IN-HOUSE RATING SCALE DEVELOPMENT

Simon Dawson

ABSTRACT
The rating scale is a central part of the assessment of spoken and written performance. Publicly available scales often do not fit the needs of specific assessments and therefore a unique scale is often created from scratch. The development of a valid and reliable rating scale is time-consuming, especially if it is done without experience or guidance. This paper attempts to provide guidance to practitioners on rating scale development. To do this, the paper gives an account of an attempt to revise a rating scale for use in the assessment of oral presentations. Using action research, the study focuses on the lessons that were learnt and ends with a set of recommendations.

摘要
等级量表是评估口语和写作表现的核心组成部分。然而，可获取的公开量表往往不能满足特定评估的需要，因此，经常需要从头建立量表。 开发一个有效且可靠的定级量表很耗时，尤其是在没有经验或指导的情况下。本文旨在为等级量表开发者提供指导。基于此目的，文章详述了一个针对口头报告测试而修改一个定级量表的尝试。使用行动研究方法，本研究聚焦所取得的经验和教训，并在结尾部分提出了一组建议。

INTRODUCTION

A sound rating scale is an essential part of a reliable and valid assessment of spoken and written language; it clearly sets out the construct being measured, provides accurate descriptions of the typical performance expected at different levels, and guides raters in making consistent judgements that are in line with an agreed standard. For these reasons a great deal of time and money is invested in the development of rating scales (e.g. Galaczi, ffrench, Hubbard, & Green, 2011; Fulcher, 2003; North, 2003; Turner, 2012). Some larger organisations make their scales available to the public (e.g. IELTS and TOEFL) and so these can be used in institutions without the resources to develop their own. However, a rating scale is designed for a specific purpose and the more generic, publicly available scales often do not suit the needs of specific educational settings. For this reason, institutions will often develop rating scales in-house, often with limited time and resources. While there is some useful guidance for practitioners regarding scale development, for example, the Council of Europe (2011) publication on the development of the Common European Framework of Reference (CEFR), this type of guidance is limited. This paper, then, attempts to provide guidance to practitioners who find themselves faced with the task of creating or developing a rating scale with limited resources. To do this, the researcher has taken an action research approach to gain first-hand insight into in-house rating scale development. The development of a rating scale for the purpose of assessing an EAP speaking exam is described first, and then recommendations based on the experience are presented. The research setting is the Centre for English Language Education (CELE), a department of the University of Nottingham, Ningbo, a Sino-British university in China. CELE has a large EAP program of approximately 1,500 pre-undergraduate students and 170 pre-masters students.

BACKGROUND INFORMATION: RATING SCALE DEVELOPMENT METHODOLOGIES

The literature on rating scale development methodologies outlines two main approaches: the armchair approach and the empirical approach (Lim & Galaczi, 2013). In the armchair approach, a scale is created based on an expected range of performance and draws solely on expert knowledge. Also called the intuitive method (Council of Europe, 2011, p.208) and the a priori method (Fulcher, Davidson, & Kemp, 2010), this method has the advantage of being based on theoretical views about the development of second-language (L2) ability or from learning objectives set out in a course curriculum. The problems with this approach stem largely from the fact that expectations often do not match reality. In contrast, the empirical approach, also known as performance data-based methods, (Fulcher, Davidson, & Kemp, 2010) is data-driven with evidence ▷

guiding the creation of the scale and the writing of descriptors. Qualitative methods (workshops, observation) as well as quantitative methods, for example, discriminant analysis and item response theory (Council of Europe, 2011, p. 210), are used to develop the assessment instrument. The big advantage with the empirical approach is that the scale not only reflects actual learner performance but also the way they are referred to by assessors.

