System 50 (2015) 56–68

Contents lists available at ScienceDirect

System

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

Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism

Qin Xie*

Hong Kong Institute of Education, Hong Kong

ARTICLE INFO

Article history:
Received 18 January 2014
Received in revised form 7 March 2015
Accepted 10 March 2015
Available online 4 April 2015

Keywords:
Washback
Test design
Test preparation
Test-taker perceptions
Structural equation modeling

ABSTRACT

This study utilized Structural Equation Modeling to investigate the washback mechanism, focusing on two design aspects of an English language proficiency test: component weighting (weight assigned to different test papers) and testing methods (item format), and their washback on test preparation. Two months before taking the test, a large sample of test-takers (N = 1000) were surveyed regarding their perceptions of the two design aspects and their test preparation activities. Their official test scores were collected when they were available. Data was analyzed to estimate the washback effects of perceived changes on test-taker time management and approaches to test preparation, and their test performance. The study found that test-takers spent more time on the papers with higher weight and less on those with lower weight. Reporting component scores seemed unable to adjust this tendency. Meanwhile, favorable perception of test validity was associated with a higher level of engagement in both desirable language learning activities and focused test preparation (drilling and cramming). This suggests that favorable perceptions may not be able to reduce negative washback, but may be able to promote positive ones.

1. Introduction

Washback, or backwash, refers to the influence of high-stakes external testing on teaching and learning within the classroom (Alderson & Wall, 1993; Cheng & Curtis, 2012; Green, 2013; Qi, 2011). Existing studies on high-stakes English language proficiency tests (e.g., Alderson & Wall, 1993; Cheng, 2005) provide insights into the complex nature of washback. However, the exact mechanism of washback remains unclear (Cheng, Watanabe, & Curtis, 2004). This is certainly because of the complex nature of washback, where multiple factors interact and shape the extent and nature of the washback getting into the classroom. It may also be partly due to the methodological limitations of existing washback studies. Most existing studies adopt qualitative methods (Cheng et al., 2004; Green, 2007; Matoush & Fu, 2012; Qi, 2005; Zhan & Andrews, 2013); very few studies employ quantitative methods. While qualitative methods can effectively identify various factors affecting the washback getting into classrooms, they are not as effective for unraveling the network of relationships among these factors. Quantitative methods, on the other hand, can answer questions such as in what ways and to what extent do different factors associated with testing affect teaching and learning in classrooms? Such questions are essential to understand the

* B4-2/F-12, Hong Kong Institute of Education, NO. 10 Lo Ping Road, Tai Po Market, New Territories, Hong Kong. Tel.: +852 2948 8368.
E-mail address: qxie@ied.edu.hk.

http://dx.doi.org/10.1016/j.system.2015.03.002
0346-251X/© 2015 Elsevier Ltd. All rights reserved.

mechanism of washback. This study surveyed a large sample of student test-takers and utilized Structural Equation Modeling to explore the mechanism underlying the washback of high-stakes testing. The study was situated within the context of China's reform on the College English Test, focusing on two aspects of test design and their washback effects on test-taker time management, approaches to preparation, and test performance. The following section reviews relevant literature on washback, component weighting, and Structural Equation Modeling.

2. Literature review

Many studies on test washback were conducted following the introduction of new language tests (or new test features) to investigate whether changes made to the tests were followed by desirable changes in teaching and learning (e.g., Cheng, 2005; Hung, 2012; Wall, 2000; Wall & Alderson, 1993). These studies found washback to be highly complex, because multiple factors and multiple stakeholders coexist, and their complex interactions largely determine the extent and nature of washback getting into the classroom (Matoush & Fu, 2012; Shih, 2010; Spratt, 2005; Watanabe, 2004b; Zhan & Andrews, 2013).

