Journal of Applied Research in Memory and Cognition xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning?

Andrew C. Butler*

Washington University in St. Louis, United States

Multiple-choice tests are arguably the most popular type of assessment in education, and much research has been dedicated to determining best practices for using them to measure learning. The act of taking a test also causes learning, and numerous studies have investigated how best to use multiple-choice tests to improve long-term retention and produce deeper understanding. In this review article, I explore whether the best practices for assessment align with the best practices for learning. Although consensus between these two literatures is not a foregone conclusion, there is substantial agreement in how best to construct and use multiple-choice tests for these two disparate purposes. The overall recommendation from both literatures is to create questions that are simple in format (e.g., avoid use of complex item types), challenge students but allow them to succeed often, and target specific cognitive processes that correspond to learning objectives.

Keywords: Multiple-choice, Testing, Assessment, Learning

Over the past 100 years, the multiple-choice test has come to dominate student assessment in the United States and many other parts of the world. In its simplest form, a multiple-choice item consists of a stem (i.e., the context, content, or question the test-taker is required to answer) and a set of potential responses, only one of which is correct (the other responses are commonly referred to as lures or distractors). The origins of the multiple-choice test are complex, with roots in early efforts to measure intelligence, an educational reform movement seeking to create a fairer, norm-referenced system of evaluation, and a pressing need for efficiency in test administration due to increasing student enrollments and other demands such as the testing of soldiers during World War I (Madaus & O'Dwyer, 1999). Many scholars credit Frederick Kelly (1916) with inventing the multiple-choice item and mark the start of its rise in popularity to its adoption in educational (e.g., the Scholastic Aptitude Test by the College Board) and intelligence testing (e.g., the Stanford-Binet Intelligence Test) during the 1920s (cf. Rogers, 1995). The ubiquity of multiple-choice testing in education today stems from the many advantages that it offers relative to other assessment formats. For example, multiple-choice tests are relatively easy to score, offer greater objectivity in grading, and allow more content to be covered by reducing the time it takes test-takers to respond to questions. Students also tend to prefer multiple choice to other assessment formats because they think it is easier (Zeidner, 1987; for review, see Struyven, Dochy, & Janssens, 2005).

Given the popularity and utility of multiple-choice tests, it is not surprising that a great deal of research has been conducted on how best to construct and use them (see Haladyna, 2004). Most of this research has focused on issues related to assessment, such as how to improve the reliability and validity of multiple-choice

Author Note

I would like to dedicate this article to Eugene Winograd, who was my undergraduate mentor and sparked my interest in memory research; fittingly, our first project focused on people's memory for the distractors used on a multiple-choice test. I would also like to thank all the many mentors, colleagues, and students that I have benefited from working with thus far in my career. In particular, I want to give a special thanks to Roddy Roediger (my doctoral advisor) and Elizabeth Marsh (my postdoctoral advisor), both of whom have profoundly influenced my thinking and facilitated my success. I also want thank Nathaniel Raley Woodward, Roddy Roediger, the reviewers, and the editor for their helpful comments and suggestions. While writing this article, I was supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition – Collaborative Grant No. 220020483.

* Correspondence concerning this article should be addressed to Andrew C. Butler, Washington University in St. Louis, United States. Contact: andrew.butler@wustl.edu

Please cite this article in press as: Butler, A.C. Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning? *Journal of Applied Research in Memory and Cognition* (2017), https://doi.org/10.1016/j.jarmac.2018.07.002

tests. Synthesizing the findings of these studies, researchers have identified numerous best practices in an effort to provide practical advice for educators (e.g., Haladyna, Downing, & Rodriguez, 2002; Moreno, Martínez, & Muñiz, 2006). Although some of this advice applies to assessment more generally (e.g., avoid trick questions; omit irrelevant information; use simple vocabulary; check for spelling, punctuation, and grammar), much of it is specific to the multiple-choice format (e.g., make each possible answer to a question about the same length; rotate the position of the correct answer across items). As an aside, it is interesting to note that despite wide-spread communication of these best practices, significant flaws remain common in most multiple-choice assessments (DiBattista & Kurzawa, 2011; Downing, 2005).