REPORTING OF RESEARCH ACTIVITY: REVISION OF RATING SCALE FOR ACADEMIC ORAL PRESENTATIONS

The project used in this paper is the revision of a rating scale used for the assessment of academic oral presentations (AOP). The scale is relevant to several users: students to know how they will be assessed and what their scores mean; tutors in preparing students for the assessment; assessors to guide decision making during the assessment; and trainers who use the scale in the training of tutors to carry out assessments of AOPs. With such wide use, the document plays an important role in a course that is compulsory for over 1,500 students. Because of the importance of the document, it is under close scrutiny and therefore under constant review. In its latest review, two areas in need of revision were identified: 1. the level of detail of the descriptors, and 2. the number of components to score.

Looking firstly at the level of detail of the descriptors, the scale descriptors were designed to be sparse for ease of use. In the scale descriptors, the only real change across bands is the adjective used to describe the particular features of the category. For example, in the category of Cohesion (Structure and Linking Points), the differentiation between the five levels of organization is made simply by changing the adjective, i.e. Effective organization changes to Satisfactory organization. Such sparse description does not give users sufficient information about the performance expected at different levels across the scale and there was call from tutors in particular for more detail in descriptions.

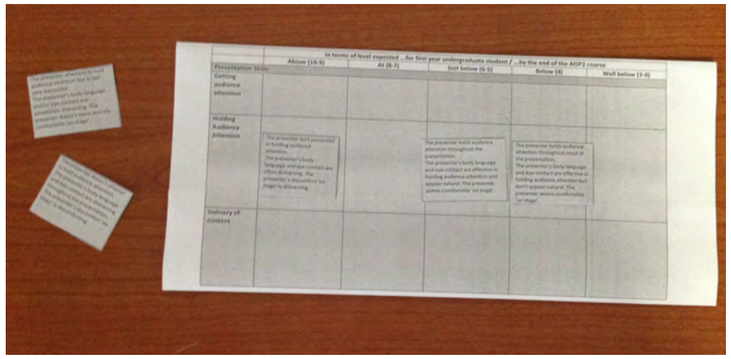
The second aspect identified as in need of change was the number of components to score. The assessment required assessors to provide eleven component scores. As reported by assessors and experienced first-hand by the researcher, assigning such a high number of scores with any degree of consistency or accuracy during a live assessment was extremely challenging and often simply not possible. The aims of the revision then were to 1. expand the descriptions of performance to provide better guidance, and 2. reduce the number of components to be scored to reduce the load on assessors.

OVERVIEW OF THE SCALE DEVELOPMENT ACTIVITIES THAT TOOK PLACE

STEP 1: CREATING A DRAFT SCALE

AIM 1: REDUCE THE NUMBER OF COMPONENTS TO BE SCORED

To reduce the number of components to be scored would require either dropping components or merging them. As a summative assessment, the AOP construct is basically the course learning outcomes: a list of the things students should be capable of by the end of the course. The person responsible for the course required that all the components were retained in the assessment so none could be dropped; that left the second option which was to merge components. To do this the researcher went back to the construct to see if it could be reorganised. Banerjee and Wall (2006) outline a procedure that can be used to establish a construct. To begin with, a list of all the relevant features that need to be incorporated into the assessment is compiled, then the items are organised according to natural groups. Following this idea, the researcher set up a focus group with EAP tutors from the department who taught and assessed on the AOP course, and were therefore familiar with the terms used in the course learning outcomes. As in Banerjee and Wall's (2006) study, a list of the relevant features to be incorporated in the assessment was compiled and then physically cut into pieces. Participants were asked to group the pieces in any way they felt was intuitive. The three tutor pairs all came up with different groupings, which was interesting as it indicated that there is no single correct way to group the statements, as the researcher had hoped there would be. Nevertheless, the groupings created and the justifications given during the session helped to inform a revised construct framework which was considered to better represent the construct and had a slightly reduced number of scorable components.



◀ Figure 1. Uncovering problem areas.

AIM 2: EXPAND PERFORMANCE DESCRIPTIONS

With the components set, the next issue to address was the descriptors. Due to limited time, the arm-chair approach was adopted and a set of descriptors was written by the researcher using knowledge from teaching the AOP course and assessing AOP performances. In developing the descriptors, the researcher discovered the difficulty of attaining consistent labelling throughout and keeping wording brief whilst also clear for a user.

