6 Potential Solutions to Practical Equating Issues

Alina A. von Davier

Educational Testing Service

I work on problems in statistics that I can solve.
—Rupert Miller (Stanford, Department of Statistics) to Paul Holland, circa 1964

6.1. Introduction

Test equating methods are used to produce scores that are interchangeable across different test forms that are built to the same specifications (Holland, Chapter 2; Holland & Dorans, 2006; Kolen, Chapter 3). It is the most stringent form of score linking because it claims score interchangeability, not merely comparability, as do concordances and predictions (see Holland & Dorans and Holland, Chapter 2, for more details and definitions of types of score linking). Other types of score linking might use the same computations as test equating but do not result in scores that are interchangeable. A linking typically does not qualify as an equating when the test forms are not constructed to the same specifications or when the test forms measure different constructs. Test equating places several stringent requirements on the content and statistical properties of the test forms and on the samples of test-takers involved and is vulnerable to deviations from these requirements. These deviations might result in scores that are not interchangeable. In these circumstances, the intended test equating becomes a weaker form of test linking and the lack of interchangeability of scores can lead to unintended unfairness to some test-takers.

A good equating is like good cooking: It starts with good ingredients, the right tools, sound knowledge, and a bit of talent. Some “stumbling blocks to equating” (Cook, Chapter 5) appear when the assumptions

The opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.
required by an equating method are not fulfilled—for example, when the population invariance assumption fails (Dorans & Holland, 2000). Other stumbling blocks arise when the samples available for equating are too small and when large differences exist in the abilities of the groups that take the two test forms to be equated. In these situations, the equating issues are further exacerbated by poor test or anchor-test construction. In an attempt to address these stumbling blocks, researchers have measured the impact on equating of failures of assumptions (population invariance studies, studies on the quality of the anchor) and have developed new strategies to cope with design and data difficulties (equating with small samples, new approaches to anchor-test construction, and new equating models).

This chapter outlines some of this new research and discusses how it can improve test equating practice.

Before embarking on this investigation of the usefulness of new methodologies, we need to remember that, so far, no systematic theory of test equating has been outlined. Over the years, methods have been developed in response to the need to create comparable test scores in practical circumstances. In order to evaluate these methods, Dorans and Holland (2000), Holland and Dorans (2006), Kolen and Brennan (2004), and Lord (1980) have laid out a framework that defines a good equating procedure. This framework is based on the following five requirements on the test forms and on the equating functions: the same construct, equal reliability, symmetry, equity, and population invariance requirements. “This is not much of a theoretical underpinning for test equating,” said Dorans and Holland (2000, p. 283). Moreover, many of these requirements are vague or arguable. In addition, in most situations, a failure of any of these requirements is hard to detect using the available data. The combination of the lack of a theory and difficulties in detecting bad equating results in practical settings create a challenging situation for a practitioner.

The research overviewed in this chapter is mostly focused on observed-score equating methods and investigates the following equating issues:

1. The population sensitivity of equating functions
2. Small samples equating
3. Addressing the differences in ability of the two groups of test-takers by matching on an anchor test and by constructing the anchor test in nontraditional ways
4. Addressing the stability of the equating results by implementing new equating models such as kernel equating (KE) and by applying the KE framework