Studies in Educational Evaluation 44 (2015) 1–8

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students

Heidi Hyytinen [a,*], Kari Nissinen [b], Jani Ursin [b], Auli Toom [c], Sari Lindblom-Ylänne [a]

[a] Centre for Research and Development of Higher Education, Institute of Behavioural Sciences, The University of Helsinki, Finland
[b] Finnish Institute for Educational Research, The University of Jyväskylä, Finland
[c] Faculty of Behavioural Sciences, The University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form 10 November 2014
Accepted 11 November 2014
Available online 6 December 2014

Keywords:
Critical thinking
Performance assessment
Constructed-response task
Performance task
Multiple-choice questionnaire
AHELO

ABSTRACT

This article compares the test results of two different performance-based assessments of critical thinking: a constructed-response task from the Collegiate Learning Assessment (CLA) and a multiple-choice questionnaire (MCQ). These tests ostensibly measure the same critical thinking skills, such as analysing, interpreting and evaluating information and problem solving. The study utilised a mixed-method approach to explore the differences in students' (n = 330) test scores. The results showed that the correspondence between the CLA and the MCQ was fully comparable in 45.5% of the students' test performances. Ten percent of the students had completely opposite test results. Explanations for the inconsistent results are discussed in detail.

Introduction

Although critical thinking is considered a vital skill for learning and for coping with ever-changing working environments (Halpern, 2014), there is increasing evidence that all students do not improve their critical thinking skills during their studies in higher education (e.g., Arum & Roksa, 2011; Bok, 2006; Pascarella, Blaich, Martin, & Hanson, 2011). Yet higher education has been identified as an extremely adequate context in which to facilitate learning of these skills (Rapanta, Garcia-Mila, & Gilabert, 2013). Graduate students themselves have reported that their critical thinking skills are inadequate for their future work (Tynjälä, Slotte, Nieminen, Lonka, & Olkinuora, 2007), and employers have expressed similar concerns regarding students' preparedness for working life (Tynjälä, 2008).

Recently, the focus of research on critical thinking has moved to authentic performance assessment. As Stes and her colleagues (Stes, Min-Leliveld, Gijbels, & van Petegem, 2010, 48) have aptly stated, "more attention should be given to studies researching behavioural outcomes, thereby drawing not only on self-reports of participants, but also measuring actual changes in performance". One attempt to develop a direct measurement of students' learning outcomes is the feasibility study carried out by the Organisation for Economic Cooperation and Development (OECD) known as AHELO – the Assessment of Higher Education Learning Outcomes (Coates & Richardson, 2012). The aim of the study was to determine "whether it is practically and scientifically feasible to assess what students in higher education know and can do upon graduation within and across these diverse contexts" (Tremblay, Lalancette, & Roseveare, 2012, 9).

AHELO was implemented at a time when there was pressure to carry out more assessments of students' performance in higher education (Coates & Richardson, 2012; see also Stes et al., 2010). At the same time there was widespread concern about the effects of testing. For example, testing schemes are assumed to exert considerable influence on what universities emphasise in teaching and the qualities of learning they promote. As Brooks (2012, 606) notes: "outcomes of college learning will be those mandated by the test". Test results may also have significant implications for policy-making, especially educational policy within the area of higher education (Douglass, Thomson, & Zhao, 2012; Morgan & Shahjahan, 2014). Another criticism concerns the focus of the tests. Banta

* Corresponding author at: Centre for Research and Development of Higher Education, The University of Helsinki, 00014 University of Helsinki, Finland. Tel.: +358 456781039.
E-mail addresses: heidi.m.hyytinen@helsinki.fi (H. Hyytinen), kari.nissinen@jyu.fi (K. Nissinen), jani.p.ursin@jyu.fi (J. Ursin), auli.toom@helsinki.fi (A. Toom), sari.lindblom-ylanne@helsinki.fi (S. Lindblom-Ylänne).

http://dx.doi.org/10.1016/j.stueduc.2014.11.001
0191-491X/

and Pike (2012) have argued that the skills and outcomes measured by testing represent only a small part of what is important in higher education and in working life (see also Sackett, Borneman, & Connelly, 2008). With these criticisms in mind, we assume that there is a need for studies that investigate more deeply the results of critical thinking tests.

In the area of critical thinking, previous studies have compared the results of self-reports and performance-based assessments (e.g., Bowman, 2010; Bowman & Seifert, 2011). These studies have shown that self-reports and performance assessments measure different aspects of students' abilities and therefore yield different pictures of those abilities. The present study addresses this issue by comparing the test results of two different performance-based critical thinking tests, both of which were used in AHELO. These two approaches measure critical thinking still differ from each other although they both represent the performance-based assessments. The aim of this study is to explore how closely the measures of these two tests are aligned. A strong variance in the test results would have profound implications, as this would mean that the form of assessment substantially affects the findings about student outcomes (Bowman, 2010). Information about the differences between various performance assessment instruments is valuable for interpreting the results of critical thinking tests.

