Second Language Testing for Student Evaluation and Classroom Research

Second Language Testing for Student Evaluation and Classroom Research

by

Greta Gorsuch

and

Dale T. Griffee

INFORMATION AGE PUBLISHING, INC.
Charlotte, NC • www.infoagepub.com

Library of Congress Cataloging-in-Publication Data

CIP record for this book is available from the Library of Congress
http://www.loc.gov

ISBNs: 978-1-64113-011-0 (Paperback)

978-1-64113-012-7 (Hardcover)

978-1-64113-013-4 (ebook)

Cover image: *Survivor*, 1995, oil on canvas, 30 x 84 in. by Tom Palmore, born 1944 Ada, Oklahoma; lives Santa Fe, New Mexico, Albuquerque Museum. Museum purchase, 1993 General Obligations Bonds, 1995.30.1

Printed in the United States of America

DEDICATION

to J. D. Brown, our teacher

Praise for Language *Testing for Student Evaluation and Classroom Research*

Gorsuch and Griffee address a need with this book this for an introduction into language testing. Although the field is extensive, all future teacher/researchers will start at the same point. The chapters of *Language Testing for Student Evaluation and Classroom Research* provide a wealth of information from tests for classroom, commercial, and research purposes. I believe it is important for all readers who are new to language testing be exposed to the concepts in all aspects of the field. In addition to explaining key concepts, Gorsuch and Griffee provide instruction for designing test tasks or items. This can be especially challenging when item or task design is done by a single person (e.g., your typical classroom teacher). The authors include chapters on how to check item quality (Reliability and Test Validity) and that the information from the test is understood correctly (Standard Setting). Each of these three concepts have several methods for measuring them. The authors do an excellent job of providing a well-balanced evaluation of each method so that the reader can make an informed decision in their own practice. The accompanying student workbook provides readers with firsthand exposure to reliability, validity, and standard setting methods so that they can determine the extent they agree with the authors' evaluations. Where many other textbook authors will focus on only one of these areas for the majority of their books, Gorsuch and Griffee discuss the entire cycle of the test development process. My career trajectory so far has taken me from a teacher in my Master and Doctoral schools to an analyst for a large-scale testing organization (Defense Language Institute Foreign Language Center). To this day, I look through every book I have purchased on the subject of language testing. I am confident that this book will become a resource readers will carry with them for years.

— **Jeremy Gevara,** PhD, Instructional Systems Specialist, Test Analysis and Design, Defense Language Institute, Monterey, California

Intended mainly for Second/foreign language teachers and students of applied linguistics, *Language Testing for Student Evaluation and Classroom Research* discusses language testing principles fundamental to the day-to-day performance of a language teacher. The book introduces a variety of language testing concepts in a comprehensive and accessible manner, assuming no formal training in language testing. Gorsuch and Griffee build on their vast experience in the field of language teaching to effectively walk practicing teachers and the students of applied linguistics through the theory and practice of second/foreign language testing and help them design, develop, and validate classroom tests as needed by their teaching practices. The step-by-step approach of the book also enables the teachers to evaluate the existing tests and adapt them to their particular teaching contexts. The companion workbook, with discussion questions linked to the concepts presented in the text, is another valuable feature of this book that not only deepens the students' engagement with the text but also helps them to study independently by using the Answer Key. I strongly recommend *Language Testing for Student Evaluation and Classroom Research* to educators and students who have an interest in or are in need of improving their classroom evaluation practices despite lacking formal training in language testing.

—**Shahrzad Saif,** PhD, Professor, Language Testing and Assessment, Université Laval, Quebec, Canada

CONTENTS

PREFACE

Our goal in writing this testing book is to give a balance between theory and practice for second language teaching, student evaluation, and classroom research. Both of us teach testing to applied linguistics graduate students, but we are primarily classroom teachers, having between us a combined 70 years of experience teaching English as a second or foreign language. We believe that language teachers are, first of all, classroom teachers who are responsible for evaluating students, including assigning grades and giving feedback. We also believe that teachers are researchers who have their own creative ideas and often make their own materials. At the same time, we think classroom teachers are educators and evaluators. Teachers often find themselves in full- or part-time administrative roles, and thus are educators working within an organization. Teachers also find themselves wanting to investigate the effectiveness of the courses they are teaching, and hence they are evaluators. The multiple roles we point out—those of Teacher, Researcher, Educator, and Evaluator—can be expressed as TREE. Thus, we refer throughout this book to you, our readers, as TREEs.