STEP 2: UNCOVERING PROBLEMS IN THE DESCRIPTORS

With a draft set of scale descriptors written, the next stage taken by Galaczi et al. (2011) was to uncover problem areas. They did this using both quantitative and qualitative methods. Rasch analysis was used in the quantitative study to establish which descriptors were consistently applied to which performances. Due to lack of experience, we decided to opt for a qualitative approach: a focus group. Participants were asked to re-order the descriptors which had been cut into separate pieces of paper. In preparation for the session, the scale was divided and a blank grid created with the scoring criteria on the y-axis and a 5-point scale on the x-axis. The individual descriptors were distributed to tutors one set at a time and tutors were asked to put them along the scale at the point they felt they best represented. To make the task slightly more challenging, and therefore force participants to think carefully about where to place the descriptors, participants were not given a full set of descriptors so they also had to identify where the gap created by the missing descriptors should be placed (Figure 1). The premise was if participants were able to re-order them in the same way they are ordered on the master scale, then this would support the descriptors. Indeed, problem descriptors were identified quickly by participants as they were the ones which participants were unable to order easily. The two main problems were lack of clear differentiation from adjacent descriptors and lack of clarity in wording. This method was an efficient way to have tutors familiar with the assessment look closely at the draft descriptors and be forced to make decisions based on their appropriate placing on a scale. After the session, the problem descriptions were revised and a second draft scale was produced.

In their second study, Galaczi et al. (2011) used a verbal protocol to see whether (and how) assessors refer to the descriptors during an assessment. In the present study, a similar method was used with EAP tutors attending a trialling of the second draft of the scale. Four participants familiar with assessing AOPs watched two video presentations and used the descriptors to rate the videos. Participants were provided with note paper to record comments as they completed the task. In this session, rather than giving participants the full 8-component scale (this was seen as unrealistic as the participants were unfamiliar with the new descriptors), they were given descriptors for only three components to score. The researcher clarified meaning of the annotations while collecting them and then used these to inform the creation of a full scale ready for trial.

STEP 3: TRIALLING THE FULL SCALE TO SEE IF RATERS ARE ALIGNED

Similar to the second study run by Galaczi et al. (2011), this study sought to see to what extent raters were in agreement when using the scale to rate an AOP video performance. A secondary aim was to gather further feedback on how closely the descriptors reflect actual performances.

Eight departmental EAP tutors used the trial scale to rate two video oral presentations. Participants independently registered their ratings by highlighting statements across the categories they felt best matched the performance they saw. It was found that raters showed consistent agreement in their evaluation, which suggests the rating scale works well. Raters also reported the scale was an improvement on what they had previously used with the descriptors providing more concrete guidance to decision making. Moreover, it was commented that for any new scale, a period of familiarization and moderation is needed for assessors to begin using the tool as it is designed to be used.

RECOMMENDATIONS RESULTING FROM THE RESEARCH

While the outcome of the scale development was useful for the purposes of this paper (i.e. a revised scale for use in the large-scale assessment of academic oral presentations), of greater importance are the recommendations resulting from the experience, as follows:

1. WORK FROM ACTUAL PERFORMANCES

In the same way that Galaczi et al. (2011) began the development process with experienced assessors identifying areas in need of revision and a committee setting the parameters, so did this study. The research committee agreed that the number of components should be reduced and the descriptions expanded. In hindsight, the importance of reducing the number of components to fewer than five, as recommended by Luoma (2004), was apparent. The scale development should have begun not with a discussion of changes that need to be made but with a viewing of the task being performed (i.e. an academic presentation being given). This would help to ensure that the assessment ▷

	10	8	6	4	2	0
Eye contact	x	xxxx	x			
Use of visuals		x	xxxxx	x	x	
...						

