Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–10
DOI: 10.3102/10769986211072308

https://journals.sagepub.com/home/jeb

What Is Actually Equated in "Test Equating"? A Didactic Note

Wim J. van der Linden
University of Twente, Enschede, the Netherlands

The current literature on test equating generally defines it as the process necessary to obtain score comparability between different test forms. The definition is in contrast with Lord's foundational paper which viewed equating as the process required to obtain comparability of measurement scale between forms. The distinction between the notions of scale and score is not trivial. The difference is explained by connecting these notions with standard statistical concepts as probability experiment, sample space, and random variable. The probability experiment underlying equating test forms with random scores immediately gives us the equating transformation as a function mapping the scale of one form into the other and thus supports the point of view taken by Lord. However, both Lord's view and the current literature appear to rely on the idea of an experiment with random examinees which implies a different notion of test scores. It is shown how an explicit choice between the two experiments is not just important for our theoretical understanding of key notions in test equating but also has important practical consequences.

Keywords: *measurement scale; probability experiment; random variable; sample space; test equating; test score*

Introduction

The process of equating two different forms of the same test is often referred to briefly as "test equating." The custom obviously is an example of casual use of language, as test forms are physical entities left untouched during the process of equating. Casual language is unproblematic as long as all parties are aware of the actual meaning of the words used in their communications. But this is precisely where the current literature on test equating seems to disagree with a foundational paper published some 70 years ago.

One of the frequently cited introductory texts to test equating defines it as

> *a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. (Kolen & Brennan, 2014, p. 2)*

The definition appears to reflect a common view in the literature. Almost exactly the same words are used, for instance, in the introductions by Braun and Holland (1982, p. 12), Holland and Dorans (2006, p. 187), and in an opening quote in von Davier et al. (2004, p. 1). And the same idea of score comparability appears to be the point of departure of nearly every recent research paper on the subject of test equating.

However, in an early introduction to the problem of test equating by Frederic M. Lord, one of the founding fathers of modern test theory, we meet a different view of what is actually adjusted during an equating. Taking the frequency distributions of scores on two different test forms as his point of departure, this author used the following analogy to explain the process of equating:

> *We may imagine each of the original two frequency distributions of scores to be drawn on a perfectly elastic surface. It will then always be possible to stretch and compress the scale (as it is drawn on the surface) of either one of the distributions in such a way that this distribution will become identical with the distribution on the other test. This stretching and compressing of the score scale of measurement transforms it into a new converted-score scale of measurement. The converted-score scale on the first test and the unconverted score scale on the second test will now be "comparable." (Lord, 1950, p. 2)*

Keywords in these two apparently different views of equating are the *scores* assigned to examinees and the *scale* on which they are scored. The more recent view focuses on an adjustment of the former whereas Lord's introduction requires an adjustment of the latter. The only exception known to the author are González and Wiberg (2017, section 1.2.3) whose statistical foundation of the equating problem supports Lord's view. Also, observe that Lord's quote points at identical score *distributions* on both test forms as the desired effect of his comparability of scale, not at comparable scores.

From a statistical point of view, the notions of scale, score, and score distribution are fundamentally different. The purpose of this note is to analyze these differences in more detail and show how they relate to the problem of test equating. The results of the analysis reveal that the view offered by Lord some 70 years ago appears to be more in line with standard statistics than the current literature on equating. As for the notion of a score distribution, both views appear to rely on a probability experiment with distributions different from what follows from test theory.

Basic Statistical Concepts

Test equating, as most of test theory, is indebted to statistics for its main concepts and methods. It is thus no surprise that notions as scale, score, and score distribution have their parallels in statistics. In fact, with different names, they underly all of statistics. In the following, we analyze these parallels; for a

more technical introduction to the statistical concepts used in our analysis, Casella and Berger (2002, chap. 1) is recommended.

