduate students of English from all five Austrian English departments agreed to take part in a mock speaking test involving a presentation and the role play task. All of them were non-native speakers of English, mostly with a German-speaking background, and in the final semesters of their BA programme or at an equivalent stage in their teaching degree programme. Finally, eight language teaching experts, six from the University of Vienna and one each from the University of Innsbruck and the University of Klagenfurt, took part in the descriptor-performance matching activity. They were all qualified teachers of English as a foreign language, some with postgraduate qualifications. Six of them were native speakers of German and two were native speakers of English.

Two types of data were collected in this phase: performance data and rating data. As regards the former, it was decided to organise mock exams without any raters present rather than record performances in live exams. One reason was that there were no common oral exams across Austrian degree programmes that would have met the specifications for this research project. Another reason was that a number of contextual and task-related variables, such as timing and setting, had to be controlled for, which would have been impossible in real exam situations. In the first part of the mock exam, the participating students were required to give a five-minute academic presentation on a topic of their choice supported by visuals. In the second part, the students interacted in groups of four. The interaction task took the form of an unrehearsed role play discussion of a controversial topic, in which the candidates adopted a given position and pursued specific communicative goals. A total of 153 student performances were videorecorded, including 75 presentations and 78 interactive performances. The videos were made available to the raters via an internet platform, and the raters did the rating task online.

The instrument designed to collect the rating data was similar to the behavioural observation scales used for the evaluation of work performance (Latham & Wexley 1977). Like in the sorting task questionnaire, here too the descriptor units were listed vertically in a random order. This time, however, a five-point scale was attached horizontally for the judges to indicate their perception of how accurately each descriptor unit represented the observed proficiency level of the candidates. The purpose of this instrument was to generate data that allowed me to conduct another multi-faceted Rasch analysis and calibrate the descriptor units onto an arithmetic scale. In addition, a manual with instructions on how to match the descriptors with performances and how to complete the rater questionnaires was provided for the raters.

The performances were distributed according to a rating plan that ensured sufficient overlap between the ratings for the FACETS analysis to be meaningful. Since a fully crossed rating design, in which all raters match up all performances with all descriptor units, would have been highly impractical, a less data-intense judging plan was created. This rating plan included a reduced set of overlapping observations, which ensured sufficient connectedness between raters, performances and descriptor units. This procedure yielded a total of 21,909 valid data points for the multi-faceted Rasch analysis.[8]

5. Interpretation and synthesis of the findings

The procedures outlined above yielded different sets of descriptor unit hierarchies, each one in itself defining a continuum of increasing speaking proficiency. After checking the psychometric dimensionality of the descriptor units, I was in a position to interpret the meaningfulness of the hierarchies. Special attention was paid to surprising calibrations, which were odd in the sense that the estimated difficulty appeared counter-intuitive or contrary to current theory. Such units possibly describe something that is not quite in line with the main construct. At the same time, the analysis produced isolated calibrations that deviated considerably from the intended position along the continuum, but on reconsideration seemed to be meaningful, indicating that the scale developers may have misconceptualised the language activities represented by these units.

The next step was to set cut-off points and divide the calibrated lists of descriptor units into five roughly equal-interval band levels. Once the cut-off points had been set, the content of the resulting levels was checked. To this end, the calibrated descriptor units were arranged in tabular form, grouped according to content traits and listed in descending order of their difficulty values. The resulting charts facilitated the interpretation of the progression and the identification of gaps along the continuum. The focus was on the content integrity of the levels, that is, the extent to which the levels described proficiency in a consistent and coherent way, both vertically along the proficiency continuum and horizontally across

[8] For further details on the research design and methodology, see Berger (2015: 129-261)

different construct traits. At the same time, the vertical progression of the units was checked to see if they advanced meaningfully from one level to the next one.

