Research Report

Uncommon Measures Revisited

ETS RR–20-04

Neil J. Dorans

December 2020

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Uncommon Measures Revisited

Neil J. Dorans

Educational Testing Service, Princeton, NJ

This report, which is based on an invited presentation given at the 2015 meeting of the Association of Test Publishers, is a response to the continuing proliferation of scale linking studies that have occurred since the publication of *Uncommon Measures* in 1999. The report has four parts. First, I restate the conclusions made in *Uncommon Measures* about linking the scales of state assessments to the National Assessment of Educational Progress scale and summarize points made by Thissen with respect to such linkages. Then I reiterate the important role played by the features of testing situations on the type of linkages made by Kolen and note how these features interact with the taxonomy of score linkages provided by Holland and Dorans. Next, I summarize findings from a 2010 National Council on Measurement in Education symposium that described the linkage studies conducted in 2008 to update the concordances between the 2005 version of the *SAT*® test and the ACT, and I discuss their implications for linking score scales in general. Finally, I offer some concluding advice pertaining to linkages in general.

Keywords Linking; scale aligning; concordance; NAEP; SAT; ACT

doi:10.1002/ets2.12287

Scores are the end products of assessment processes that focus on measurement of individual test takers. Scores are used for admissions, placement, diagnosis, and a variety of other purposes. Their properties are sometimes misunderstood, taken for granted, ignored, or presumed to be something they are not. Yet they have an impact on decisions that affect individuals and institutions. Scores from different assessments are often treated as if they were interchangeable, even when they are not. The urge to make comparisons compels people—even those who should know better—to forget that each assessment is a tool crafted for a specific purpose.

When can scores from different assessments be viewed as the same, essentially the same, pretty much the same, sort of the same, hardly the same, or the same in name only? Dorans et al. (2007) attempted to answer that question from the perspective of a linking taxonomy suggested by Holland and Dorans (2006). The book provides guidance about how to answer the question in practice. It examines linking issues ranging from the relatively easy task of producing interchangeable scores on alternative versions of a test to the daunting challenge of aligning state standards with the National Assessment of Educational Progress (NAEP) scales. This report revisits the reasons why Feuer et al. (1999) chose *Uncommon Measures* as their title.

Uncommon Measures and Linkages to National Assessment of Educational Progress

Late last century, the National Research Council published *Uncommon Measures*. The motivation for that report was the debate in the late 1990s between those who favored voluntary national tests, which President William Clinton proposed in his 1997 State of the Union Address, as a means of assessing the educational progress of students across the nation and those who believed that statistical linkages among existing tests could be used to achieve that purpose. The volume examined the feasibility of linking the results of commercial and state tests to compare one student's achievement with national and international benchmarks as well as with those of students in other places.

Four major conclusions were reached:

1. Comparing the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale, is not feasible. (p. 4)
2. Reporting individual student scores from the full array of state and commercial achievement tests on the NAEP scale and transforming individual scores on these various tests and assessments into the NAEP achievement levels are not feasible. (p. 4)

Corresponding author: N. J. Dorans, E-mail: ndorans@ets.org

3. Under limited conditions it may be possible to calculate a linkage between two tests, but multiple factors affect the validity of inferences drawn from the linked scores. These factors include the content, format, and margins of error of the tests; the intended and actual uses of the tests; and the consequences attached to the results of the tests. When tests differ on any of these factors, some limited interpretations of the linked results may be defensible while others would not. (p. 5)
4. Links between most existing tests and NAEP, for the purpose of reporting individual students' scores on the NAEP scale and in terms of the NAEP achievement levels, will be problematic. Unless the test to be linked to NAEP is very similar to NAEP in content, format, and uses, the resulting linkage is likely to be unstable and potentially misleading. (p. 5)

The first conclusion deals with linking score scales from different assessments to a single scale, such as the NAEP scale. The second conclusion follows from the first conclusion, basically restating the first conclusion, which is about linking score scales, in terms of comparing individual students who receive scores on these linked score scales. The third conclusion notes that there may be special cases where linkages are possible but that the validity of inferences based on these linkages would depend on many factors. The fourth conclusion states that a problematic linkage between an existing test scale and the NAEP scale implies that comparing scores or achievement levels could be misleading.

