Listening. Learning. Leading.®

Research Report
ETS RR–11-10

Sources of Score Scale Inconsistency

Shelby J. Haberman and Neil J. Dorans

March 2011

Sources of Score Scale Inconsistency

Shelby J. Haberman and Neil J. Dorans
ETS, Princeton, New Jersey

March 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Daniel R. Eignor

Technical Reviewers: Jinghua Liu and Longjuan Liang

Abstract

For testing programs that administer multiple forms within a year and across years, score equating is used to ensure that scores can be used interchangeably. In an ideal world, samples sizes are large and representative of populations that hardly change over time, and very reliable alternate test forms are built with nearly identical psychometric properties. Under these conditions, most equating methods produce score conversions close to the identity function. Unfortunately, equating is sometimes performed on small non-representative samples with variable distributions of ability, and administered tests are built to vague specifications. Here, different equating methods produce different results because they are based on different assumptions. In the nearly ideal case, there are smaller deviations from the identity function because great effort is taken to control variation. Even when equating is conducted under these desirable conditions, the random variation in form-to-form equating, when concatenated over time, can produce substantial shifts in score conversions, that is, scale drift. In this paper, we make distinctions among different sources of variation that may contribute to score-scale inconsistency, and identify practices that are likely to contribute to it.

Key words: scale drift, systematic error, random error, reliability, seasonality, random walk

Table of Contents

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 variability in the raw-to-scale conversions across editions of a test are minor and due to random sampling error. This stability reflects the fact that score equating procedures based on very large samples are accomplishing their intended purpose. In an ideal world, many circumstances mesh. Tests are parallel or nearly so. Populations remain constant over time. Samples are representative and sufficiently large so that sampling error has minimal effect on equating. Likewise, the number of test administrations is small. Dorans, Moses, and Eignor (2010) discussed these and other best practices for score equating.

Reality differs from the ideal in several ways that may contribute to scale inconsistency which, in turn, may contribute to the appearance of or actual existence of scale drift. Scale drift is defined to be a systematic change in the interpretation that can be validly attached to scores on the score scale. Among these sources of scale inconsistency are inconsistent or poorly defined test construction practices, population changes, estimation error associated with using small samples of examinees, accumulation of errors over a long sequence of test administrations, use of inadequate anchor tests, and equating model misfit. In this paper, we make distinctions among different sources of variation that may contribute to score scale inconsistency. In the process of delineating these potential sources of scale inconsistency, we indicate practices that are likely either to contribute to inconsistency or to attenuate it.

1. Sources of Score Scale Inconsistency

1.1 Test-construction Practices

When tests are built successfully to very tight specifications, it is reasonable to expect that the conversion that maps a raw score onto a score-reporting scale will be the same for all versions of a test. This raw-to-scale consistency is viewed as evidence of scale stability and lack of scale drift. Is variability in raw-to-scale conversions or scale inconsistency evidence of scale drift? Perhaps, but raw-to-scale inconsistency is not necessarily due to scale drift. In general, unstable raw-to-scale conversions can also 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 overly vague specifications. If the test assembly process does not follow a precise blueprint or if the resources needed to execute the blueprint are not available, variations in tests can occur that lead to variations in raw-to-scale conversions. This phenomenon is not scale drift; too much variation in the raw-to-scale conversions, however, 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 induce 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 Brennan, 2007 and Liu & Walker, 2007 for a discussion of what to look for with tests in transition.)

Sometimes new versions of a test are less reliable than older versions because a reduction in testing time leads to a reduction in the number of test questions. Less reliable tests cannot be equated to more reliable tests (Holland & Dorans, 2006). As a consequence, the linking of a less reliable test to a more reliable test is likely to be subpopulation dependent. This subpopulation dependence can manifest itself as increased variability in raw-to-scale conversions. In addition, linking scores from the less reliable test to scores from the more reliable test will align scores that have different meanings with each other. This result is a form of scale drift.

1.2 Subpopulation Shifts and Changes in Populations

In practice, mean scores change over time. 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 necessarily imply scale drift? It does not.

Sometimes this change is seasonal within a fixed testing time. SAT® scores exhibit a within-year seasonality (Haberman, Guo, Liu, & Dorans, 2008) that is fairly consistent over years. This seasonality is an example of subpopulation shifts within a given population. Over a long period of time, cyclical seasonal shifts in subpopulations

may occur against a backdrop of significant changes in the population that may affect score meaning.

During the 1960s and 1970s, average SAT scores declined. 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). Was the decline due to scale drift, or did factors such as the availability of financial support and threat of the draft alter the composition of the population of students who applied to college? During the 1980s, average SAT scores increased. Was this increase in means dues to scale drift or to a population shift associated with less available financial aid, the end of the draft (and the war), or other factors that might alter the probability of applying to college?