The final step was to synthesise the findings from all three procedures in order to establish the most effective set of descriptor units and reintegrate them into the revised versions of the ELTT scales. To this end, a systematic approach to evaluating the overall quality of the descriptor units was developed. Four quality criteria were deemed relevant to the purpose at hand, including – in descending order of importance – the soundness of the calibrations, statistical model fit, consistency across procedures in terms of level allocation, and congruence with the scale developers' original intention. Each descriptor unit was evaluated according to these criteria. Depending on the extent to which a descriptor unit satisfied the quality criteria, it was classified as either *excellent*, *good*, *adequate*, *poor* or *problematic*. This evaluation procedure identified the most effective descriptor units in a way that reflected the weighted importance of the quality criteria and determined which units were reintegrated into coherent level descriptions. While *excellent* and *good* units were included in the final set exactly as they were, *adequate* and *poor* units were included with modifications – the latter, however, only in exceptional cases if dropping them would have caused serious construct underrepresentation in the final scales. *Problematic* units were removed altogether. This process resulted in the modified versions of the two ELTT scales.[9]

6. Main results

In answer to the central research question, the findings suggested that, on the whole, the ELTT level descriptions represent an adequate continuum of increasing speaking proficiency. This was evidenced to varying degrees by the results from the three procedures. In the descriptor sorting part, overall, the correlational analyses showed moderate consistency and agreement among the raters and a fair amount of match between the ratings and the original order of the ELTT descriptor units. In the descriptor calibration and performance-matching procedures, multi-faceted Rasch analysis showed a good amount of fit to the psychometric model, indicating that the descriptors can be calibrated meaningfully onto a common scale. In response to research question 1a (*To what extent does the ELTT speaking construct represent a single [psychometric] dimension?*), it was therefore concluded that the ELTT speaking construct is sufficiently unidimensional in a psychometric sense. This does not mean, however, that all descriptor units met the model's expectations. It was particularly the task-related descriptor units referring, for example, to the content of presentations, genre-specific presentation skills, and interaction skills, which deviated from the unidimensional ideal, but not to an extent that would suggest that the construct is psychometrically multidimensional.

With regard to the empirical hierarchy of the ELTT descriptors, the findings suggested a largely systematic progression. The two Rasch-based procedures yielded hierarchies of performance descriptions that were clearly meaningful and generally in line with the scale

[9] For the modified scales, see Berger (2015: 294-297).

developers' intentions. Some isolated cases appeared calibrated at unexpected yet meaningful positions along the proficiency continuum, pointing to misconceptions on part of the scale developers. The findings showed that, in the most general terms, speaking proficiency increases along a continuum from the ability to use a broad range of language with a relatively high degree of accuracy and the ability to apply a number of basic presentation and interaction strategies appropriately on to the ability to use a very large range of (complex) language with an ever-increasing degree of precision and ease and the ability to use presentation and interaction strategies with great flexibility and for maximum effect. Thus, with regard to research question 1b (*What does the empirical hierarchy of the ELTT scale descriptors look like?*), the results revealed that the proficiency progression was meaningful, with only a few isolated descriptors that appeared at unexpected positions.

Groupings of descriptor units and gaps along the calibrated scale became apparent when the units were arranged into trait categories and the categories inspected separately. The expectation was that related descriptor units would cluster around specific points on the scale while at the same time leaving gaps between adjacent groupings. Such gaps usually appeared at roughly equal intervals, which supported the division of the scales into five equidistant bands. Research question 1c (*Are there any natural gaps and groupings on the vertical scale of descriptors that would suggest cut-off points to allow equidistant bands?*) was therefore answered in the affirmative. The gaps and groupings within individual trait categories were clear enough to suggest five distinct proficiency levels.