Thissen (2007) noted that for several decades, surveys of educational achievement that provide aggregate results, for example, NAEP and the Trends in International Mathematics and Science Study, have been linked to each other and have been linked to other assessments. The desire to make international comparisons with national standards, to report statistics for schools or districts on the NAEP scale, and to compare state performance standards have been cited as reasons for linking.

Thissen (2007) reviewed several studies involving linkage with NAEP, discussed the procedures used, and then examined some questions that have arisen about the validity of the results. Thissen concluded that statistical procedures for accomplishing such linkages had advanced considerably. He noted, however, that the continuing use of linkages between disparate assessments had not made the results easier to rely on or even to interpret. He believed that the most difficult problem to solve may be variation in the unknown level of motivation with which students approach any standardized test or survey that lacks direct personal consequence. He noted that the stability of test linkages over time or populations is an empirical question and that it is important to observe the extent to which the relationships between scores on the two tests being linked are constant or variable over time or between groups.

Subsequent to *Uncommon Measures*, research efforts focused on the development and application of statistical indices that empirically evaluate the quality of scale linkages. Dorans and Holland (2000) initiated a compilation of subpopulation sensitivity values across different types of test pairs. Holland (2005) expanded the Dorans and Holland compilation in a chapter subtitled "What Has Happened Since *Uncommon Measures*?" A special issue of the *Journal of Educational Measurement* was dedicated to subpopulation linking sensitivity (Dorans, 2004). A special issue of *Applied Psychological Measurement* also examined subpopulation invariance of linking functions (von Davier & Liu, 2008).

In addition, Dorans et al. (2007) edited a volume that examined different kinds of linking functions. Kolen (2007) and Kolen and Brennan (2004) asserted that the quality of linking different assessments is affected by three key features of the test administrations: similarity of test content, similarity of conditions of measurement, and similarity of examinee population. These three key features work together to define what is actually measured by the test.

How do these features of testing conditions mesh when a state assessment is compared to a survey? First, it is very important to look beyond score titles, one of the points raised by Braun and Mislevy (2005) in their critique of the intuitive test theory that is often employed by users of tests and test scores. This requires looking at the blueprints used by the assessments that are being linked. It is also important to look at instructions, scoring, and other conditions of measurement. As Thissen (2007) noted, it is also important to look at the target populations and potential difference in the motivation of test takers who respond to different assessments.

A Linking Framework

Holland and Dorans (2006) provided a framework for linking that considered key aspects of the measurement process, including similarity of constructs assessed. They divided linking into three basic categories called predicting, scale aligning, and equating. It is important to distinguish between these categories because they are often seen as similar or identical when, in fact, they are not. The distinctions between different types of linkages are related to the similarities and differences between the tests as well as to the ways in which the linked scores are intended to be used and interpreted. The terms used for the various methods of score linking should have restricted meanings so that they may be used with precision. Understanding the distinctions among these categories can prevent violations of professional practice.

This report focuses on scale alignment. The goal of scale aligning is to transform the scores from two different tests onto a common scale. Scale aligning has many subcategories, including activities such as *battery scaling* (Kolen, 2004), *anchor scaling* (Holland & Dorans, 2006), *vertical scaling* (Harris, 2007; Kolen & Brennan, 2004; Patz & Yao, 2007; Yen, 2007), *calibration* (Holland & Dorans, 2006), and *concordance* (Pommerich & Dorans, 2004). The subcategories of scaling form a continuum starting with situations where the tests measure different constructs all the way to those where the tests measure similar constructs. The taxonomy also takes into account similarity of reliability and difference in populations of testing takers taking the test.

Linking group-level scores to individual-level scores is a scale alignment approach that presents a vexing set of challenges. Linkages of state assessments to the NAEP scale probably involve different constructs, different populations, and tests of different reliability administered to test takers with different motivations. These are less than ideal linking conditions. See Braun and Qian (2007), Ho and Haertel (2007), and Koretz (2007) for a discussion of these issues.