A typical statistical treatment begins with the explicit definition of a *probability experiment* in the outcomes of which we are interested. Examples of such experiments are throwing a die, watching a tennis match, recording the effect of a new pharmaceutical drug, and, indeed, observing an examinee taking a test. As an illustration, we first consider the well-known case of throwing a die. The set of all possible outcomes of the throw—that is, the set of integers $\{1, 2, 3, 4, 5, 6\}$—is known as the *sample space* of the experiment. When the experiment is conducted, we do not observe the entire sample space, only one of its elements, for example, integer 4. However, due to its random nature, if the experiment is replicated (i.e., independently repeated) a number of times, different outcomes occur. If it is replicated a large number of times, the relative frequency of each of the possible outcomes stabilizes, a phenomenon which, for the case of a discrete sample space, is explained by the laws of large numbers. The limiting numbers to which these relative frequencies converge can be treated as *probabilities*. The larger the probability of an element of the sample space, the more frequently it will tend to be observed.

If the experiment can be described in this probabilistic way, statistical analysis becomes possible. Typically, we then begin by deriving a model for the *probability distribution* of the outcomes from the nature of the experiment. For example, if the assumption of a fair die holds, the observed outcomes are from a uniform distribution. If the die is not fair, the distribution must be a member of the multinomial family with unequal probabilities. And if it is replaced with a fair coin, the distribution is know to be binomial.

The importance of the distinctions between all these elementary concepts is reflected in their notation. The sample space for the die experiment is the set $S = \{1, 2, 3, 4, 5, 6\}$, with arbitrary elements. It is important to be aware of the nature of x as a mathematical variable running over the numbers 1 through 6 in the sample space. Specifically, there is nothing probabilistic about x; its sole purpose is to represent each of the possible outcomes of the experiment as defined by the number of dots on the sides of the die. However, the notation becomes fundamentally different when the die is actually thrown and one of the possible outcomes is observed. We then use capital X to represent the observed outcome from the experiment. If the outcome happens to be 2, the observation is recorded as $X = 2$. This new variable X is an example of a *random variable*; that is, a variable with a probability distribution over the sample space. For an arbitrary outcome, the observation is denoted as $X = x$ and we refer to X taking the value of x as a *realization* of random variable X. The difference between capitals and lower cases is thus fundamental. Without it, we would be unable to make any distinction between a mathematical representation of the sample space of an experiment and a statistical analysis of its observed outcomes.

The notation for the probability distribution associated with X needs a little more explanation. For technical reasons, statistics operates with the cumulative distribution function, which is denoted a $F_X(x)$. The argument of the function reflects the fact that it runs over the sample space; for each value of x, it gives us the probability of observing $X \leq x$ during a replication of the experiment. Although the function thus has a known mathematical variable x as argument, it is not yet identified; it is always possible to meet another experiment with a different probability distribution over the same sample space (as a case in point, consider the examples of throwing a fair and unfair coin). Hence, the use of X as index in $F_X(x)$ to identify the proper function.

Probability Experiment of Test Taking

We now consider the experiment of an arbitrary examinee p taking a test where our interest is in the number of correct responses. The experiment served as the point of departure of Novick's (1966) early axiomatic formalization of classical test theory shortly thereafter included in Lord and Novick (1968, section 2.2).

For a test of n items, the experiment has sample space $S = \{1, 2, ..., n\}$.[1] Due to inherent unreliability of the test (replicated administration is unlikely to result in exactly the same score for each examinee), the experiment must also be treated as probabilistic. For an arbitrary examinee p, the experiment thus implies a random variable X_p with realization $X_p = x$. Observe that the left-hand side of this equation has index p but the right-hand side has not. For each examinee, a distinct random variable X_p with its own probability distribution is required (for instance, to reflect the fact that a more able examinee has a distribution more toward the upper end of the sample space than someone who appears to be less able). But the sample space itself is the same for all examinees. For a group of examinees that takes a test, we thus have as many random variables and distributions as examinees. Each of these distributions is known to be a different member of the family of compound binomial distributions. Unlike its special case of the binomial family, which holds for the earlier experiment of the number of heads observed when tossing a single coin repeatedly, the family does not have an explicit mathematical expression for its distribution function $F_{X_p}(x)$, but we do know how to calculate the probabilities for each possible value of x (Lord & Wingersky, 1984).