Concerning the validation methodology, one must acknowledge the different nature of the data and the statistical paradigms used in the three phases. The descriptor sorting procedure analysed experiential teacher judgements within the realm of classical test theory, the descriptor calibration subjected these judgements to probability-based multi-faceted Rasch analysis, and the descriptor-performance matching procedure Rasch-analysed teacher judgements based on real performance data. The descriptor sorting resulted in two provisional three-band scales containing those descriptors that reached the highest frequency figures in the sorting task. The descriptor calibration and the descriptor-performance matching procedures, in comparison, yielded two five-band scales of calibrated descriptor units with sample-free difficulty estimates. Dividing these two scales into five approximately equidistant band levels resulted in similar yet different level descriptions. The comparison of the two Rasch-based procedures revealed a strong positive correlation between the two sets of difficulty estimates, but not an equally close correspondence between the level allocations produced by the two procedures. In conclusion, research question two (*Does the validation methodology have an effect on the hierarchy of descriptor units?*) was answered in the affirmative. The validation methodology was not a neutral factor but had a clear effect on the resulting rating instruments. It was therefore suggested that the different procedures should be considered complementary rather than alternative validation methods.

Finally, the analysis showed that the original versions of the ELTT scales were not ready for operational use yet. Instead, problematic descriptor units had to be identified and eliminated. Only the psychometrically most stable descriptor units with consistent level allocations were deemed productive in an assessment situation. With regard to research question three (*Which rating scale descriptors are the most effective ones?*), it was concluded that the most effective descriptor units were characterised by sound calibrations, good fit to the psychometric model, consistent level allocations, and congruence with the scale developers' original intentions. In this view, descriptor units referring to academic presentations generally turned out to be more effective than those referring to interactions. Furthermore, the chances of a descriptor being effective were highest when it related to linguistic rather than task-related aspects of the construct, when it was formulated in the indicative mood, and when it closely resembled the descriptor formulations in the CEFR.

7. Discussion

The findings from this study make several contributions to the description of academic speaking proficiency, the conceptualisation of performance assessment, and models of rating scale development. Most significantly for the field of foreign language teaching, learning and testing, the study represents a major step towards specifying academic speaking proficiency. The ELTT descriptors for lexico-grammatical resources and fluency, pronunciation and vocal impact, content and structure, presentation skills, content and relevance, and interaction skills were plotted on a vertical scale with different difficulty values, thus providing an empirical description of a proficiency continuum. After this continuum had been divided into five approximately equal intervals, the performance descriptions at every level turned out to be remarkably coherent across the different scale categories. It was therefore possible to abstract common reference points from the specific descriptors and describe the salient characteristics at every level. These characteristics were subsumed under the labels 'Effective Operational', 'Full Operational', 'General Academic', 'Advanced Academic', and 'Full Academic Proficiency'.[10]

Effective Operational Proficiency is the level at which speakers can use the language with sufficient control to participate productively in formal university settings. Their linguistic repertoire is large enough to express conceptually complex ideas clearly and appropriately. Essential academic functions in presentations and interactions can be performed appropriately, with occasional disfluencies due to linguistic planning and repair.

Full Operational Proficiency is the level at which speakers can use language entirely appropriately in formal university settings. The progression is characterised by the growing control and sophistication of the linguistic resources. Speakers have access to a larger range of language that enables them to reduce the number of errors and pauses for linguistic planning and repair. At the same time, they show a new degree of awareness of contextual factors such as audience and task requirements.

General Academic Proficiency is where speakers can use language appropriately and effectively for most academic purposes. This level is characterised by a new quality dimension, both linguistically and strategically. Disfluencies mainly occur on account of

[10] The interested reader is referred to Berger (2015) for the complete list of ELTT descriptors.

linguistic refinement as opposed to planning and repair. There is a new focus on cognitive skills integral to academic practice, including argumentation and reasoning, and the speakers' attentional focus seems to shift from communicative appropriateness to strategic effectiveness. Attributes of professional public speakers begin to feature at this level. Regarding interaction skills, there is a new degree of adaptive responsiveness; speakers can react to unanticipated circumstances by discerning contextual factors and adapting their contributions accordingly.