Concordances Between SAT and ACT

In contrast to the linking of state assessment to the NAEP scale, which is like climbing a vertical cliff, developing concordances between ACT scales and scales from the *SAT®* test is like hiking a hilly terrain. Concordances are links between tests that are designed for similar uses, measure similar constructs, and are administered under similar testing conditions to similar populations in which the test scores exhibit similar difficulty and yield similarly high reliabilities in a common population (Holland & Dorans, 2006, p. 190, figure 6.2; p. 193). The SAT and ACT are two tests that produce scores that may be concorded.

Dorans et al. (1997) reported on the results of a collaborative study between ETS staff, on behalf of the College Board, and ACT staff who linked the 1994 version of the SAT to the ACT. A convenience sample of more than 100,000 scores on both ACT and SAT was used to establish concordances between ACT and SAT scores. Dorans (1999) delved into the issues associated with this concordance in greater depth. Pommerich et al. (2004) also examined issues associated with concordances in a special issue of *Applied Psychological Measurement* (Pommerich & Dorans, 2004).

Dorans (1999) noted that construct similarity is very important to examine. Content evaluation is a first step. It is also important to evaluate the empirical relationship between the tests to be linked. In addition, the correlation between scores should be high, and subgroup differences on the two measures should be in accord with each other.

When the SAT was altered in 2005, the College Board and ACT cooperated by agreeing to supply scores, administration dates, and gender and ethnic/racial data from their 2006 cohorts to a third party for matching to produce a joint cohort that was the intersection of the SAT and ACT cohorts. Concordances between ACT and SAT scores and composites were computed on this joint cohort. A series of collaborative studies between the College Board and ETS staff was reported on at the 2010 National Council on Measurement in Education annual meeting. (New concordances have been developed between the ACT and the SAT revision of 2016). What follows is dated but amply illustrative of the issues involved in such linkages.

Table 1, based on SAT and ACT Math scores from the 2006 joint ACT – SAT cohort, demonstrates the need for concordances. Comparisons of scores at the same percentile rank in the SAT and ACT cohorts would be misleading because the 2006 SAT cohort was more able than the 2006 ACT cohort, as can be seen when comparing the 75th, 50th, and 25th percentiles for Math scores in each of these cohorts to their percentile rank in the joint cohort. In the joint cohort, a score of 600 on the SAT was at the same percentile rank (68.4) as a score of 26 on the ACT. In the SAT cohort, the 600 was at the 75th percentile, while a 24 was at the 75th percentile in the ACT cohort. In the joint cohort, a 24 (75% in the ACT cohort) on ACT concorded to a 560 on the SAT, which is 40 points lower than a 600 (75% in the SAT cohort).

In that 2010 National Council on Measurement in Education symposium, Dorans and Petersen (2010) asked the following question: To what extent do concordances achieve score interchangeability? They answered this question by reviewing the five critical requirements for equating and indicated the extent to which concordances meet these requirements. While judgmental analyses of the content of the tests that produce concorded scores is an important consideration, they focused on empirical matters, such as the important role that the correlation between scores plays in evaluating whether the scale

Table 1 ACT–SAT Scale Linkages Across Three Distinct Cohorts

Percentile in SAT cohort	SAT M score	Percentile SAT/ACT cohort	ACT M score	Percentile in ACT cohort
75	600	68.4	26	
	560	58.5	24	75
50	520	41.6	22	
	480	31.6	20	50
25	440	20.0	18	
	400	10.8	16	25

Note. ACT M = ACT Math; SAT M = SAT Math.

alignment provided by a concordance can be sensibly interpreted. They maintained that indices of subpopulation invariance are also important in distinguishing concordances from equatings. They also provided an overview of the types of issues that must be considered when examining the concordance between scores.