THE TEST AND I

Everybody has a story about tests and testing, and we're no exception. In Japan when we moved from teaching in commercial language schools to university teaching, we were required to administer more formal tests and

Second Language Testing for Student Evaluation and Classroom Research, pp. ix–x

then assign grades. This change forced us to reflect on our attitudes toward testing. We concluded that we did not like the idea, practice, or results of testing. However, we both had to admit that testing was never going to go away. Given the urban, global lifestyle that is emerging around the world, would testing, and the reliance on testing as a means of gathering information about people, be more likely to increase or decrease? The answer seemed obvious. Testing would not and will not go away or decrease in importance; in fact, it is likely to increase. Then and now, the only reasonable response for a language teacher is to learn as much as possible about tests and testing. At the time we were coming to this conclusion, J. D. Brown started offering a testing course at Temple University Japan. Dale enrolled and entered the world of testing, and what he talked about after class was so interesting that Greta enrolled the next time the course was offered. As we applied what we had learned, we became more intrigued with tests, their purposes, their construction, their analyses, and their validation.

ACKNOWLEDGMENTS

It is hard to know how far back to acknowledge people who formed our ideas on testing. We have our grade school teachers, Mrs. Melbone and Mr. Kiefer, who taught us averages, an idea at the heart of much of this book. J. D. Brown's course was our formal introduction to testing. Charles Alderson was and is an inspiration to all of us TREEs. Over the years, William Lan at Texas Tech University has assisted us with interesting conversations on statistics and testing, and Ruth Maki allowed us to attend her cognitive psychology course on experimental design. Roman Taraban, also from the Texas Tech psychology department, went far beyond normal collegial cooperation in introducing us to the idea of test effect. Jeremy Gevara has argued, instructed, and collaborated with us on testing projects. We are both deeply indebted to education, testing, and measurement colleagues, some known personally but many known only through their published works. Other debts of gratitude include Herb Miller for early morning conversations at Starbucks and gifts from his library, and Laura Valentine Rivers, Amanda Kirk, and Heather Thomas for their interesting questions. Thanks to Camille Vilela for her timely help with Chapter 11 on Tests and Teaching. We are truly indebted to our Texas Tech University students in LING 5340 (Second Language Testing) for being our muses, our reasons for writing.

Second Language Testing for Student Evaluation and Classroom Research, pp. xi–xii

WHO THIS BOOK IS WRITTEN FOR

- Practicing language teachers who want to know more about tests with a view to make and improve their classroom tests, and thus more accurately know what their students can do now and what they still need to learn to do.
- Graduate students or late-career undergraduate students in language teaching programs. This book might well be their introduction to testing.
- International students studying in language teaching or applied linguistics programs at English-medium universities. Their English is good, but they are not familiar with many idiomatic phrases. For this reason, we explicitly explain such phrases and include them in the glossary.
- Students who are being supported through graduate school as foreign language instructors in Arabic, Chinese, English, Italian, Japanese, Russian, and so on. They want and need practical application of testing principles to do their jobs. They pushed us to include more formulas in the text.
- Practicing language teachers who do not like testing, especially the idea of statistical analysis.

INTRODUCTION

Testing as We Know It

Tests are everywhere. From pregnancy tests to determine whether life is there, to autopsy tests to determine why life is not there, we are subject to tests from cradle to grave. Throughout life, as Hanson (1993) puts it, we are awash in tests. Schools in particular are sources of testing. We take tests to get into school, we continue to take tests throughout school life, and finally we take tests to get out of school. Although we are primarily interested in second language classroom tests in this book, Teachers, Researchers, Educators, Evaluators (TREEs) cannot be isolated from the larger context of testing both nationally and internationally. So, what has been happening in the larger context?

TESTING AS WE KNOW IT MAY BE OVER

Testing as we have known it may be over. So, first, what is testing as we know it? Second, why might testing as we know it be over? Finally, what might come next for testing?

TESTING AS WE KNOW IT

Testing as we know it began in the 20th century, although the roots of testing go back to the 19th century. Some of the statistical procedures necessary for testing were born in the 18th century, and standardized tests were

Second Language Testing for Student Evaluation and Classroom Research, pp. xiii–xix

developed in China at least by 500 B.C. Nevertheless, testing, *as we know it*, can be explained as a 20th-century phenomenon. Good sources to understand this history are Hanson (1993) in general testing, Spolsky (1995) in second language testing, and Gould (1996) in measurement.

The heart and soul of testing as we know it is the **intelligence test**.[1] This is true for several reasons: The intelligence test is a product of the 20th century. It has not changed significantly since its inception, it is the basis for most large-scale educational testing such as college entrance tests (in the United States at least), and it is the model for proficiency tests such as the Test of English as a Foreign Language (TOEFL). One result of the prolonged use of the intelligence test is that many of us believe that the intelligent quotient (IQ) can be expressed as a single number and is an important indicator of intelligence.

What Is an IQ?

The first modern intelligence test was created by in 1905 by Alfred Binet, who was working as a consultant for the French department of education. His aim was to develop a test that would identify mentally deficient students so they could be helped (Spolsky, 1995). Binet created his test by establishing a series of tasks, each one more difficult, and then matching scores on the tasks (mental age) to the age of children (chronological age) to determine which students were in need of help. Subsequent researchers created a ratio of mental age divided by chronological age, and the IQ was born.