It is now clear that the idea of a score scale in test theory (in test equating: the number-correct scale) is exactly the same as that of a sample space in statistics. It refers to the entire set of possible outcomes of the test; nothing more, nothing less. As for the idea of a test score, for an arbitrary examinee p, this is just a random variable X_p with a different probability distribution $F_{X_p}(x)$ over the same scale for each examinee (sometimes also referred to as the support of the distribution, e.g.,

Casella & Berger, 2002, section 2.1.). The result actually observed when an examinee does take the test is one realization $X_p = x$ of their random variable.

Although straightforward, the match between all these concepts is easily blurred due to unfortunate connotations of the terms "observed score" and "observed-score distribution" in use throughout test theory. The former is easily taken to suggest the number of items correct actually observed when an examinee takes a test. However, as just highlighted, the observed score of an examinee p is a random variable X_p with possible realizations ranging over the entire scale, as opposed to the *true score,* which is its fixed expected value (e.g., Lord & Novick, 1968, section 2.3). What actually is observed for the examinee in a test administration is one *realized* observed score.

The term "observed score," when taken at face value, is thus potentially misleading. In fact, when its actual meaning is ignored, it is a small step to take "observed-score distribution" to refer to the frequency distribution of the numbers of items correct actually obtained for the entire group of examinees in the test administration. However, the term should be used for the distributions of the random variables X_p for each of the examinees, *not* for an empirical distribution of the single realizations recorded for each of them as in Lord's quote. The best way to keep track of all necessary distinctions is to remember that, paradoxically, the observed score of an examinee is never observed, only one of its possible values is.

Experiment of Test Equating

It is now time to apply the standard statistical terminology and notation to the problem of test equating. The first step is to realize that the problem actually involves two distinct probability experiments, one for an old test form X and the other for a new form Y, both assumed to measure the same ability. (We use Latin capitals to denote test forms and italics for the random variables associated with them.) When we equate back in time, only the experiment for the new form is conducted. The experiment for the old form is hypothetical; it would have been conducted if the new examinees had taken the old form as well (but then, obviously, no equating problem would have been left).

For notational simplicity, consider the case of forms with the same number of items, n. The experiment for the old form has sample space $S_X = \{1, 2, ..., n\}$ with an arbitrary element x; for the new form, it is sample space $S_Y = \{1, 2, ..., n\}$ with an arbitrary element y. Although both spaces are represented by the same set of integers, their similarity is only nominal. Unlike our earlier example where x represented the number of dots on the sides of a single die, x and y now refer to the performances of examinees on two sets of items with *different* properties (otherwise, again, we would have no equating problem left). As we have two different sample spaces for each examinee p, we also have two random variables X_p and Y_p with distinct distribution functions $F_{X_p}(x)$ and $F_{Y_p}(y)$. The equating problem is to find the transformation that, for each of the examinees,

makes the sample space, random variable, and distribution function for form Y identical to those for X (or, using a more abstract, measure-theoretic concept, maps the *probability space* for Y onto the one for X; see, e.g., Capiński & Kopp, 2004, section 2.6.1). The same definition of a test equating transformation as a mapping between the two sample spaces is found in González and Wiberg (2017, section 1.2.3).

The necessary transformation is found setting the two distribution functions equal to each other for each p. Thus, setting

$$F_{X_p}(x) = F_{Y_p}(y), \tag{1}$$

and making x explicit, we obtain it as

$$x = F_{X_p}^{-1}(F_{Y_p}(y)). \tag{2}$$

Actually, the dependence of the transformation on p is unnecessarily restrictive. A standard assumption underlying all of test equating is a common ability measured by the number-correct scores for the two test forms. As the other characteristics defining the identity of the examinees can be ignored, Equation 2 can therefore be replaced with

$$x = F_{X|\theta}^{-1}(F_{Y|\theta}(y)), \tag{3}$$

where $F_{X|\theta}(x)$ and $F_{Y|\theta}(y)$ are the distribution functions for the scores of arbitrary examinees on test forms X and Y with common ability level θ.

The equating transformation just derived is thus a mathematical function from scale y of the new test form to scale x of the old form, exactly as claimed by Lord (1950). Distribution functions $F_{X|\theta}(x)$ and $F_{Y|\theta}(y)$ serve as parameters of the function identifying the transformations for examinees with different ability levels in the equating study.