Moses et al. (2010) examined an important challenge with any ACT–SAT concordances, namely, that the cohort on which they are developed (i.e., examinees who took the ACT and the redesigned SAT tests) will differ from the cohorts for which they are intended (i.e., examinees who took either the ACT or the redesigned SAT test, but not both). Because concordances are population dependent, an important consideration is how to adequately represent the "either–or" examinee groups within the "ACT and SAT" data. A related consideration is the role of testing order in the concordance data, as examinees who took the ACT before the SAT may differ from those who took the SAT before the ACT. Test takers who took both tests were divided into two groups. The SAT-first group took the SAT before they took the ACT; the ACT-first group sat for the ACT before they took the SAT. Moses et al. (2010) described evaluations of the effects of testing order on linking functions and the extent to which the concordance sample represented the target cohort of examinees who took either the ACT or the redesigned SAT test. The evaluation of order effects on linking functions suggested that the SAT-first subsample appeared to represent the intended ACT-only and SAT-only populations more closely than the ACT-first subsample. Differences between concordances computed from the complete ACT–SAT cohort data and from the SAT-first subsample were deemed small enough to support the use of the entire ACT–SAT cohort data rather than the more representative SAT-first subsample.

The sample of examinees used to produce the ACT–SAT concordances contained data from many examinees who took the ACT and the revised SAT tests more than once. Dorans et al. (2010) described evaluations of the ACT and SAT scores that informed the ACT–SAT concordances, including examinees' first ACT and SAT scores, their most recent ACT and SAT scores, their ACT and SAT scores that were closest in time, and averages of all their ACT and SAT scores. From an empirical reduction of uncertainty/correlation perspective, the average score was deemed superior to the other three pairs of scores examined. These correlational analyses suggested that ACT–SAT concordances would be of highest psychometric quality when these concordances are based on averages of the cohort sample's ACT and SAT scores.

These three papers addressed issues pertaining to the definition of the analysis sample, the assessment of order effects, the determination of which test scores to concord, and the determination of which pairs of multiple pairs of ACT–SAT scores to use. A decision was made to use all data regardless of order and to use the average of multiple pairs of SAT scores.

The ACT–SAT concordances were computed using the equipercentile linking method applied to the raw, unsmoothed examinee data. The equipercentile method defined the ACT–SAT concordances as the set of SAT scores that have percentiles that are equal to the percentiles of the ACT scores. For the two concordances with the ACT Composite (ACT C), the ACT Sum (English + Math + Reading + Science Reasoning) score was used in the equipercentile computations. The ACT Sum was used rather than the ACT C in the equipercentile computations because the ACT Sum contained a larger range of possible integer scores (4–144) than the ACT C (1–36), which increased the precision of the equipercentile computations. These concordances of the ACT Sum scores were converted into concordance tables for ACT C scores by dividing the ACT Sum by 4 to produce scores that ranged from 1 to 36 in intervals of 0.25. The linking function used to concord ACT C scores to SAT Sum scores involved only the integer values of the ACT C scores because noninteger values of the average of the four ACT scores are rounded up to the nearest integer to produce the ACT C score.

An important gauge of the quality of the ACT–SAT concordances is how representative they are for examinees of different genders, races, and ethnicities. Liu et al. (2010) examined subpopulation sensitivity by comparing the ACT–SAT concordances computed from the complete ACT–SAT population to those computed using data from

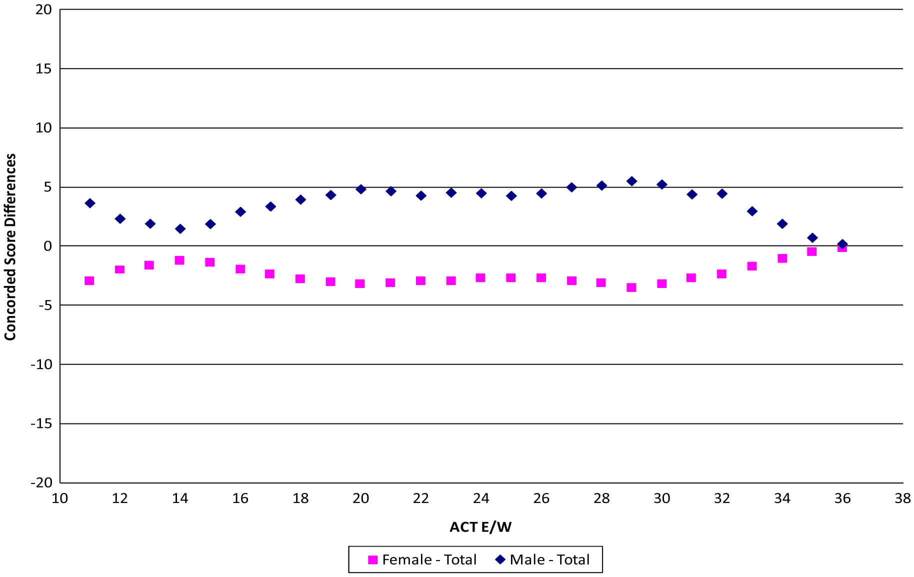

Figure 1 Differences in Concordances Between ACT English/Writing to SAT Writing.