Advanced Academic Proficiency represents the stage at which learners can use language fluently and accurately on all levels pertinent to academic presentations and interactions. They have access to a very broad linguistic repertoire, which enables them not only to speak without constraints that would restrict the effectiveness of the communication but also with greater efficiency and ease. Speakers can communicate naturally, fluently and accurately even when their attention is otherwise engaged. There is an unprecedented degree of meta-cognitive awareness that enables speakers to perform their tasks more consciously, while at the same time monitoring and optimising their performance. In functional terms, there is a greater capacity for persuasion as opposed to mere argumentation and for interaction management as opposed to mere participation.

Full Academic Proficiency reflects the highest degree of control and naturalness in all respects. The speaking process has become completely automatic so that the full attentional capacity is available for the communicative effect.

These common reference points are significant for the field of foreign language teaching, learning and testing in at least two major respects. Firstly, they help to flesh out the notoriously vague proficiency descriptions of the CEFR at levels C1 and C2, and provide finer, meaningful and concrete level distinctions within the higher proficiency range. While the CEFR (Council of Europe 2001: 35-36) defines the fixed common reference points at the C1 and C2 levels as "good access to a broad range of language, which allows fluent, spontaneous communication" and a "new degree of discourse competence", which shows itself in an ever more fluent use of co-operating strategies, coherence/cohesion and negotiation (C1), and a "degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners" (C2), the study outlined here provided a more fine-grained distinction for the upper proficiency range. As such, these reference points provide a sound basis for the development of course curricula, the formulation of teaching and learning objectives, materials design, and other educational purposes at tertiary level. Secondly, the common reference points also provide useful criterion statements for future construct definition projects. If the proficiency continuum is to be refined further, such benchmark statements can serve as a starting point in an endeavour to add new aspects of speaking proficiency in a transparent and coherent way.

In terms of the conceptualisation of performance assessment, the study has shed light on the question of whether rating scale validation procedures should be considered a key constituent of performance assessment. All current theories agree that rating scales are not neutral factors in the rating process but may exert a considerable influence on the scores awarded, yet no model to date has taken scale validation procedures into consideration. However, the results of this research suggest that validation procedures should be mapped onto theoretical models of performance assessment as they potentially play a major part in shaping the rating instrument. Figure 2 shows an extended model of performance assessment based on Fulcher (2003) and Knoch (2009) with the express purpose of factoring in validation procedures. The major expansions and additional relationships are indicated by dark grey shading and dashed arrows, respectively.

