9

Interpreting Test-Score Comparisons

Randy E. Bennett, *Educational Testing Service*

CONTENTS

INTRODUCTION

This chapter focuses on making sense from test-score comparisons. The chapter begins with some basic premises. It then proceeds to a discussion of factors that can weaken the tenability of test-score comparisons. Finally, the chapter offers some suggestions for responsibly interpreting and communicating comparisons. Much of the content draws upon ideas and examples from preceding chapters.

BASIC PREMISES

This chapter proceeds from the premise that getting meaning from assessment results inevitably requires some type of comparison. Without a benchmark or reference point, an assessment result can become an uninterpretable abstraction. To lend meaning to the results for an individual, the results may be referenced, or compared, to those of other test takers, to past performance, to the types of tasks that characterize performance at a particular score level, or to some absolute standard like a cut point indicative of broader domain proficiency.

Not only does deriving meaning from assessment results require some type of comparison, but some common comparative frame is usually needed for results to be

aggregated. That is, we cannot sensibly compute an average score for a group unless each member of that group has a result that is *comparable*.[1]

Comparisons are strongest when the same measure is given under substantively the same conditions to analogous student samples at the same point in time. In the case of comparisons of performance to an absolute standard, the similarities of conditions, student sample, and time point are with the conditions, time point, and student group assumed in setting the cut point. Comparisons become weaker as the measure, assessment conditions, student samples, or the time of administration begin to diverge. The more severe and numerous the divergences, the less defensible the comparison is likely to be.[2]

As defined above, strong comparisons will necessarily be limited to a subset of the comparisons assessment users may want to make. For that reason, it is important to identify each source of divergence and how that divergence might affect the tenability of the comparison.

WEAKER COMPARISONS

In this section, three types of divergence are briefly discussed. They are divergence due to instruments (i.e., assessments) that are nominally the same, to dissimilar instruments, and to different examinee populations.

Instruments That Are Nominally the Same

The "same" instrument can, in practice, appear in several different forms. Each of those forms can introduce divergences that weaken our ability to make comparisons. There are at least three senses in which an instrument can appear in nominally different forms. One sense is literal and refers to the presentation of examination content. For example, the Programme for International Student Assessment (PISA) delivered its 2015 science examination in 90 different language versions (OECD, 2018). Comparisons across language versions pose challenges because ideas are not always directly translatable in forms that are similar in meaning, vocabulary level, or syntactic complexity, potentially affecting the difficulty of questions (see Chapter 8, Comparability in Multilingual and Multicultural Assessment Contexts). Moreover, the same content may require relatively little text to represent it in one language but a lengthier exposition in another language, differentially affecting reading demand.

The literal form of an assessment can change via the method chosen for its delivery: paper or computer. That change may be relatively minor, as when the multiple-choice questions from a paper test are presented in similar fashion on screen. The change is more significant, however, if the online version employs item types that the paper test does not (e.g., technology-enhanced items or simulation tasks) or if the modes of

[1] Group-score assessments that use direct estimation are an exception (e.g., National Assessment of Educational Progress [NAEP] and Programme for International Student Assessment [PISA]. Such estimation is, however, not used when individuals must be awarded scores, such as on state assessments.

[2] Somewhat different considerations apply when the same student is tested repeatedly over time to measure growth, for example, through annual state assessments placed on a vertical scale. These considerations might include the types of scores compared, the effectiveness of the scaling, and the degree to which the tested constructs overlap.

response are substantially different (e.g., answering an essay question on paper versus on a computer). These more significant differences may affect the difficulty of questions and perhaps even the skills measured (Bennett, 2003; Bennett et al., 2008; Horkay, Bennett, Allen, Kaplan, & Yan, 2006).

The same instrument can also take a different literal form through the provision of accommodations for students with disabilities or for English learner students (see Chapter 6, Comparability When Assessing English Learner Students, and Chapter 7, Comparability When Assessing Individuals with Disabilities). An obvious example would be the translation of the examination into Braille or provision of portions of the test in American Sign Language.

An examination also can take a different form when the presentation of the assessment is literally the same but there is divergence in how constructed-response questions are scored. A good example is found in the Smarter Balanced Assessment Consortium, in which member states choose a scoring vendor and whether that vendor uses human grading, machine grading, or both methods. Those choices may not necessarily produce the same results across states even within the same method, depending on the resolution procedures used when raters disagree or upon the particular machine-scoring algorithm that is employed (Bennett & Zhang, 2016).

