Chapter 13
Evaluating the Assessment Qualities of Teacher-Created Tests

Lauren M. Rein
University of Northern Iowa, USA

ABSTRACT

Due to programmatic pressures and professional reasons, language teachers and program administrators are becoming more interested in the analytical side of test evaluation. Program-wide tests are often maintained by program administrators, while teacher-created tests too often remain in the classroom and are never effectively analyzed for their assessment qualities. This chapter will attempt to describe several practical and approachable methods to plan and evaluate teacher-created tests without complicated and advanced statistical measures. This chapter argues for a harmonious existence of both standardized exams and classroom assessments built around student learning outcomes. After outlining well-known assessment foundations and defining terminology, the chapter will describe the practical necessity of aligning with outcomes in order to meet programmatic standards. Finally, this chapter will provide a variety of qualitative measures to evaluate teacher-created tests, along with a small selection of basic statistical measures.

INTRODUCTION

Language teachers, who manage the daily language instruction of their students, as well as their program administrators, are under increasing pressure to provide evidence for the worthiness of their classroom assessments. However, traditional language assessment research does not always provide achievable methods of doing so. Many researchers pointed out that a great deal of language testing research is historically rooted in advanced quantitative statistics (Lazaraton, 2004; Lumley & Brown, 2005; Rhaman, 2017; Richards, 2009; Tsushima, 2015). Researchers and linguists promote the efficacy of Chi-squared and ANOVA, focusing on the objective data and losing sight of the practical and human needs of classroom testing. Classroom teachers cannot wait weeks for analyses and reports before making curricular
decisions. Nor does the teacher, with the workload of grading, instruction, staff meetings, and student conferences, have limitless time to manage complex analyses on classroom tests. Teachers are left with the demand to prove that their assessments are valid and reliable, without practical tools to meet expectations established by programs or accrediting bodies.

In coincidence with the classroom teacher’s perpetual concern of limited time is the recent broadening of scope of issues that researchers are investigating, as well as the expanding methods of research (Lumley & Brown, 2005; Turner, 2017). What were once narrowly-defined fields of reliability and validity in a norm-referenced testing context have expanded, with concepts and methodologies that overlap and intersect. This growing attention and research into language assessment prompts some classroom teachers to further their training through professional development opportunities (Katz, 2012; Sartor, 2017). Teachers are pursuing a deeper understanding of assessment literacy, as well as ways to evaluate the classroom assessments that they design themselves.

The objective of this chapter is to provide classroom teachers and program administrators some basic tools to evaluate both the qualitative and quantitative qualities of teacher-created tests in order to ensure dependable and valid assessments that align with course objectives.

BACKGROUND

Classroom Assessment

Language teaching and language testing existed on the periphery of academia and research until the late twentieth century, when investigation in the field increased, especially in the direction of qualitative testing (Lazaraton, 2004; Richards, 2009). Traditionally, when assessment researchers described methods in analyzing testing qualities, language tests fell into the paradigms of norm-referenced and criterion-referenced testing. Norm-referenced testing (NRT), such as the TOEFL or IELTS tests, seeks to compare the test-taker to a norm or standard distribution of the collection of all test-takers. NRT is large-scale testing, where a test-taker’s skills are evaluated without classroom instruction. Students are evaluated on a scale and given a score in reference to his or her position in relation to the other test-takers.

Criterion-referenced testing (CRT), on the other hand, measures a test-taker’s specific knowledge in reference to specific skills or a standard of performance, rooted in classroom instruction. In other words, classroom teachers have a list of objectives to teach, and they must evaluate how students are progressing in mastery and skill of these outcomes by using a variety of classroom assessments. Classroom teachers continuously evaluate student performance in achieving objectives, making decisions about the direction of what to teach next based on classroom assessment, both formative and summative. Summative and formative assessment are described more extensively in Fulcher and Davidson (2007). Brown and Hudson (2002) also introduced an alternative to criterion-referenced testing: objective-referenced testing. Objective-referenced testing (ORT) measures the students’ achievement of specific objectives of the class. These tests are designed to evaluate the specific learning objectives as pre-determined by the course designer. See their book for a discussion on the differences of NRT, CRT, and ORT (along with other testing terminology) in more detail.