specific subpopulations. Standard measures of subpopulation invariance (Dorans, 2004; Dorans & Holland, 2000; Liu & Dorans, 2013) were used. The results showed that concordances of the ACT–SAT subtests were more invariant across subpopulations than concordances of ACT–SAT composite scores. A subset of those results is reported in the following pages.

In theory, subpopulation invariance is a necessary but not a sufficient condition for an equating relationship. In practice, subpopulation-free linking is impossible (Holland & Dorans, 2006; Lord, 1980). Even so, equating relationships should be essentially subpopulation invariant. Concordances are expected to be more subpopulation sensitive than score equatings. To assess subpopulation invariance, a standard practice is to compare the linking relationships of subpopulations to the linking relationship in the total population.

The concordance conducted in the total group to produce a total group concordance function is denoted by $s_P\,(x)$. Concordances are also produced for each subpopulation of interest, $s_{P_g}(x)$. Then differences between the subpopulation and total group concordance functions for each subpopulation are computed, $s_{P_g}(x) - s_P\,(x)$.

It is informative to look at difference plots of concordances to compare a subpopulation concordance to the total population concordance at each score level. For example, Figure 1 is the plot for ACT English/Writing (ACT E/W) to SAT Writing (SAT W) that was reported in Liu et al. (2010).

Note that for Writing, the conversion for ACT to SAT for males is slightly higher than the total group conversion, while the conversion for females is lower than the total group conversion. Given that the SAT scores are reported in 10-point increments from 200 to 800, differences of less than 5 points are not considered problematic (Liu & Dorans, 2013). The results for the concordance between SAT Math and ACT Math were even more invariant (Dorans, 2015).

Two simple summary statistics can be used to average the differences in concordances seen in these plots. The root expected square difference (RESD) is an index that summarizes how close each subpopulation's concordance function is to the total population concordance function. For group g, the RESD is

$$\mathrm{RESD}_g = \sqrt{\sum_x f_{gx}\left[s_{P_g}(x) - s_P\,(x)\right]^2},$$

where f_{gx} is the relative frequency of test scores at score level x in subgroup g. The mean difference (MD) between the subgroup concordances and the total group concordances is computed via

$$\mathrm{MD}_g = \sum_x f_{gx}\left[s_{P_g}(x) - s_P\,(x)\right].$$

For the Writing concordances, these summary statistics were less than 5 scale-score points (Table 2).

Table 2 Average Differences Between Subpopulation and Population Concordances Linking ACT E/W to SAT W

	MD	RESD
Total	0.00	0.00
Female	−2.74	2.79
Male	4.22	4.31

Note. ACT E/W = ACT English/Writing; MD = mean difference; RESD = root expected square difference; SAT W = SAT Writing.