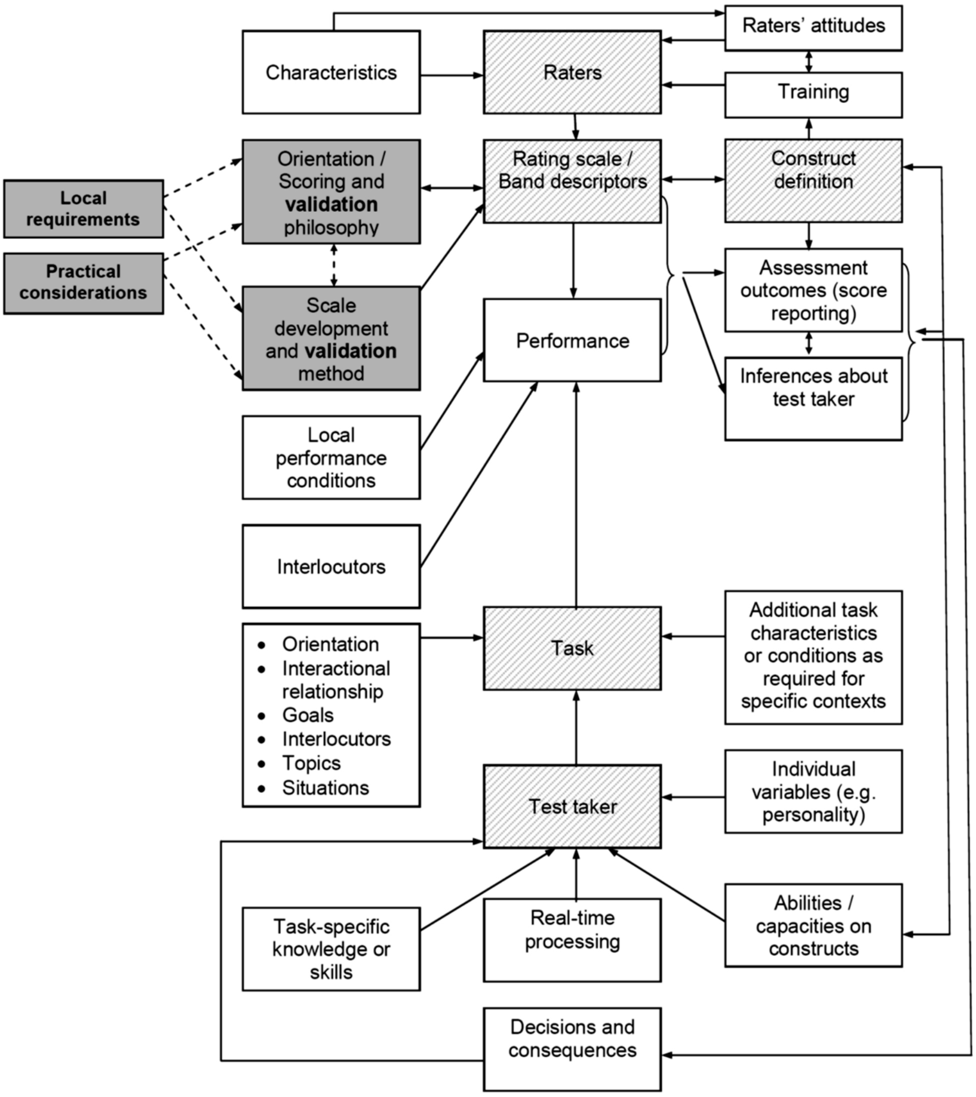

Raters' attitudes
Characteristics
Raters
Training
Local requirements
Practical considerations
Orientation / Scoring and validation philosophy
Rating scale / Band descriptors
Construct definition
Scale development and validation method
Performance
Assessment outcomes (score reporting)
Inferences about test taker
Local performance conditions
Interlocutors
Task
Additional task characteristics or conditions as required for specific contexts
Orientation
Interactional relationship
Goals
Interlocutors
Topics
Situations
Test taker
Individual variables (e.g. personality)
Task-specific knowledge or skills
Real-time processing
Abilities / capacities on constructs
Decisions and consequences

Figure 2: An expanded model of performance assessment

Due to the potential impact that validation methods may have on the final rating instrument, in practice, different scale validation methods should be used in combination rather than isolation whenever possible, and the utility of the methods should not be evaluated independently but in terms of their combined effect.

The third major contribution concerns rating scale development. The study has underlined the complexity of rating scale development and made a strong case for the integration of intuitive and data-driven approaches into the design process. Galaczi et al.'s (2011) model for the development of assessment scales, which presents the process along a chronological and a methodological dimension, was considered to be a helpful starting point. At the same time, this model turned out to be too linear in the conceptualisation of the development process and of limited practical use for small-scale projects, which are usually run within severe financial and practical constraints. It is particularly the methodological dimension of the model that wants wider applicability. It was therefore concluded that theoretical models for the development of rating scales need to be flexible enough to be useful in contexts beyond large-scale high-stakes testing. Accordingly, Galaczi et al.'s (2011) two-dimensional grid was converted into a three-dimensional model space unfolding between a chronological, a functional and a methodological dimension. The additional dimension should make it easier for scale developers to distinguish more explicitly between intuitive, data-based, qualitative and quantitative methods for research, development and consultation purposes.

