Scale Consistency, Drift, Stability: Definitions, Distinctions and Principles

Shelby Haberman

Neil J. Dorans

ETS, Princeton, NJ

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME)

April 13-17, 2009, San Diego, CA.
Abstract

For testing programs that administer multiple forms within a year and across several years, score equating is used to ensure that scores can be used interchangeably. In an ideal world, sample sizes are large and representative, populations are constant, and very reliable alternate test forms are built with nearly identical psychometric properties. Under these conditions, most equating methods produce score conversion close to the identity function. Unfortunately, equating is sometimes performed on small non-representative samples of varying ability that are administered tests built to vague specifications. Here, different methods produce different results because they are based on different assumptions. When individual equatings are concatenated over time, there are shifts in score conversions. In the nearly ideal case, there are smaller shifts because great effort is taken to control variation. Sometimes, there are many entangled sources of variation. In this paper, we make distinctions among terms often used to describe shifts in raw-to-scale conversions, list different sources of variation, and suggest principles that should guide the test-construction and test-scheduling processes.
1. The Ideal

In an ideal world, measurement is flawless, and score scales are properly defined and well maintained. Shifts in performance on a test reflect shifts in the ability of examinee populations and any inconstancy in the raw-to-scale conversions across editions of a test are minor and reflect the fact that score equating procedure based on very large samples are accomplishing their intended purpose. In an ideal world, many things need to mesh. Tests are parallel or nearly so. Populations are fairly steady. Samples are representative and sufficiently large so that sampling error has minimal effect on equating. Likewise, the number of test administrations is small. When these ideal conditions are not met, the phrase “scale drift” may be heard.

2. The Real

Reality differs from the ideal in several ways that may contribute to the appearance of and actual existence of scale drift, where this drift is defined to be a change in the interpretation that can be validly attached to scores on the score scale. Among these sources are population shifts, inconsistent or poorly-defined test-construction practices, estimation error associated with small samples of examinees, and inadequate anchor tests. In addition, a sequence of sound equatings can produce non-random drift.

2.1 Population Shifts

In practice, mean scores change over time. Sometimes this change is seasonal within a fixed testing time, e.g., lower-scoring students tend to take the SAT® at the end of the calendar year. Other times, it can represent a shift in the population. Whenever score distributions shift in one direction over time, there is a tendency to wonder whether the score scale has remained intact. Does a shift in score distributions imply scale drift? It does not. The shift could be due to scale drift and/or to other factors.

During the 1960s and 1970s, average SAT scores declined, perhaps because educational opportunities were made available to the less academically prepared. During the 1980s, average SAT scores increased, perhaps because of a reversal in self-selection trends observed a decade earlier. The declines of the 1960s led to investigations of “scale drift” by Modu and Stern (1975, 1977) as part of a broad investigation of the SAT score decline (Wirtz, 1977).
2.2 Test Construction Practices

When tests are built successfully to very tight specifications it is reasonable to expect that the conversion that takes a raw score onto a score scale are the same for all versions of a test form. This raw-to-scale consistency is viewed as evidence of scale stability and lack of scale drift. Is variability in raw-to-scale conversions evidence of scale drift? Perhaps, but raw-to-scale inconsistency, however, is not necessarily due to scale drift.

In general, unstable raw-to-scale conversions can be due to a variety of sources such as differences in test difficulty, differences in test content, changes in test reliability, and the instability associated with sampling error.

Variable raw-to-scale conversions can be due to loose test-construction practices or to vague specifications. If the test assembly line does not follow a precise blueprint or if the resources specified by the blueprint are not available, variations in tests can occur that lead to variation in raw-to-scale conversions. This phenomenon is not scale drift; however, too much variation in the raw-to-scale conversions may lead to scale drift.

If changes occur in the blueprint either as a result of proactive planning or in response to shortages in items, shifts in the raw-to-scale conversions should be expected. Theses shifts may qualify as scale drift if the construct being measured has changed enough so that a score based on the old blueprint is different from a score based on the new blueprint (see Liu and Walker, 2007 and Brennan, 2007 for a discussion of what to look for with tests in transition.)

Less reliable tests can not be equated to more reliable tests (Holland and Dorans, 2006). As a consequence, the linking of a less reliable test to a more reliable test is likely to be population dependent. This population dependence can manifest itself as increased variability in raw-to-scale conversions. In addition, linking the less reliable test to the more reliable test on the will align scores that have different meanings with each other. This is a form of scale drift.