The IQ Test Comes Into Being

Binet's test was introduced to America by H. H. Goddard, who accepted Binet's test and the IQ as a single number and claimed (unlike Binet) that the IQ represented a single entity called intelligence (Dweck, 2000a). He also created the term **moron** to describe high-functioning mental defectives and attributed this condition to biological family inheritance rather than social or medical conditions. At this point, the IQ test was poised to take off. The next step was to modify and popularize the test, and Lewis Terman, a professor at Stanford University, accomplished this task by revising the test items and standardizing the scale so that the average IQ score was set at 100 and the standard deviation was 15 (Gould, 1996). The resulting test became known as the **Stanford-Binet** test and is familiar to us today. Nevertheless, the test still had to be administered by individual examiners to individual respondents. The final step to use the test large

1. For all terms in **bold**, see Glossary.

scale was taken by R. M. Yerkes, who proposed to the U.S. Army in World War I (1914–1918) that they test all recruits so they could be identified for various tasks, including officer selection and training (Hanson, 1993). The technological breakthrough that enabled this massive testing project was the invention in 1915 of the multiple-choice test item. As a result of millions of men taking the **Army Alpha Test**, intelligence testing was launched as a normal activity undergone by normal persons. The IQ ratio was established as a single number that was a measure of one's intelligence, and testing as we know it was born.

Five Reasons Testing as We Know It May Be Over

We have argued that the foundation of modern testing was born in the early 20th century. We now argue that the seeds of its destruction and transformation are being born now, in the early 21st century. So, what are the complaints against testing as we know it, and what might be coming next?

1. Testing Requires a Lot of Money

Testing has become an industry that takes money away from solving our real educational problems. In a letter to the *New York Times*, Krashen (2012) stated that tests do not increase achievement. What is true, Krashen argued, is that testing is expensive: "New York City plans to spend over half a billion dollars on technology in schools, primarily so that students can take the electronically delivered national tests." This money will not be spent on fighting poverty—the real cause of low educational achievement. This money should be spent on health care for children and for books. Our unhealthy obsession with tests and testing prevents that from happening.

2. Testing Is Unfair

Writing specifically about intelligence tests, Hanson (1993) said:

> Intelligence tests are designed in part to promote equal opportunity, but it happens that test scores are perfectly correlated with mean family income: those who score highest on tests have the highest average family income, and those who score lowest come from families with the lowest average income. Thus instruments that aim to promote equal opportunity in fact systematically favor the advantaged to the detriment of the disadvantaged. (p. 6)

The two issues of fairness are test preparation and test construction. Hanson argued persuasively on the first issue of test preparation: Students who enjoyed parental involvement, attended good schools, were in good

health, and possibly attended test preparation academies did better than those who did not. Evidence for this theory came from data released by Educational Testing Service (ETS) as reported in Fallows (1989). The ETS report grouped family income of test takers into four levels (Educational Testing Service, 1980). Then ETS provided average test scores from the test takers from each of the four income levels. The correlation results, seen in Table I.1, are striking.

Table I.1.
The Relationship Between Income and Test Scores

Income levels	$0–$5,999	$6,000–$11,999	$12,000–$17,999	$18,000
Aver test scores	403	447	469	485

Fallows (1989) explained the implications of the correlation:

> If all you knew about two students was how much money their respective families had, you would be able to predict that the student from the richer family would probably get the higher score. (p. 229)

The problem is not with parental involvement, good health, or attending good schools, and we would wish that for all students. The real problem is that the test is supposed to be measuring student abilities, not the factor of family income and all of the advantages that good income creates. These data are from the 1980s, but there is no reason to believe the situation is different now. Of course, the correlation results in Table I.1 would not tell you the results for any particular student. For the reason why, see Chapter 7 on correlation.

The second issue is test construction. Test construction involves the structure of the test. Testing has become associated with large-scale identification of groups of persons that places them in categories of low/middle/high using normal distribution measured in units of standard deviation (see Chapter 6 on descriptive statistics). Failure is **baked into the cake** because for some students to do well on these tests, other students must do poorly.

3. Testing Hides Its Social Role

The charge is that tests have social implications that are not acknowledged by test makers and administrators. The argument is that tests have three aspects: psychometrics, validation, and social practices, with most emphasis on **psychometrics**, some emphasis on **validation**, and little or no emphasis on social practices (McNamara, 2008). Psychometrics is the combination of "psychology" and "metrics" and in practice is the application

of statistical analyses to a test. Two examples are **reliability** (see Chapter 8 on reliability) and **item bias** (see Griffee & Gevara, 2012). Validation (see Chapter 9 on validity) is the presentation of evidence that the **construct** (that which is being tested) is in agreement with the stated test purpose and test use.