Alternative Experiment of Test Equating

As already noted, though Lord and the later equating literature differ in their view of the equating transformation, they agree on a notion of test scores different from our earlier statistical definition. The reason resides in a different probability experiment assumed to represent the problem of test equating. As the choice of experiment is the fundament upon which all of statistical modeling rest, the difference is not inconsequential.

To the author's knowledge, the first explicit definition of the probability experiment currently present throughout the equating literature is the one in Lord (1982, p. 165), where it was adopted to derive an asymptotic standard error of equipercentile equating. The experiment is as follows: A test form X has been administered to a group of examinees randomly sampled from some specified population. A different form Y has just been administered to a separate random sample from the same population. The problem is to equate the scale of Y back to

the scale of X. Observe that the experiment for form X is no longer hypothetical, an *actual* sample of examinees is now supposed to have taken the form. The two experiments, with additional specifications to allow for different data collection designs, have been the automatic point of departure in the equating literature ever since. (Lord already considered the case of a single random sample taking two different forms. Because of its similar assumption of random sampling of examinees, it is not further considered here.)

The alternative choice of experiment immediately leads to different random variables for the scores of the examinees. Generally, data obtained through a simple random sample of size n from a specified population imply a model with random variables $Z_1, ..., Z_n$, one for each observation but with a common distribution $F_Z(z)$. As all draws are made independently, the case is known as the one of independent and identically distributed (iid) variables (e.g., Casella & Berger, 2002, section 5.1). For equal sample sizes, the experiment introduced by Lord implies thus one set of iid variables with distribution $F_X(x)$ for the group that took form X and another set $Y_1, ..., Y_n$ with distribution $F_Y(y)$ for the group with form Y. Our earlier experiment, however, though defined on the same two sample spaces, implies the alternative case of one set of *non-iid* variables X_p each with a different distribution $F_{X_p}(x)$ and another set with different distributions $F_{Y_p}(y)$. It is important to observe the difference in interpretation between the indices of the random variables in the two experiments. For the earlier experiment, the indices are to identify each of the individual examinees in the equating study; for the experiment with iid variables, they are just used to distinguish between the random draws that constitute the two samples.

As the two experiments differ in their choice of the source of randomness present in test equating, they can be labeled as experiments with *random scores* versus *random examinees.* The former clearly is the experiment of choice when the interest is in test scores of individual examinees with their less than perfect reliability, for instance, as they are used to make an admission or selection decision for each of them. Experiments with random subjects typically occur in such areas as opinion polling, survey analysis, and marketing research with their interest in quantities as population means, quantiles, standard deviations, and so on. (Just for the sake of completeness, experiments both with random examinees and test scores do exist, for instance, in studies of group-based assessment of educational progress. They involve actual sampling of students from well-defined educational populations with scores treated as random too. Their statistics involves the use of two-level models, with a separate level accounting for each source of randomness. However, group-based educational assessment is not the typical application addressed in the observed-score equating literature.)

Most likely, Lord's (1982) choice of experiment just followed the statistical tradition of his day. All introductory texts used in Statistics 1.01 in the educational and behavioral sciences programs at the time, explicitly or more implicitly,

motivated the use of statistics assuming the case of random sampling with its iid variables (in fact, most of them still do!). Nevertheless, his choice still is a bit of a puzzle. As demonstrated by the careful presentation of the probability experiment of test taking in the introduction to classical test theory in Lord and Novick (1968, section 2.2), he must have been aware of the differences between the two alternative experiments.

Practical Consequences

The choice of a probability experiment of test taking, with its subsequent implications for the treatment of test scores as random variables and the specification of their distribution, is not without practical consequences. Two of these consequences, the choice of scale transformation and the standard error of the equated scores, are briefly discussed.

The scale transformation in Equations 2 and 3 was derived for the equating experiment with non-iid random variables for the two groups of examinees. For the alternative experiment, setting its two distributions equal

$$F_X(x) = F_Y(y), \tag{4}$$

and making x explicit, it follows as

$$x = F_Y^{-1}(F_X(x)). \tag{5}$$

The result is one common transformation for all examinees. The transformation is appropriate when the interest is in an "average equating" for examinees sampled from a "synthetic population." It is less so when we are concerned with the equated scores for each of the examinees that actually took form Y.