Another example of a shift in population is the increase in the proportion of test takers who take a test measuring mathematical or science proficiency that is administered in a language that is not their primary language. For some of these test takers, performance on the test may be more a function of their language level than their competency in the math or science domain. While small shifts in the language composition of the population do not necessarily affect score scales (Liang, Dorans, & Sinharay, 2009), changes in the composition of the population may, however, be large enough to change the construct.

1.3 Sampling Errors

With finite samples of examinees, there is random error in equating due to the estimation process. 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. It is a source of scale inconsistency that is nonsystematic. As noted in 1.4, however, accumulation of these random errors can induce a substantial shift in the meaning of the scale.

One potential source of systematic error is nonrandom sampling of examinees. For example, selection on the basis of the tests to be equated affects the raw-to-scale conversions. Suppose scores below the 20th percentile of one form were deleted from the analysis. The linking of scores on that test form to another other test form in that truncated sample would differ from the linking between the tests in the full population. 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.

1.4 Accumulation of Random Equating Error

Accumulation of random error over many successive equatings can produce scale drift. This drift can be the consequence of a random walk (Feller, 1968). Because random sampling error is nonsystematic, the expected value of accumulated errors associated with a sequence of linkages is zero. The variance of this accumulated error, however, is a function of the number of steps in the walk, or in this case the number of links in an equating chain. As the number of links in the chain increases, the probability of a substantial amount of drift increases. In addition, the degree of drift tends to be correlated across successive linkings. Hence we state that the accumulation of random error via a random walk over successive equatings can produce scale drift. Alho and Spencer (2005) illustrate the effect of random walks on the forecasting of demographic trends, while Malkiel (2003) provides an accessible treatment of random walks in the context of the stock market. Haberman et al. (2008) and Haberman (2010) demonstrate that linking test forms across many administrations will produce a phenomenon that is similar to a random walk.

This accumulation of error may be the bane of continuous testing. With any testing program that has a fixed level of demand, an increase in the 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 sizes, the net effect is a doubling of random scale drift. Ignoring this important relationship among total test volume, the number of administrations, and scale drift can lead to practices that undermine the scale of a test very rapidly. Accumulation of random scale drift can have effects very similar to those of systematic scale drift. In typical data collection, equating results obtained within a small time interval are much more similar to each other than they are to results derived over a long time interval.

1.5 The Role of the Anchor

The anchor-test design is subject to more sources of drift than a well executed equivalent-groups design. The role of the anchor is to convert the anchor-test design into

equivalent groups, either via a chaining process or via poststratification methods (Holland & 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 adequately 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.

1.6 Model Misfit

We have discussed several factors that contribute to inconsistencies in raw-to scale conversions. Most have dealt with the tests and the people that take the tests. How the data are analyzed is another factor. Equating procedures apply statistical models to data to produce equating functions that are concatenated across time to place test scores from different test forms on a common score reporting scale. When the data collection is well designed, equating methods tend to give convergent results. For example, when tests are parallel and large representative equivalent samples of test takers are administered these parallel forms, the resultant equating is likely to be approximated well by an identity function. If these conditions do not hold, the identity is unlikely to hold, and different equating methods will give different results. Bias associated with equating model misfit is likely to contribute directly to scale drift; small sample sizes are likely to contribute directly to scale inconsistency and indirectly, via accumulated error, to scale drift.

2. Summary of Sources of Inconsistency

We have attempted to clarify that neither differences in score distributions nor inconsistencies in raw-to-scale conversions need indicate scale drift. In addition, we have tried to identify various sources of variability in score distributions and conversion tables that may or may not induce drift. For example, there are times when tests do not meet specifications because of random errors associated with pretesting with small samples. Other times, the specifications cannot be met on a regular basis because they are unrealistic or have been changed. The former type of failure to meet specifications is

unsystematic or random, while the latter is systematic and chronic and more likely to induce scale drift. Both types are included in Table 1, which contains a summary of the previously described systematic and nonsystematic sources of scale inconsistency. Scale drift is a shift in the meaning of a score scale that alters the interpretation that can be attached to score points along the scale.

Shifts in score distributions or raw-to-scale conversions cannot be used as a definition for scale drift because these shifts can occur for reasons unrelated to drift. These kinds of shifts are labeled *No* under *Induce scale drift?* in Table 1. These shifts, however, may be due to scale drift. Teasing out scale drift shifts from nondrift shifts requires data collection designs 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 because of changes in the environment of the examinees due to modifications of curriculum, public attention to portions of test content, changes in scientific knowledge, etc.