By emphasizing psychometric concerns, test makers ignore the social effects of tests. For example, one result of early IQ tests was to claim that not only was the IQ of individuals hereditary, but that whole groups and countries could be rated according to innate intelligence (Gould, 1996). As a result, politicians representing groups that were already established in a country used these ratings to establish restrictive immigration quotas of groups they wanted to keep out. One of the groups thus identified were European Jews, who in the 1930s were restricted to a strict quota and not allowed into the United States. This is not to say that early test makers were against any group, but it is to say that tests can be used by persons who are.

To use an analogy, from a psychometric point of view, a test is like a car, and test makers are like car mechanics who are only concerned with the condition of the car, the workings of its engine, and the functionality of its design. The car mechanics are not concerned with other aspects of the car, such as who drives the car (students who take the test), the driving conditions (who passes and fails and the effect this has on their life), or how the car is used (teachers and how the test affects teaching).

4. Tests Are Instruments of Power and Control

This argument was well articulated by Shohamy (2001) when she posited that tests are instruments from above that force compliance on test takers below: "Tests are used as a method of imposing certain behaviors on those who are subject to them" (p. 17). On a small scale, this classroom practice is common. If a teacher believes that students are for some reason not mastering the material, she may announce a test, which she believes will promote learner study and review. On a national level, a test may be used to maintain or change teaching. Shohamy (2001) gave the example in Israel of a national English as a Second Language (ESL) test in the 1980s, when **communicative language testing** was first being introduced into school curriculums. This resulted in an emphasis on oral, spoken English. The purpose of the test was to motivate teachers to move away from grammar teaching to an emphasis on oral English. This move would be easy for new teachers recently graduated from training programs but more difficult for experienced teachers.

The test was effective because the teachers, like many teachers, **teach to the test**. However, no change in the curriculum occurred, and no training on new teaching techniques was provided to the teachers. Rather than

invest time and money to enable teachers to change their curriculum and teaching, the new test was used as brute force to coerce teachers to change. No wonder teachers resent tests imposed from on high.

5. Tests Create a Dysfunctional Elite

The intelligence test in its many forms serves to identify the best and brightest so they can form the **meritocracy,** from which a society can staff the highest positions in the church, military, business, and government. However, does this elite serve society? Hayes (2012), writing from an American perspective, cited the war in Iraq, the misconduct of individuals such as John Edwards and Bernie Madoff, institutions such as Enron, the banking crisis, a dysfunctional Congress, and the sex scandal in the Catholic Church. He argued that our elites have failed us, and we can assume neither good faith nor competence. Meritocracy has failed us because it assumes **a level playing field** where everybody has a fair chance. However, according to Hayes (2012), this assumption has been subverted by a multimillion dollar test preparation business that is successful to the point that "one of the best ways to predict a student's SAT score is to look at his parents' income: the more money they make, the higher the score is likely to be" (Hayes 2012, p. 38). The flaw in the meritocracy model is that tests strive to find and enable **the 1%** rather than help **the 99%**. We argue that with the failure of the meritocracy, we may be turning more to use tests to assist and teach the 99%.

In that sense, testing as we know it may be at an end.

What Might Come Next?

Messick (1989) helped to rethink test validation. He concluded that tests have social consequences that must be addressed. Thus, two scenarios may occur: one bad and one good. The bad scenario involves **throwing the baby out with the bathwater**. That would involve reducing or eliminating tests and adopting an anti-psychometric stance, claiming that validity and reliability are **positivistic** and for that reason should be ignored. In other words, people are more than numbers and scores on tests.

The good scenario involves the use of testing to assist the 99%. We should work to have high-quality tests that serve our learners' needs and our social needs, and this is the main aim of this book. This goal, of course, is easier said than done. But the first step is for TREEs to know about and how to do tests. Consider this: Most second language teachers (TREEs) probably do not think their classroom tests influence anyone outside their class. On the contrary, this testing book first assumes that tests are used to

evaluate students through grade assignment, which do have an effect on students' lives and the lives of their families and are of legitimate concern to others in the school. Second, this book assumes that tests can be used as part of course evaluation, an applied type of research (see Griffee & Gorsuch, 2016). For instance, if students in a foreign language program are all getting Cs based on test scores, then a department head, program head, or teacher (all TREEs) may use the low scores to perform evaluations of the courses. Why are students doing so poorly? Does the curriculum support students to the desired outcomes? What changes might be needed? Finally, this book assumes tests can be used for research purposes aside from evaluation, which is essential for a course, program, or teacher to evolve. For example, if learners are taught to process test scores and teacher comments as feedback, will they be more able to improve future test-specific performance? Because all three purposes can affect society, and because testing is not likely to go away, we can work to make our tests be as helpful as possible to the 99%.