Chapter 7
Comparative Judgement as a Promising Alternative to Score Competences

Marije Lesterhuis
University of Antwerp, Belgium

San Verhavert
University of Antwerp, Belgium

Liesje Coertjens
University of Antwerp, Belgium

Vincent Donche
University of Antwerp, Belgium

Sven De Maeyer
University of Antwerp, Belgium

ABSTRACT

To adequately assess students' competences, students are asked to provide proof of a performance. Ideally, open and real-life tasks are used for such performance assessment. However, to augment the reliability of the scores resulting from performance assessment, assessments are mostly standardised. This hampers the validity of the performance assessment. Comparative judgement (CJ) is introduced as an alternative judging method that does not require standardisation of tasks. The CJ method is based on the assumption that people are able to compare two performances more easily and reliable than assigning a score to a single one. This chapter provides insight in the method and elaborates on why this method is promising to generate valid, reliable measures in an efficient way, especially for large-scale summative assessments. Thereby, this chapter brings together the research already conducted in this new assessment domain.

INTRODUCTION

Competence based education has an important share in the curricula in current higher education studies (Heldsinger & Humphry, 2010). As performances are the most direct manifestation of competences, most scholars agree that performance assessments are most suitable to evaluate these competences (Darling-Hammond & Snyder, 2000; Pollitt, 2004). This obviously brings along some challenges (Baker, O'Neil, & Linn, 1993). For a performance assessment to be valid, there needs to be close similarity between the type of performance the student has to execute for the test and the performance that is of interest (Kane, Crooks, & Cohen, 1999). Close-ended tasks and multiple choice exams, for example, are often very much focussed on knowledge reproduction and are quite limited in scope and complexity (Pollitt, 2004). In other words, there is a big difference between asking someone to describe how to do something and actually asking that person to do so. Furthermore, not every competence can be tested via knowledge reproduction alone. Therefore, close-ended tasks and multiple choice exams are less suitable for some types of performance assessment.

Open-ended tasks are more suited for performance assessment. But in open-ended tasks there possibly lies a big challenge, because answers on open-ended tasks are less predictable than those on close-ended tasks. Thus, students' answers will vary to a greater extent, as they have more freedom in the interpretation of the task and how they execute the task. Because there is more variation and unexpected responses of students, human scorers are needed who are able to interpret students' work (Brooks, 2012).

To guide the human scorers, rubrics are mostly used to score performance assessments. Rubrics consist of several criteria or categories concerning aspects or sub dimensions of the competence. Criteria are introduced because they are believed to assure that all assessors look at the same, predefined aspects (Jonsson & Svingby, 2007). However, problems with validity arise as it is almost impossible to formulate all relevant criteria in advance (Sadler, 2009a). In other words, criteria are too reductionist in nature (Pollitt, 2004). In addition to validity, reliability may also be at stake. It is often shown that people differ in how they score tasks, due to differences in severity or leniency (Andrich, 1978; Bloxham & Price, 2015). Also, it has been shown that assessors attribute different scores to the same tasks, dependent on their mood, the moment they assess the task, or the order in which the tasks are being evaluated (Albanese, 2000). Consequently, the use of rubrics does not guarantee high inter-rater reliabilities (Jonsson & Svingby, 2007). Although extensive effort in training assessors upfront can help, this is still not always sufficient in order to gain high reliabilities especially for performance assessments.

To summarize, a lot of different factors influence the validity and reliability of assessments. Unfortunately, rubrics, as an attempt to increase assessment reliability, do not necessarily lead to more reliable assessments and can impede validity. This chapter will introduce an alternative method for scoring in performance assessments, namely Comparative Judgement (CJ). CJ is based on the assumption that people are more reliable in comparing than in assigning scores to single performances (Thurstone, 1927). In CJ, various assessors independently compare several performances of students and decide each time which of them is best with regard to the competence. Based on this, the performances can be ranked from worst to best on a scale. Because the rank-order is based on decisions of several assessors, the scale represents the shared consensus of what a good performance comprises (Pollitt, 2012a). This method seems to be very promising. However, because it has only recently been introduced, many questions regarding its application, advantages and disadvantages remain unanswered. Based on previous research insights, this chapter will firstly provide some background by describing how CJ works, followed by a step-by-step description of how a CJ assessment can be set up in practice. Subsequent sections will further discuss the method with regard to its validity, quality measures and efficiency.

Related Content

Challenge-Based Learning in Higher Education: A Malmö University Position Paper
Cecilia E. Christersson, Margareta Melin, Pär Widén, Nils Ekelund, Jonas Christensen, Nina Lundegrenand Patricia Staaf (2022). *International Journal of Innovative Teaching and Learning in Higher Education (pp. 1-14).*
www.irma-international.org/article/challenge-based-learning-in-higher-education/306650

Young Adults' Civic Engagement in Community-Based Organizations
Amy Grzybowski (2023). *Co-Constructing and Sustaining Service Learning in Graduate Programs: Reflections from the Field (pp. 11-29).*
www.irma-international.org/chapter/young-adults-civic-engagement-in-community-based-organizations/328883

Learning While Teaching: Harnessing the Potential of Peer-to-Peer Learning to Enhance Language Learning
Sandra Vieira Vasconcelos (2023). *Fostering Pedagogy Through Micro and Adaptive Learning in Higher Education: Trends, Tools, and Applications (pp. 93-110).*
www.irma-international.org/chapter/learning-while-teaching/328743

Supporting ESL Students in Distance Learning Courses: A Case Study
Sarah D. Korpi (2018). *Promoting Ethnic Diversity and Multiculturalism in Higher Education (pp. 59-78).*
www.irma-international.org/chapter/supporting-esl-students-in-distance-learning-courses/199152

The Resurrection of the First Accounting Course: The Case for Blended Teaching in Financial Accounting
Gregory J. Krivacek (2023). *International Journal of Innovative Teaching and Learning in Higher Education (pp. 1-17).*
www.irma-international.org/article/the-resurrection-of-the-first-accounting-course/333627