2.3 Sampling Errors

With finite samples of examinees, there is random error in equating due to estimation. The standard deviation of this error is approximately proportional to the reciprocal of the square root of the sample size. This random error introduces random
noise into the equating process. Over links of equatings, drift can occur. This is drift due to a non-systematic source. It can be called scale instability. Scale drift has systematic sources.

One potential source of systematic error is non-random sampling of examinees. For example, selection on the basis of the tests to be equated can be shown to affect the raw-to-scale conversions. When the sample is not representative of the population, systematic error can be induced, especially in the absence of an anchor. These sources of systematic error can induce scale drift.

2.3 The Role of the Anchor

The anchor-test design is subject to more sources of drift than a well-executed equivalent-group design. The role of the anchor is to convert the anchor-test design into equivalent groups either via a chaining process or via post-stratification methods (Holland and Dorans, 2006). Much can go wrong with this design. The groups may be too far apart in ability. The anchor may not have a strong enough correlation with the total tests to compensate for the lack of equivalence between the samples for the old and new forms. The anchor may possess different content than the tests. All of these factors can result in raw-to-scale conversions that vary as a function of equating method. These variations can induce scale drift, and the set of anchor-test influences may in fact be the largest contributing factor to scale drift.

3. Accumulation of Error

Accumulation of random error over many successive administrations can produce random scale drift. This accumulation is the bane of continuous testing. With any testing program that has a fixed level of demand, an increase in number of administrations is accompanied by a decrease in the sample sizes available for equating. For the special case where a doubling of administrations is accompanied by a halving of sample size, the net effect is a doubling of random scale drift. Ignoring this important relationship between total volume, the number of administrations, and scale drift can lead to practices that undermine the scale of a test very rapidly. Random scale drift can have effects very similar to those of systematic scale drift. In typical data collection, equating results within a small time interval are much more similar to each other than they are to results derived in the distant past.
4. What is a Large Effect?

When placed in the context of the non-random error associated with a given test form, the conditional error of measurement or its average, the standard error of measurement, the amount of drift induced by any and all sources can seem to be small. For example, when the standard error of measurement for a test on a 200-to-800 point scale is 40 points, then a drift of 10 points might seem small in comparison. It is an inappropriate comparison. The former is random error, which means that on average across all people, it is expected to be close to zero. Drift on the other hand is in one direction. Across all people it is 10 points rather than zero points. This distinction is important to keep in mind as we attempt to tease out systematic sources of error from random sources.

5. Summary of Effects

Table 1 contains a summary of the previously-described systematic and non-systematic sources of variability in equating settings. For us, drift means a shift in the meaning of the score scale, which can only be induced by systematic sources of variability, which includes the accumulation of random errors.

<table>
<thead>
<tr>
<th>Source</th>
<th>Systematic</th>
<th>Random</th>
<th>Scale Drift?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Shift</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Test Difficult Shift</td>
<td>No</td>
<td>Maybe</td>
<td>No</td>
</tr>
<tr>
<td>Construct Shift</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Reliability Shift</td>
<td>Yes</td>
<td>No</td>
<td>Maybe</td>
</tr>
<tr>
<td>Random Sampling</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Accumulated Random Error</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Non-random Samples</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Non-representative Samples</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Inadequate Anchors</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

In sum, shifts in score distributions or raw-to-scale conversions can not be used as definition as for scale drift because these shifts can occur for non-drift reasons such as
shifts in population composition or alterations in test difficulty specifications. These shifts, however, may provide evidence of scale drift. Teasing out scale drift from non-drift shifts requires data-collection design in which an old test is administered to a new population. Ideally this type of experiment would be replicated several times. In practice, this direct comparison may be impractical due to changes in the environment of the examinees due to modifications of curriculum, public attention to portions of test content, changes in scientific knowledge, etc.

6. Overview of the Session

Scale drift is a shift in the meaning of score scale that alters the interpretation that can be attached to score points along the scale. We have attempted to clarify that trends in score distributions and inconsistent raw-to-scale conversions do not necessarily indicate scale drift. In addition we have made tried to identify various sources of variability in score distributions and conversion tables that may or may not induce drift. In the next paper in this session, Haberman, Lee and Qian (2009) use jackknifing techniques to evaluate equating accuracy. This paper looks at the effects of both sample size and choice of anchor on the stability of raw-to-scale conversions. Haberman, Guo, Liu and Dorans (2009) examine trends in SAT means and the consistency of SAT raw-to-scale conversions over about eight years of SAT administrations. Liu, Curley and Low (2009) use both an equivalent-groups design and a non-equivalent-anchor-test design to examine scale drift over 11 years. With the exception of the study by Liu, Curley and Low (2009), the studies in this session are describing shifts in score distributions or variability in conversion tables.
References