Figure 2. Example of scoring using Distributional Assessment (DA) Model

fits the reality, rather than vice versa, which is a point Alderson (1991) makes. The committee, along with experienced assessors, should watch the performance and take notes of features that stand out. These features, with reference to the course learning objectives, should be compiled to make up the components to be scored. In this way, it would have been established early on that it is not possible for an assessor to successfully handle more than five components during a 10-minute presentation. As well as the number of components to score, the number of bands in the scale (5) was set to fit neatly into the university scoring system of a 0 – 100 scale. In fact, for some components, there clearly were not five sufficiently differentiated levels of performance (e.g. presentation opening or referencing). This again supports the importance of working from examples of performance early on so that only true distinctions are incorporated, rather than forced ones.

Similarly, in the writing of descriptors, working from actual performances with tutors would have helped capture better the true points that distinguish presentations from one another and the meta-language used by assessors when evaluating performance. A method to do this is described in the Council of Europe (2011) document in which workshop participants rank performances, explaining their ranking. This method captures the most salient features used by assessors used to differentiate between levels as well as the language used by the raters to describe performances. This would help to ensure distinctive differences are included in the descriptors, which in turn will help guide assessors better in their assessment judgements.

2. UNDERSTAND HOW ASSESSORS USE A MARKING SCALE AND LET THAT INFORM THE SCALE DESIGN

Considering the fact that a scale is a tool used for guiding decisions during assessment (and for spoken performance, the assessment is usually live), then the way it is used in practice should be considered when designing the scale. From personal experience of live assessment, a tally system in which scores for each component are continuously awarded during the performance is a practical form of record keeping when there are numerous components (the overall score for the component is then made by averaging the scores at the end). Kane (1986) put forward a similar approach for assessment of performance: the Distributional Assessment (DA) Model. With this model, rather than having the assessor observe a performance and try to simultaneously retain global impressions of several independent criteria, the assessor records every judgment they make as they witness the relevant behavior (figure 2).

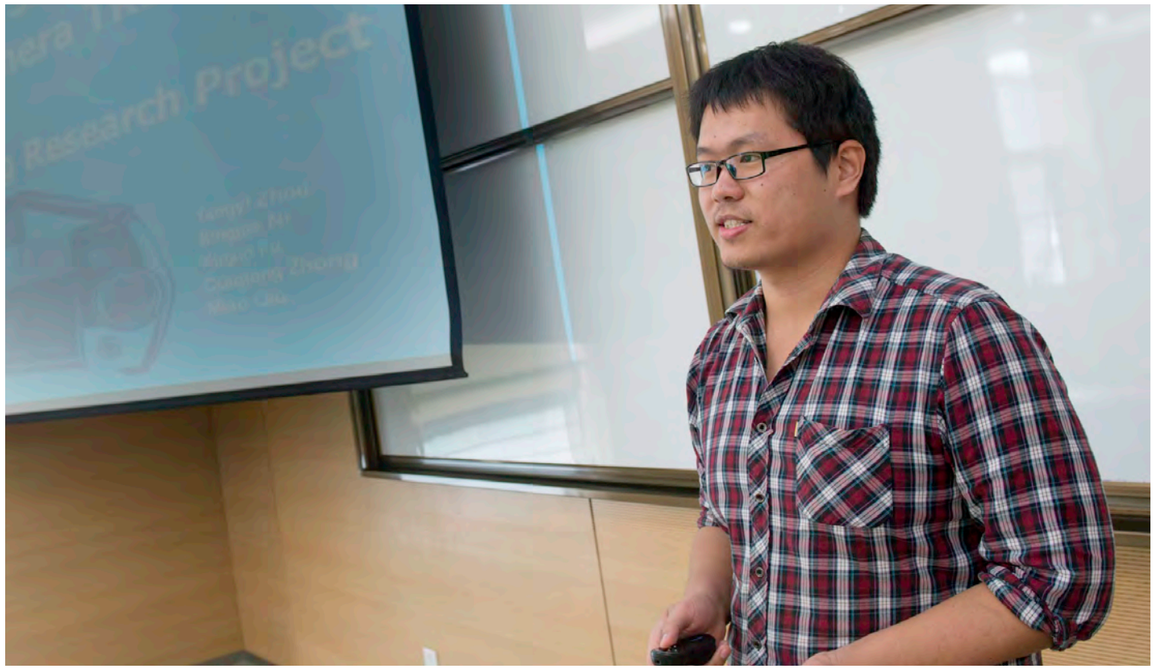
It is therefore worth investigating how assessors approach the task without any guidance and then using that understanding in the assessment tool design.