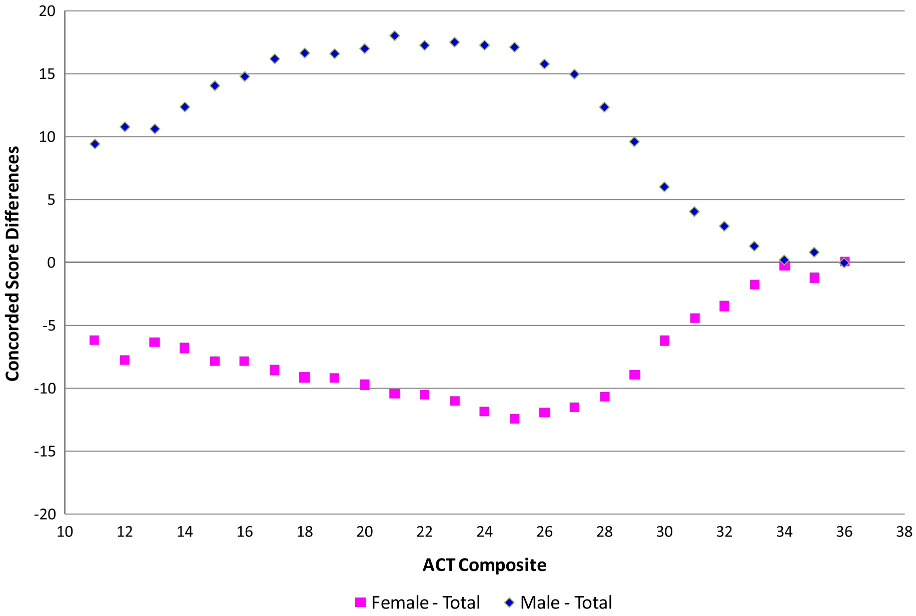

Figure 2 Differences in Concordances Between ACT Composite to SAT Critical Reading + Math.

Table 3 Two Average Differences Between Subpopulation and Population Concordances Linking ACT C to SAT CR + M

	MD	RESD
Total	0.00	0.00
Female	−9.71	9.97
Male	14.32	15.03

Note. ACT C = ACT Composite; MD = mean difference; RESD = root expected square difference; SAT CR + M = SAT Critical Reading + Math.

Differences between gender-specific concordances and the total group concordances for the composite scores are not that small, as seen in Figure 2, which appeared in Liu et al. (2010). Here the differences between the ACT C to SAT Critical Reading + Math (SAT CR + M) concordances between the male subgroup and the total group are between 10 and 20 points; differences between the female and total group are between 5 and 10 points. For the ACT C to SAT CR + M concordances, the differences in summary statistics were larger as well (Table 3).

In sum, the concordance results were mixed with respect to sensitivity to gender. The Math to Math was the most invariant with respect to gender. In that sense, it was the concorded score most like an equated score. Writing was next in line with respect to invariance. The composite concordances were more population sensitive. In the case of ACT C to SAT CR + M, the gender results suggested a sensitivity that is consistent with the different weights placed on Math scores.

In Table 4, which is based on the concordances that produced the difference plots depicted in Figure 2 and Table 3, we see what happens when a test score is linked to itself via another score and via linkages that are subpopulation dependent. Here an SAT CR + M score of 1160 is linked to a 25 on ACT C via a concordance based on males only. When concatenated with a linkage based on females only for ACT C to SAT CR + M in which a 25 is linked to an 1130, the score 1160 is concorded to a score of 1130. For this score of 1160 and the SAT CR + M score of 1130, taking SAT CR + M scores to the

Table 4 Chained Linkage of SAT CR + M to Itself Through ACT C Across Male and Female Subpopulations

Concordance (males)		Concordance (females)	
SAT CR + M	ACT C	ACT C	SAT CR + M
1160	25	25	1130
1130	24	24	1100

Note. ACT C = ACT Composite; SAT CR + Math = SAT Critical Reading + Math.

SAT CR + M scale through a concatenation of the SAT to ACT concordance based on males only with the ACT to SAT concordance based on females only suggests that SAT CR + M = SAT CR + M − 30. Linking ACT C to itself via a reserve linkage yields ACT C = ACT C + 1. (The standard deviation for ACT C is about 5 points.) Chaining concordances based on different subpopulations incorrect inferences.

ACT C and SAT Sum are uncommon measures used for similar purposes. The definition of the composite is the issue here. One-half Math (SAT CR + M) and one-quarter Math (ACT Sum/4) are not comparable for males and females as groups (nor for most individuals).

One important implication of these ACT–SAT concordances for concatenating links between scales of tests that are more challenging to link than the ACT and SAT, for example, linking state assessment scales with reliable scores obtained from mostly motivated students to scales from educational surveys such as NAEP, is that concatenating concordances across subpopulations (gender or state) is not likely to bring you back to where you started.