Finally, the project resulted in a set of practical recommendations for descriptor formulation. In addition to the criteria set out in the CEFR (Council of Europe 2001: 205-207), including positiveness, definiteness, clarity, brevity and independence, all of which proved to be highly relevant quality characteristics in the study at hand, three further criteria were suggested in the light of the findings. Firstly, *uniqueness* refers to the principle that descriptors should be unique to a particular level rather than appear at several levels in the same scale. Multi-level descriptors, which occur in exactly the same form at several levels, are not productive in an assessment situation. This is in line with what North (2014) termed the 'salient features' approach to formulating descriptors. The basic assumptions of this approach are that the salient performance features are mentioned at whichever level they start to be new and criterial, that these features apply to speakers at the level concerned as well as to speakers at a higher level, and that repeating a descriptor at levels above the one concerned is therefore unnecessary. Secondly, the criterion of *directness* refers to the observation that descriptors tended to function better if they were formulated as *do* statements as opposed to *can-do* statements. Descriptors may be more readily interpretable if they refer directly to some observable behaviour as the manifestation of an ability rather than the underlying ability itself. And thirdly, the criterion of *specificity* means that scale descriptors should be as specific as possible in terms of both the language activity described and the speakers' degree of skill in performing this activity. If a descriptor mentions a language activity alone without referring to the circumstances or conditions under which speakers can perform it, or the effect produced by performing that activity, it may well be too vague to be productive in an assessment situation.

7. Conclusion

The project reported here is significant in a number of ways. It continued the pioneering work in language testing at Austrian universities initiated by the ELTT group. Never before have experts from all five Austrian English departments collaborated in a language testing project. Furthermore, it produced extensive empirical evidence about the validity of a language testing instrument. Never before has a language assessment scale been so thoroughly researched in the Austrian university context. Thus, the study makes an important contribution to the professionalisation of assessment practices in Austrian university language programmes.

The project is relevant also beyond the national context, not least because it addresses speaking assessment as a research area that is relatively underrepresented in the validation literature. The significance of the research for the wider field of foreign language teaching, learning and testing lies in its contribution towards specifying academic speaking proficiency. The study has demonstrated how a local rating scale development project can enhance our understanding of academic speaking and the progression between levels C1 and C2. While the CEFR descriptions of speaking ability at levels C1 and C2 are notoriously vague and underspecified in many respects, this study has provided more fine-grained subdivisions of the upper proficiency range. Those ELTT scale descriptors that turned out to be the psychometrically most stable ones can complement the CEFR descriptors, and the salient characteristics of the scale levels offer a specification of academic speaking, adding concrete details to the higher reference levels. Such finer distinctions satisfy the pedagogical need in many local tertiary contexts to capture and report even small gains in language proficiency. They can also serve as useful benchmarks or criterion statements for future construct definition projects. Thus, the potential beneficiaries include both researches in the field of foreign language teaching, learning and testing and practitioners in tertiary language education who are engaged in the task of adjusting common frameworks for language teaching and learning to the needs of the local context.

Furthermore, the specific research design developed in this study allowed conclusions about the applicability and usefulness of the scale construction and validation methods, particularly for smaller projects in local university contexts. Two theoretical models resulted from the findings. Firstly, an extended model of performance assessment was proposed. While earlier models featured only scale *development* methods as key variables, the expanded model also factors scale *validation* procedures in to the assessment process. Secondly, a three-dimensional model for rating scale development was created, which in contrast to existing ones is more precise and flexible enough to be applied to local, small-scale projects.

Ultimately, it is hoped that the project can go some way towards promoting the idea that valid language testing and assessment is both a pedagogical desideratum and a professional responsibility.

References

Abedi, Jamal; Baker, Eva; Herl, Howard. 1995. *Comparing reliability indices obtained by different approaches for performance assessments*. Los Angeles: CRESST. http://www.cse.ucla.edu/products/reports/TECH401.pdf (2 Aug. 2016).

Berger, Armin. 2012. "Validating analytic rating scales: Investigating descriptors for assessing academic speaking at C1 and above". MA thesis, Klagenfurt University.