It may look difficult to estimate the earlier transformations with the conditional score distributions in Equation 3, but it is not. The available options include item-response theory (IRT) observed-score equating, the use of the conditional observed-score distributions on X and Y given the score on anchor items in the two forms, conditioning on other proxies for the common ability measured by the two forms, single-group equating, or test forms assembled with preequated number-correct scales. For a review of these examples, performed under the name of local equating, see González and Wiberg (2017, chap. 6) and van der Linden (2011, 2018). One of the advantages capitalized on in each of the examples is the absence of the necessity to reconcile the differences in ability distribution between the two groups of examinees that took form X and Y, a problem inherent in the experiment with random examinees generally resolved by performing the equating in Equation 5 adopting arbitrary sampling weights for an assumed synthetic population. Also, for each of these examples, the assumption of a common ability measured by the two forms suffices. Specifically, it is not necessary to have forms of equal length, equal reliability, or to be concerned about the requirement of equity (van der Linden, 2019).

It is required statistical practice to report estimated quantities along with measures of their accuracy. Lord's (1982) asymptotic standard error of equating serves the goal for an equating based on the experiment with random examinees. However, if the interest is in equated scores for the individual examinees that took form Y, the experiment of random scores applies and a different standard error is required for examinees with different levels of ability to account for the differences in random error between their equated scores. The choice between the two is completely analogous to the one between a constant error of measurement for all examinees and conditional errors given their level of ability discussed extensively in the history of testing. The required standard errors of equating for the experiment with random scores could be calculated similarly to Lord (1980), replacing his marginal distributions of the scores X and Y on the two forms by their conditional distributions given the common ability underlying the forms.

Conclusion

It thus appears to be a statistical fact that Lord's answer to the question of what actually is equated in "test equating" is correct. Just as claimed in his quote, any proper equating transformation is directly from the scale of one form of a test to the scale for another. As for the role of observed-score distributions in the equating, the situation is different though. The distributions in the earlier quote from Lord (1950) were the empirical frequency distributions of realized scores collected for two test forms in the equating study. In his later derivation of the asymptotic standard error of equipercentile equating, they are random variables with identical distributions for each of the examinees that took the two forms. The same assumption is made throughout the current equating literature.

Both the notions of scale and observed-score distribution deserve to be reconnected with the basic statistical concepts of probability experiment, sample space, and random variable introduced in every modern introduction to statistics. Possible reasons for the current lack of connection might be the tradition in the educational and behavioral sciences of introductory texts motivating the use of statistics only by considering the case of random sampling with iid variables as well as the possible confusion created by the connotation of the terms "observed score" and "observed-score distribution" discussed earlier in this note.

Note

1. To account for existing differences between test items, the statistical approach should actually begin with the modeling of test taking as a set of n experiments, one for each item with sample space $\{0, 1\}$. As the interest in the equating literature is almost exclusively in number-correct scoring, this more fundamental first step is omitted here.

References

Braun, H., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holl & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Capiński, M., & Kopp, E. (2004). *Measure, integral and probability* (2nd ed.). New York: Springer.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

González, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. New York: Springer.

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.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (Research Bulletin 50-48). Princeton, NJ: Educational Testing Services.

Lord, F. M. (1980). *Applications of item response theory*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, *7*, 165–174.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *8*, 452–461.

Novick, M. R. (1966). The axioms and principe results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.

van der Linden, W. J. (2011). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking* (pp. 201–223). New York: Springer.

van der Linden, W. J. (2018). IRT observed-score equating. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 3: Applications* (pp. 143–164). Boca Raton, FL: Chapman & Hall/CRC.

van der Linden, W. J. (2019). Lord's equity theorem revisited. *Journal of Educational and Behavioral Statistics*, *44*, 415–430.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Author

WIM J. VAN DER LINDEN is professor emeritus of measurement and data analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; email: wjvdlinden@outlook.com. His interests include test theory, applied statistics, and research methods.

Manuscript received May 31, 2021
Revision received October 26, 2021
Accepted December 11, 2021