Implications for Other Linkages

Recent work by Reardon et al. (Reardon, Ho, & Kalogrides, 2019; Reardon, Kalogrides, & Ho, 2019) attempting to link the results of various state assessments to NAEP scales has demonstrated that efforts to link uncommon measures (estimates of state assessment performance obtained from different assessments) will continue to arise in the future. One purpose of this report is to remind potential users of the outcomes of such linkages (e.g., average test scores in The Educational Opportunity Project at Stanford University, 2019) that the issues raised in *Uncommon Measures* remain as relevant as ever.

The conclusions of *Uncommon Measures* were reached before Dorans and Holland (2000) proposed indices that could be used to check whether such linkages generalize. These conclusions have been strengthened empirically by population sensitivity studies, such as those cited in Dorans (2004) and von Davier and Liu (2008). Published studies should inform the discussion about the merit of conducting these links. As such compilations of results have expanded, we are better able to address empirically the consequences of complying with the reoccurring call to uncommon measures.

The ACT–SAT concordances summarized in this report were between very reliable scores from two testing programs that produce well-crafted tests administered in a similar mode of administration to motivated test takers from similar populations. In addition, the scores are used for similar purposes. These conditions are as good as it gets for concording. Even so, order effects exist for all measures, and subpopulation sensitivity affects the linkages between composite scores.

When linking measures that differ in reliability, differ in how the "same construct" is defined and measured, differ in how subpopulations perform on them, and differ in consequences for individuals, the concerns raised in *Uncommon Measures* and echoed by Thissen (2007), Ho and Haertel (2007), and others remain salient: A lot of things can go wrong.

Concordances and other forms of scale linkages should be made available only when the relationship between the linked score scales is strong enough to support intended inferences, when linkages are comfortably population invariant, for example, the same concordances are obtained in males and females, and when the linkage leads to improved practice. Some linkages, however, engender flawed comparisons and should be avoided as bad practice.

When does it make sense to compute the particular class of linkage called a concordance? Holland and Dorans (2006, p. 190, figure 6.2; p. 193) suggested that concordance makes sense when the tests are designed for similar uses, measure similar constructs, and are administered under similar testing conditions to similar populations in which the test scores exhibit similar difficulty and yield similarly high reliabilities. Finally, the concordance should lead to improved practices. The FRANK principles (Pommerich, 2007) can be used to guide the practice of linking scores: (a) flexibility in linking practices, (b) responsibility in creating and disseminating concordance tables, (c) awareness of the limitations of concordances, (d) notification as to proper interpretation and use of results, and (e) knowledge of users and their practices.

Acknowledgments

Adapted from an invited presentation at the Association of Test Publishers' annual conference, ATP Innovations in Testing, in Palm Springs, CA, United States, on March 3, 2015.

Note

[1] The initial examinee sample for the ACT – SAT concordances contained 303,637 test takers. These examinees took the ACT at least once from September 2004 to June 2006 and the revised SAT at least once from March 2005 to April 2006. A second examinee sample was also defined from this joint cohort of 300,437 test takers for producing a concordance of the ACT and SAT Writing tests. This subsample contained data for 190,148 test takers who took the optional ACT Writing test.

References

Braun, H. I., & Mislevy, R. J. (2005). Intuitive test theory. *Phi Delta Kappan*, *86*(7), 489 – 497. https://doi.org/10.1177/003172170508600705

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 199 – 216). Springer Science+Business Media. https://doi.org/10.1007/978-0-387-49771-6_17

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-1). College Board. https://doi.org/10.1002/j.2333-8504.1999.tb01800.x

Dorans, N. J. (Ed.). (2004). Assessing the population sensitivity of equating functions [Special issue]. *Journal of Educational Measurement*, *41*(1), 1 – 2. https://doi.org/10.1111/j.1745-3984.2004.tb01154.x

Dorans, N. J. (2015, March 1 – 5). *Uncommon measures revisited* [Paper presentation]. ATP Innovations in Testing, Palm Springs, CA, United States.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281 – 306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, *73*(2), 24 – 34. https://www.aacrao.org/docs/default-source/c-u-.pdfs/73-2---fall-1997.pdf

Dorans, N. J., & Petersen, N. S. (2010, May 1 – 3). Distinguishing concordances from equating. In T. Moses (Organizer), *Updating the ACT/SAT concordances* [Symposium]. The annual meeting of the National Council on Measurement in Education, Denver, CO, United States.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. Springer. https://doi.org/10.1007/978-0-387-49771-6