For instance, the multiple-choice (MC) test format associated with modern psychometric testing is believed by many to be one source of detrimental washback (Messick, 1996), whereas open-response testing methods, which put less restriction on test-takers, are considered capable of promoting beneficial washback and enhancing test validity. The examples of open-response testing are fill-in the blanks, short-answer questions, and essay writing. Performance testing, which adopts more open-ended testing methods, is generally perceived to have higher potential for generating beneficial washback. Because tasks in performance tests are closer to real-life tasks (more authentic), they tend to receive positive responses from test users, including teachers and students, in regards to their validity in assessing language proficiencies (Struyven, Dochy, & Janssens, 2005). However, studies on performance testing found its washback to be merely superficial (Andrews, Fullilove, & Wong, 2002; Cheng et al., 2004; Wall, 2000). Changes made to these tests will in turn produce changes in the content of teaching and learning, but not in the methods. They may change the manner of test preparation but not its nature. Test washback is described as a “blunt instrument” (Andrews, 1995), which often falls short of the test designers' intentions; that is, good intentions often do not materialize in classrooms.

Compared with the number of washback studies on testing methods associated with performance testing, there are considerably fewer studies on component weighting and its effects on teaching and learning. It is a shared view that if one component is not assessed or is taken away from a test set, this component is likely to be ignored by teachers and test-takers (Kane & Case, 2004; Watanabe, 2004a). However, little has been done regarding the exact weighting of test components and their differential impact on teaching and learning. If test-takers are driven by what is viewed as valuable (Spolsky, 1995; Xie, 2013), it is reasonable to infer that they will spend more time and resources on the components with higher weighting and vice versa. However, such reasoning remains intuitive and may be simplistic. The effects of differential component weighting may be more complex than that. Test preparation is likely to be affected by forces other than rational considerations. For instance, influences from previous test experience and consideration of the cost-effectiveness of resources may affect test preparation. It has been observed elsewhere (e.g., Powers, 1987) that test-takers invest more effort in the sections considered to be “coachable” (i.e., where performance can be improved within a relatively short timeframe) in order to achieve a better return from their investment.

To discourage test-takers' tendency to ignore less weighted components, many test developers chose to report component scores along with a global score. However, this measure alone may not be sufficient for the purpose. The test also has to adopt a non-complementary or conjunctive scoring system. In a non-complementary scoring system, a minimal requirement or a threshold for each component is set so that it is necessary to pass each component in order to pass the whole test. The driving test is an example of the non-complementary scoring system (Kane & Case, 2004). To get a driver's license, test-takers must pass both a written test of traffic rules and a road test of driving skills. A high score on the written test cannot compensate for a low score on the road test, and vice versa. By contrast, in a complementary scoring system, decisions are made based on composite total scores, which are computed as a weighted sum or an average of component scores. Thus, high scores on one component can compensate for low scores on the other components (Hambleton & Slater, 1997; Kane & Case, 2004).

Most literature on component weighting is within the area of educational measurement, where the primary concern is the impact of different weighting schemes on the reliability and validity of composite scores (Rudner, 2001; Sawaki, 2007; Wainer & Thissen, 1993). In this area of studies, consequences of component weighting are mentioned in association with the concept of nominal weighting, which refers to the stated weighting-scheme of a testing program. Nominal weighting reflects policy intention and value judgment regarding the relative importance of different components within a test set, which contrasts with effective weighting, the actual weighting in the composite scores owing to differences in component-score reliability and variance. Effective weighting is a psychometric concern, whereas nominal weighting is pertinent to educational policy and curriculum innovation. Although there are many psychometric studies on effective weighting, little research has been conducted on nominal weighting and its impact on teaching and learning. Since nominal weighting of high-stakes testing conveys a clear, explicit value-message to educational practitioners and test-takers, investigations of nominal weighting schemes, especially in terms of their potential and effectiveness in directing teaching and learning, are clearly warranted. Such studies can provide useful information to facilitate curriculum innovation.

Another gap identified in the washback literature concerns the research method adopted. Most existing washback studies are qualitative; there is a general lack of quantitative studies. Although qualitative studies can identify salient factors affecting