Although testing is often conceptualized as a tool for assessment, it is also an activity that causes learning (Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pan & Rickard, 2018; Roediger & Butler, 2011; Rowland, 2014). Retrieving information in response to a test question strengthens memory, leading to better retention of that information over time; it can also change the representation of the information in memory, thereby producing deeper understanding. With respect to the multiple-choice format in particular, numerous laboratory studies have shown that taking a multiple-choice test is beneficial for learning (for review see Marsh, Roediger, Bjork, & Bjork, 2007). Importantly, such positive effects of multiple-choice testing also occur in authentic educational contexts. For example, multiple-choice testing has been found to improve retention and transfer on subsequent unit and final exams in middle school (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger, Agarwal, McDaniel, & McDermott, 2011), high school (McDermott, Agarwal, D'antonio, Roediger, & McDaniel, 2014), and college courses (Butler, Marsh, Slavinsky, & Baraniuk, 2014; Glass, 2009; McDaniel, Wildman, & Anderson, 2012). In addition, multiple-choice testing can enhance the learning of non-tested, conceptually related information (Bjork, Little, & Storm, 2014) and restore access to previously acquired knowledge that has become inaccessible (Butler et al., 2018; Cantor, Eslick, Marsh, Bjork, & Bjork, 2015).

Given that the use of multiple-choice testing both causes and assesses learning, it is important to consider whether the best practices for assessment align with the best practices for learning. The alignment of best practices between these two purposes of testing is not by any means guaranteed. The primary goal of assessment is to measure the extent to which students have acquired the skills and knowledge that form the learning objectives of an educational experience (e.g., an activity, session, or course). To do so effectively, a test needs to differentiate students who have greater mastery of the to-be-learned skills and knowledge from students who have less mastery, which is referred to as *discriminability*. Effective assessment also requires stable and consistent results, which is called *reliability*, and accurate measurement of the intended skills, knowledge, or both, which is called *validity* (Green, 1981). In contrast, the primary goal of using tests for learning is to produce knowledge and skills that are *durable*, so that they will be retained over long periods of time, and *generalizable*, so that they can be flexibly used in different contexts.

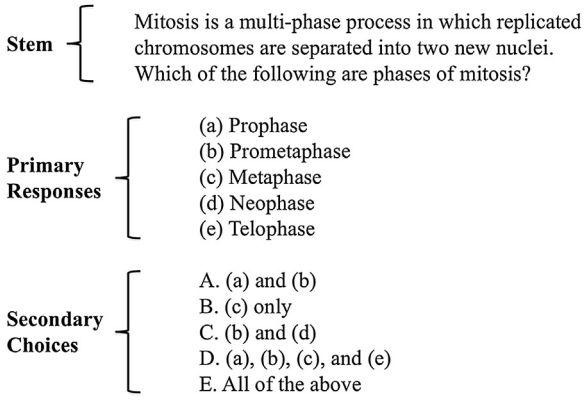

Figure 1. An example of a complex multiple-choice item ("D" is the correct answer).

The following sections explore whether there is alignment between best practices in the use of multiple-choice testing for assessment and learning. To preview the fortunate conclusion of this review, there is great consensus in the recommendations that come from these two conceptually disparate but practically related literatures. Each section focuses on a best practice derived from the set of guidelines for writing multiple-choice items for assessment put forth by Haladyna et al. (2002). The rationale for why the best practice is good for assessment is provided first, and then it is followed with a description of relevant research that helps to explain why it is also a best practice for learning.

Best Practice #1: Avoid Using Complex Item Types or Answering Procedures

Complex multiple-choice (CMC) items have become increasing popular at every level of education as test creators seek to measure higher-order thinking. CMC items are characterized by a stem (e.g., a statement or question), a set of potential responses to the stem (primary responses), and a set of alternatives that present combinations of the primary responses (secondary choices, e.g., "A and B, but not C"; for example, see Figure 1). Variations on this general structure include type K items (a particular format that has four primary responses and a set of five secondary choices that stay constant across items), multiple multiple-choice (instead of secondary choices, test-takers must select all of the correct alternatives to get credit), and multiple sentence completion (the stem contains two or more blanks and the responses contain combinations of words to complete the sentences). The general idea is that the greater complexity in format and/or answering procedure enables the assessment of greater complexity of thinking.

Within the assessment literature, there is a general consensus that CMC items should be avoided for several reasons. First, CMC items tend to be more prone to "clueing"—that is, they inadvertently enable test-takers to engage in strategic guessing. If test-takers can eliminate one or more of the primary responses, it can rapidly whittle down the number of plausible secondary choices. Due to clueing, CMC items tend to produce artificially higher levels of performance and have lower reliability relative to

Please cite this article in press as: Butler, A.C. Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning? *Journal of Applied Research in Memory and Cognition* (2017), https://doi.org/10.1016/j.jarmac.2018.07.002