Berger, Armin. 2015. *Validating analytic rating scales: A multi-method approach to scaling descriptors for assessing academic speaking*. (Language Testing and Evaluation vol. 37). Frankfurt am Main: Peter Lang.

Brindley, Geoff. 1998. "Describing language development? Rating scales and SLA". In Bachman, Lyle; Cohen, Andrew (eds.). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press, 112-140.

Burry-Stock, Judith; Shaw, Dale; Laurie, Cecelia; Chissom, Brad. 1996. "Rater agreement indexes for performance assessment". *Educational and Psychological Measurement* 56(2), 251-262.

Chalhoub-Deville, Micheline. 1995. "Deriving oral assessment scales across different tests and rater groups". *Language Testing* 12(1), 16-33.

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

Davies, Alan; Brown, Annie; Elder, Cathie; Hill, Kathryn; Lumley, Tom; McNamara, Tim. 1999. *Dictionary of language testing.* Cambridge: Cambridge University Press.

Eckes, Thomas. 2015 [2011]. *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Fulcher, Glenn. 1996. "Does thick description lead to smart tests? A data-based approach to rating scale construction". *Language Testing* 13(2), 208-238.

Fulcher, Glenn. 2003. *Testing second language speaking*. London: Pearson Longman.

Galaczi, Evelina; ffrench, Angela; Hubbard, Chris; Green, Anthony. 2011. "Developing assessment scales for large-scale speaking tests: A multiple-method approach". *Assessment in Education: Principles, Policy & Practice* 18(3), 217-237.

Hulstijn, Jan. 2007. "The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency". *The Modern Language Journal* 91(4): 663-667.

Jamieson, Joan. 2014. "Defining constructs and assessment design". In Kunnan, Antony (ed.). *The companion to language assessment: Abilities, contexts, and learners*. Chichester: Wiley Blackwell, 771-787.

Kaftandjieva, Felianka; Takala, Sauli. 2002. "Council of Europe scales of language proficiency: A validation study". In Alderson, Charles (ed.). *Common European framework of reference for languages: Learning, teaching, assessment: Case studies*. Strasbourg: Council of Europe, 106-129.

Kaftandjieva, Felianka; Takala, Sauli. 2003. "Development and validation of scales of language proficiency". In Vagle, Wenche (ed.). *Vurdering av språkferdighet*. Trondheim: Institutt for språk- og kommunikasjonsstudier, 31-38.

Knoch, Ute. 2009. *Diagnostic writing assessment: Developing and validating a rating scale for diagnostic writing assessment*. Frankfurt am Main: Peter Lang.

Knoch, Ute. 2011. “Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?”. *Assessing Writing* 16(2), 81-96.

Kramsch, Claire. 1986. “From language proficiency to interactional competence”. *Modern Language Journal* 70(4), 366-372.

Lantolf, James; Frawley, William. 1985. “Oral-proficiency testing: A critical analysis”. *The Modern Language Journal* 69(4), 337-345.

Latham, Gary; Wexley, Kenneth. 1977. “Behavioural observation scales for performance appraisal purposes”. *Personnel Psychology* 30(2), 255-268.

Linacre, John M. 1989. *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, John M. 2013a. *Facets computer program for many-facet Rasch measurement*. (Version 3.71.2). [Computer Program]. http://www.winsteps.com (23 Oct. 2013).

Linacre, John M. 2013b. *A user’s guide to FACETS: Program manual 3.71.2*. http://www.winsteps.com/a/facets-manual.pdf (23 Oct. 2013).

McKay, Penny. 2000. "On ESL standards for school-age learners." *Language Testing* 17(2), 185-214.

McNamara, Tim. 1996. *Measuring second language performance*. London: Longman.

McNamara, Tim; Ute Knoch. 2012. “The Rasch wars: The emergence of Rasch measurement in language testing.” *Language Testing* 29(4), 555-576.