3. MAKE USE OF WHAT IS ALREADY THERE

In point 1 above (Work from actual performances), it is suggested that language actually used by assessors should be employed as much as possible in scale development. Jeffrey (2015) looked to assessor comments made in coursework feedback to build descriptors for a writing marking scale. Common features assessors used to distinguish adjacent scores were identified which meant assessor meta-language was captured and there was no need to bring tutors in specifically to score and add comments to scripts. Further, the comments were made in regular assessment activities and so highly authentic. While the current assessment (AOP) does not have such a record of assessor comments, as written feedback is not given, there is a record of tutor comments given to benchmarked AOP videos, which are used for assessor training.

CONCLUSION

This study set out to develop a reliable rating scale for the assessment of oral presentations. Through this experience, several lessons were learnt including the importance of a clear understanding of the test construct, the importance of not making the descriptors too contrived (they should reflect reality, not try to dictate it) and the importance of considering the user. Overall, with well-written, empirically-grounded descriptors the task of making a fair assessment becomes easier. A rating scale that is not representative of actual performance, that is unclear or that is difficult to use during an assessment is unsuitable for use and, apart from leading to potentially unreliable or invalid assessment judgements, can create a lack of confidence in users. To make a scale that meets the needs of users requires craftsmanship and sound judgement, but the time spent in getting it right should pay off. In the end, while it was not possible to demonstrate that the rating scale created in this study was producing consistent results, the process of developing the scale has produced a piece of work that will hopefully help to guide others in rating scale development projects of their own.



REFERENCES

Alderson, J.C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.) Language Testing in the 1990s: The Communicative Legacy (pp 71-86). London: Macmillan.

Banerjee, J. & Wall, D. (2006). Assessing and reporting performances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. Journal of English for Academic Purposes, 5, 50-69. doi:10.1016/j.jeap.2005.11.003

Council of Europe (2011). Common European Framework of Reference for Learning, Teaching, Assessment. Council of Europe. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Fulcher, G. (2003). Testing second language speaking. London: Longman/Pearson.

Fulcher, G., Davidson, F., & Kemp, J. (2010). Effective rating scale development for speaking tests: Performance decisions trees. Language Testing, 28(1), 5–29. doi:10.1177/0265532209359514

Galaczi, E., ffrench, A., Hubbard, C. & Green, A. (2011). Developing assessment scales for large-scale speaking tests: a multiple-method approach. Assessment in Education: Principles, Policy and Practice, 18(3), 217-237.

Jeffrey, R. (2015). Using feedback comments to develop a rating scale for a written coursework assessment. Journal of English for Academic Purposes, 18, 51 – 63. doi:10.1016/j.jeap.2015.03.002

Kane, J. S. (1986). Performance distribution assessment. In R. A. Berk (Ed.), Performance Assessment: Methods and Applications (pp. 237-273). Baltimore: Johns Hopkins University Press.

Lim, G. & Galaczi, E. (2013). Pre-conference workshop: Rating scales and raters in speaking assessment. Language Testing Research Colloquium 2013, Seoul, South Korea.

Luoma, S. (2004). Assessing Speaking. Cambridge: Cambridge University Press.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. TOEFL Monograph, 24.

Turner, C. E. 2012. Rating scales for language tests. The Encyclopedia of Applied Linguistics. doi: 10.1002/9781405198431.wbeal1045

Author Biography

Currently EAP Tutor at University of Nottingham in Ningbo; soon to be EAP tutor at Bristol University. I teach EAP courses and design and develop EAP assessments. Recent assessment projects include the revision of rating scales for seminar discussion skills and interview speaking skills.

Email:
simon.dawson@nottingham.edu.cn