The fact that random errors, which are not direct sources of drift, can accumulate over linkings of test forms to induce scale drift is counterintuitive and has implications for continuous testing. In continuous testing, many test forms are assembled and administered within a given testing period. As a consequence, there is greater variation in test difficulty, increased likelihood that tests will not meet specifications, increased chances that test reliability will be unequal, more administrations and more subpopulations of test takers, increased error due to reduction of sample size, more opportunities for use of inadequate anchors, and greater reliance on equating models that may not fit the data adequately. Accumulation of these sources of inconsistency and instability over time is likely to produce drift fairly rapidly (Haberman, 2010). If a test form were administered at the beginning and the end of a chain of equatings, a score with a given numerical value on that form when administered at the beginning of the chain may correspond to a score on that same form that is a third of a standard deviation higher or lower at the end of the chain. In some cases where tests are administered continuously, the chain may span less than 1 year. The validity of a test score is undermined whenever it matters when the test form is administered.

The standard error of measurement is often used as a standard for assessing the amount of error in a score. Compared to the standard error of measurement, the amount of drift induced by any and all sources of scale inconsistency 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, an average drift of 10 points might seem small in comparison. This is an inappropriate comparison for several reasons. The standard error of measurement is a measure of the variability in random measurement error associated with the number of questions on the test. The average of these random errors, across test forms and across people, is expected to be zero. Drift, on the other hand, is the same for all people at a given score, and it does not cancel out. It is important to keep in mind the distinction between effects of random error on individuals versus effects on groups when comparing systematic drift to random sources of inconsistency in score scales: for a given administration of a test form, increasing the sample size for a group of examinees reduces sampling error of the group mean but has no effect on the standard error of measurement for any individual.

Table 1

Sources of Score Scale Inconsistency

Source	Systematic	Random	Induce scale drift?
Test difficulty shift - random	No	Yes	No
Test difficulty shift - chronic	Yes	No	Yes
Construct shift	Yes	No	Yes
Reliability shift	Yes	No	Yes
Subpopulation shift	Yes	No	Maybe
Population change	Yes	No	Yes
Random sampling	No	Yes	No
Accumulated random error	No	Yes	Yes
Nonrandom samples	Yes	No	Yes
Nonrepresentative samples	Yes	No	Yes
Inadequate anchors	Yes	No	Yes
Model misfit	Yes	No	Yes

3. Conclusions

The list of sources of scale inconsistency is a partial list and reflects our current thinking on the sources of scale inconsistency. The major points of this paper are the following:

1. Differences in score distributions or inconsistencies in raw-to-scale conversions may or may not indicate scale drift;
2. Sources of scale inconsistency may be random or systematic;
3. Systematic sources are more likely to induce scale drift;
4. Accumulation of many random errors, however, may induce a drift similar to what is known as random walk;
5. Alteration in the meaning of a score scale, which will eventually occur due to systematic and random errors, occurs more rapidly with continuous testing than with testing with few administrations;
6. It is important to keep in mind the distinction between effects of random error on individuals versus effects on groups when comparing systematic drift to random sources of inconsistency in score scales.

References

Alho, J., & Spencer, B. (2005). *Statistical demography and forecasting.* New York, NY: Springer-Verlag.

Brennan, R. L. (2007). Tests in transition: Synthesis and discussion. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161–175). New York, NY: Springer-Verlag.

Dorans, N. J., Moses, T., & Eignor, D. E. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). New York, NY: Wiley.

Haberman, S. (2010). *Limits on the accuracy of linking* (ETS Research Report No. RR-10-22). Princeton, NJ: ETS.

Haberman, S. Guo, H., Liu, J., & Dorans, N. J. (2008). *Trend analysis in seasonal time series models: Consistency of SAT® reasoning score conversions* (ETS Research Report No. RR-08-67). Princeton, NJ: ETS.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating*. (ETS Research Report No. RR-09-05). Princeton, NJ: ETS.

Liu, J., & Walker, M. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York, NY: Springer-Verlag.

Malkiel, B. C. (2003). *A random walk down Wall Street.* New York, NY: W. W. Norton & Company.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT score scale* (ETS Research Bulletin No. RB-75-9). Princeton, NJ: ETS.

Modu, C. C., & Stern, J. (1977). The stability of the SAT-Verbal score scale. In W. Wirtz, (Ed.), *On further examination: Report of the SAT advisory panel on the Scholastic Aptitude Test score decline.* New York, NY: College Board.

Wirtz, W. (Ed.). (1977). *On further examination: Report of the SAT advisory panel on the Scholastic Aptitude Test score decline.* New York, NY: College Board.