Milanovic, Michael; Saville, Nick; Pollitt, Alastair; Cook, Anette. 1996. “Developing rating scales for CASE: Theoretical concerns and analyses”. In Cumming, Alister; Berwick, Richard (eds.). *Validation in language testing*. Clevedon: Multilingual Matters, 15-38.

North, Brian. 1995. “The development of a common framework scale of descriptors of language proficiency based on a theory of measurement”. *System* 23(4), 445-465.

North, Brian. 2000. *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, Brian. 2002. “Developing descriptor scales of language proficiency for the CEF common reference levels”. In Alderson, Charles (ed.). *Common European framework of reference for languages: Learning, teaching, assessment: Case studies*. Strasbourg: Council of Europe, 87-105.

North, Brian. 2014. *The CEFR in practice*. Cambridge: Cambridge University Press.

North, Brian; Schneider, Günther. 1998. “Scaling descriptors for language proficiency scales”. *Language Testing* 15(2), 217-262.

Skehan, Peter. 2001. “Tasks and language performance assessment”. In Bygate, Martin; Skehan, Peter; Swain, Merrill (eds.). *Researching pedagogic tasks: Second language learning, teaching and testing*. London: Longman, 167-185.

Stansfield, Charles; Kenyon, Dorry. 1996. “Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines”. In Cumming, Alister; Berwick, Richard (eds.). *Validation in language testing*. Clevedon: Multilingual Matters, 124-153.

Turner, Carolyn. 2013. “Rating scales for language tests”. In Chapelle, Carol (ed.). *The Encyclopedia of Applied Linguistics*. Malden, Mass.: Blackwell Publishing Ltd. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1045/pdf (2 Aug. 2016).

Turner, Carolyn; Upshur, John. 2002. “Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores”. *TESOL Quarterly* 36(1), 49-70.

Tyndall, Belle; Kenyon, Dorry. 1996. “Validation of a new holistic rating scale using Rasch multi-faceted analysis”. In Cumming, Alister; Berwick, Richard (eds.). *Validation in language testing*. Clevedon: Multilingual Matters, 39-57.

Upshur, John; Turner, Carolyn. 1995. “Constructing rating scales for second language tests”. *English Language Teaching Journal* 49(1), 3-12.

How to contact VIEWS:

VIEWS c/o
Department of English, University of Vienna
Spitalgasse 2-4, Hof 8.3
1090 Wien
AUSTRIA

fax	**+ 43 1 4277 9424**
e-mail	**views.anglistik@univie.ac.at**
W3	**http://anglistik.univie.ac.at/views/** **(all issues available online)**

IMPRESSUM:
EIGENTÜMER, HERAUSGEBER & VERLEGER: VIEWS, C/O INSTITUT FÜR ANGLISTIK & AMERIKANISTIK DER UNIVERSITÄT WIEN, UNIVERSITÄTSCAMPUS AAKH, SPITALGASSE 2, 1090 WIEN, AUSTRIA. **FÜR DEN INHALT VERANTWORTLICH**: ELNORA TEN WOLDE, MARLENE SCHWARZ, KLAUS HOFMANN. **REDAKTION:** ANDREAS BAUMANN, MONIKA BONIECKI, CHRISTIANE DALTON-PUFFER, HELEN HEANEY, GUNTHER KALTENBÖCK, EVELIEN KEIZER, CHRISTINA PRÖMER, ANGELIKA RIEDER-BÜHNEMANN, NIKOLAUS RITT, BARBARA SEIDLHOFER, JULIA SKALA, BARBARA SOUKUP, UTE SMIT, STEFAN THIM, VERONIKA THIR, HENRY G. WIDDOWSON, EVA ZEHENTNER. **ALLE:** C/O INSTITUT FÜR ANGLISTIK & AMERIKANISTIK DER UNIVERSITÄT WIEN, UNIVERSITÄTSCAMPUS AAKH, SPITALGASSE 2, A - 1090 WIEN. **HERSTELLUNG:** VIEWS