Dorans, N. J., Sconing, J., & Crouse, J. (2010, May 1 – 3). Selection decisions for the ACT and SAT scores used to produce the ACT-SAT Concordances. In T. Moses (Organizer), *Updating the ACT/SAT concordances* [Symposium]. The annual meeting of the National Council on Measurement in Education, Denver, CO, United States.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. National Academy Press.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233 – 251). Springer. https://doi.org/10.1007/978-0-387-49771-6_13

Ho, A. D., & Haertel, E. H. (2007). *(Over)-interpreting mappings of state performance standards onto the NAEP scale* [Unpublished manuscript]. https://scholar.harvard.edu/files/andrewho/files/ho_haertel_overinterpreting_mappings.pdf

Holland, P. W. (2005). Assessing the validity of test linking: What has happened since *Uncommon Measures*? In C. A. Dwyer (Ed.), *Measurement and research issues in a new accountability era* (pp. 185 – 195). Erlbaum.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187 – 220). American Council on Education/Praeger.

Kolen, M. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*(4), 219 – 226. https://doi.org/10.1177/0146621604265030

Kolen, M. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31 – 55). Springer. https://doi.org/10.1007/978-0-387-49771-6_3

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4310-4

Koretz, D. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 339 – 353). Springer. https://doi.org/10.1007/978-0-387-49771-6_18

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice*, *32*(1), 15 – 22. https://doi.org/10.1111/emip.12001

Liu, J., Dorans, N. J., & Moses, T. P. (2010, May 1 – 3). Evaluating the subpopulation sensitivity of the ACT-SAT concordances. In T. Moses (Organizer), *Updating the ACT/SAT concordances* [Symposium]. The annual meeting of the National Council on Measurement in Education, Denver, CO, United States.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Moses, T., Crouse, J., & Sconing, J. A. (2010, May 1 – 3). Selection decisions for the sample data used in the ACT-SAT concordance. In T. Moses (Organizer), *Updating the ACT/SAT concordances* [Symposium]. The annual meeting of the National Council on Measurement in Education, Denver, CO, United States.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253 – 272). Springer. https://doi.org/10.1007/978-0-387-49771-6_14

Pommerich, M. (2007). Concordance: The good, the bad, and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 199 – 216). Springer Science+Business Media. https://doi.org/10.1007/978-0-387-49771-6_11

Pommerich, M., & Dorans, N. J. (Eds.). (2004). Concordance [Special issue]. *Applied Psychological Measurement*, *28*(4). https://doi.org/10.1177/0146621604265028

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement*, *28*(4), 247 – 273. https://doi.org/10.1177/0146621604265033

Reardon, S. F., Ho, A. D., & Kalogrides, D. (2019). *Validation methods for aggregate-level test scaling: A case study mapping school district distributions to a common scale* (CEPA Working Paper No. 16-09). Center for Education Policy Analysis. https://cepa.stanford.edu/sites/default/files/wp16-09-v201904.pdf

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2019). Validation methods for aggregate-level test scaling: A case study mapping school district distributions to a common scale. *Journal of Educational and Behavioral Statistics*. Advanced online publication. https://doi.org/10.3102/1076998619874089

The Educational Opportunity Project at Stanford University. (2019). *Average test scores*. https://edopportunity.org/explorer/#/map/us/districts/avg/ses/all/3.65/41.65/-95.97

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287 – 310). Springer. https://doi.org/10.1007/978-0-387-49771-6_16

von Davier, A. A., & Liu, M. (Eds.). (2008). Population invariance [Special issue]. *Applied Psychological Measurement*, *32*(1), 9 – 10. https://doi.org/10.1177/0146621607311605

Yen, W. M. (2007). Vertical scaling and no child left behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273 – 283). Springer. https://doi.org/10.1007/978-0-387-49771-6_15

Suggested citation:

Dorans, N. J. (2020). *Uncommon measures revisited* (Research Report No. RR-20-04). Educational Testing Service. https://doi.org/10.1002/ets2.12287

Action Editor: Shelby J. Haberman

Reviewers: Yue Jia and John Mazzeo