The strengths and challenges of assessments of critical thinking

Critical thinking is defined as purposeful, reasoned and reflective thinking involving an ability to make a reasoned decision between conflicting claims (Ennis, 1991). A critical thinker has the skills to evaluate the credibility of sources, identify assumptions, conclusions and reasons, ask appropriate clarifying questions, synthesise information from a variety of sources and draw appropriate explanations from particular context or type of task (Halpern, 2014). Many researchers have claimed that critical thinking cannot be defined by referring only to skills, because such thinking always involves a disposition to use these skills adequately (e.g., Bailin & Siegel, 2003; Halpern, 2014; Holma, submitted for publication). In this respect a critical thinker knows how to assess the strength of the evidence and the reasons given and, at the same time, shows the disposition to do so (Bailin & Siegel, 2003; Halpern, 2014).

In examining the critical thinking, researchers have used a variety of different tests. The assessments can be roughly divided into two main measurement protocols: self-reports and performance-based assessment. Self-reports, such as surveys, questionnaires and qualitative interviews, focus on students' perceptions of their current attributes or how these attributes have developed over time (Bowman, 2010). The validity of self-report assessment has been discussed extensively (e.g., Bowman, 2010; Bowman & Seifert, 2011; Halpern, 1993; Pike, 1995, 1996, 1999). For example, Halpern (1993, 279; see also 2014) condenses the problems of self-report instruments in the following way: "students may report that they have learned to think better when, in fact, they have not or, conversely, that they have not improved when they really have".

Performance assessment is sometimes presented as a *new* assessment approach (Andrews & Wulfeck, 2014; Dierick & Dochy, 2001). However, there is a long history of using performance-based assessment as an indicator of higher education student learning and development in order to make educational decisions (see Douglass et al., 2012; Ennis, 1991). The roots of today's performance-based assessment can be traced to the first third of the twentieth century with the beginning of standardised testing (Shavelson, 2010). What the various performance assessments have in common is the goal of eliciting what students know and can do (Andrews & Wulfeck, 2014). Performance-based assessments can be further grouped into two main approaches, namely (1) multiple-choice tests or questionnaires and (2) constructed-response tasks. Below we discuss these two main forms of performance-based assessments.

Multiple-choice tests have been a dominant testing regime within the field of research on critical thinking (see Ennis, 1991; Shavelson, 2010). In the test situation the student must analyse a question and then identify and select the correct answers from a list of given options (Popham, 2003). In contrast to the constructed-response task, multiple-choice tests are often promoted as cost effective and objective (Brown, 2001; Fellenz, 2004), as there is no need for human evaluation in scoring them. However, the cognitive demands of multiple-choice tests have been under discussion (e.g., Lindblom-Ylänne, Lonka, & Leskinen, 1996; Nicol, 2007). Many researchers have argued that a multiple-choice test does not necessary encourage students to use higher-order thinking processes (Nicol, 2007; Scouller, 1998). The reason behind that claim is that multiple-choice tests may be answered merely by low-level processing, such as factual recognition and selection (Lindblom-Ylänne et al., 1996; Nicol, 2007). The general view also suggests that it is more difficult to construct an answer than to recognise the right alternative. It is also possible to select a correct multiple-choice answer without really understanding a problem or knowing the various aspects related to it. For example, students can choose one item amongst the possible choices that best suits the question asked, and, of course, it is possible to guess the right answer from the alternatives given (Fellenz, 2004). Examinees can be assured that the correct answer is amongst the response options. Another weakness is that students "may be able to recognise a correct answer that they would never been able to generate on their own. In that sense, multiple-choice items can present an exaggerated picture of a students' understanding or competence, which might lead teachers to invalid inferences" (Popham, 2003, 81–82). Although there is evidence that by applying a well-designed multiple-choice questionnaire it is possible to measure higher-order thinking (e.g., Fellenz, 2004; Jensen, McDaniel, Woodard, & Kummer, 2014), on the basis of a student's answer it is not possible to determine how the student has processed the test questions (Lindblom-Ylänne et al., 1996). Multiple-choice test can never assess student's skill to synthesise or generate own answer either (Popham, 2003).

To address the limitation of multiple-choice tests, researchers have developed alternative assessment methods, namely the constructed-response tasks (Bennett & Ward, 1993). In the constructed-response tasks examinees create their own answers to the questions (Coates & Richardson, 2012; Rodriguez, 2003; Shavelson, 2010). This type of measures are often open-ended tasks in which students need to analyse, evaluate and synthesise complex information as well as provide reasoned explanation (see Popham, 2003; Shavelson, 2010). Therefore, the constructed-response tasks are said to promote higher-order thinking and to encourage extended problem solving more than the multiple-choice tasks. Another advantage is that the constructed response tasks can reveal the level of understanding (Popham, 2003). The constructed-response task also allows students to demonstrate their writing skills (VanTassel-Baska, 2014). These kinds of tasks are sometimes referred to as 'authentic assessment' because these tasks demonstrate the same thinking processes that individuals use when they solve complex problems in their everyday lives (Andrews & Wulfeck, 2014; Baartman, Bastiaens, Kirschner, & van der Vleuten, 2007). However, several disadvantages of the constructed-response task have been reported. The most important is the difficulty of scoring (Attali, 2014). The constructed-response assessment is characterised as subjective and open to scoring bias, because examinees' responses are traditionally scored by using human evaluation (Popham, 2003;