The third way an instrument can take a different form is less obvious. This type of divergence occurs when the instrument's presentation, response mode, and scoring are, from an objective perspective, the same for all examinees. However, for any pair of examinees that have contrasting characteristics, that assessment may appear to be as different as night and day. Consider the following pairs: (1) an examinee with sight and one with visual impairment, each presented with an unaccommodated test; (2) an examinee who routinely composes essays on a computer and one who typically writes on paper, each given an online writing examination; (3) a native English speaker and an English learner, both taking the test in English; (4) two individuals, one from the mainstream culture and the second from an environment having very different practices, both presented with a reading comprehension test presuming significant background knowledge of U.S. cultural norms; (5) two examinees who are otherwise the same, but one has seen and practiced the test items in advance; (6) two comparable examinees with the exception that only one perceives the test's consequences to be personally significant; and (7) one student having received instruction and the other having not had sufficient opportunity to learn. In all these cases, how the examinees perform and the scores they receive are facts. However, the interpretations we give could be very different and, thereby, the comparisons between those examinees (and the groups to which they belong) are weakened.

Different Instruments

In addition to divergence related to instruments that are nominally the same, comparisons can become weaker when performance on two different instruments is involved. The source of the weakened comparison is that different instruments will typically diverge in terms of the content and processes they measure, as well as the reference frames used to characterize performance.

This type of divergence occurs with some frequency. It can occur for assessment systems being used for the same purpose, such as when we try to compare the percentages

of students achieving proficiency on Smarter Balanced with those on the Partnership for Assessment of Readiness for College and Careers (PARCC) assessments. Such divergence also occurs between two different stand-alone tests used for the same purpose. One example would be use of the SAT or the ACT, and the TOEFL iBT® or the International English Language Testing System™, in making postsecondary admissions decisions; another example encompasses the many assessments used by states for classifying students as English learners (see Chapter 6, Comparability When Assessing English Learner Students). Finally, divergence can occur when measures built for one purpose are also used for and compared to assessments built for another purpose. Comparing the percentage of high school students who achieve proficiency when taking the ACT or the SAT as a state accountability measure to the analogous percentage taking an assessment built to measure a state's content standards directly might be an example (NCME, 2019).

Different Populations

Finally, comparisons become weaker when the same instrument is administered to two student samples that diverge enough from one another that they can be considered as coming from different populations (where the intent is not to compare those different populations). An infamous example is the U.S. Department of Education's attempt to evaluate school achievement across states by using ACT and SAT performance (Wainer, Holland, Swinton, & Wang, 1985). That comparison was undermined by the fact that considerably different proportions of high school students took those tests in each state. A second example is when student performance is compared across states that have different accommodation policies for students with disabilities or English learners. A last example is when the same test is administered at two points in time and the population's composition has materially changed over that period (see Chapter 3, Comparability of Aggregated Group Scores on the "Same Test").

SUGGESTIONS FOR PRACTICE

In this section, we offer suggestions for interpreting score comparisons, discussing each one in turn.

A first step is to determine why a comparison might need to be made. The wisdom of making a comparison may vary with decision-making purpose so it is important to be clear about that purpose. Comparisons can be purely descriptive, made simply for reporting what occurred. An example is in detailing how various states are ranked in terms of their students' performance on the National Assessment of Education Progress (NAEP)—a matter of fact. In practice, this is the view held by some test sponsors, who choose to report results without interpretation. For example, NAEP reports typically stay quite close to the observed results.

Descriptive purposes can, however, quickly turn (or be turned) into inferential ones because we naturally want, and often automatically do, imbue facts with interpretation. Those interpretations, by definition, entail inferences, which together provide the basis for using results in decision making.

Interpretation is, in fact, what state policy makers, the press, and the public do with the descriptive results that come from NAEP. One or more of those groups could, for

example, infer that the observed differences among states were due to differences in teacher competency, the rigor of state education standards, the policies and practices for teacher evaluation, the population demographics, or some combination of factors. Each of these inferences, of course, has particular action implications. However, a more reasoned approach is to regard descriptive results as an opportunity for posing questions that, in turn, motivate the generation of additional evidence to help distinguish among competing interpretations.

A second basic step for interpreting test-score comparisons is to ascertain what methods might have been used to make the desired comparisons tenable. Comparisons can often be made more defensible by using statistical techniques as part of generating assessment results (e.g., making adjustments to allow scores from one test form to be compared with those from another test form). Methods for facilitating comparability vary in their requirements and the degree to which they produce exchangeable scores. As a consequence, some methods may be more suitable for particular decision-making purposes than others. Equating, concordance, and prediction are examples that range from stronger to weaker in their requirements and in the results that they produce. Other chapters in this volume describe these methods (see Chapter 2, Comparability of Individual Students' Scores on the "Same Test," and Chapter 5, Comparability Across Different Assessment Systems), as well as related technical concerns (see Chapter 4, Comparability Within a Single Assessment System). For interpreting score comparisons, we suggest identifying whether the method used (if any) supports the desired comparison.

A third step is to consider how and to whom results will be reported and how comparative claims will be made. Comparative claim statements can appear (or be implied in) score reports, press releases, websites, and other communications, all of which afford opportunities to help audiences make sensible comparisons and avoid untenable ones.

In preparing to report comparative results, it is best to determine first whether the same test was used, and whether it was administered under the same conditions to comparable student samples at the same point in time. If these circumstances do not hold, the specific divergence(s) should be identified and the impact of those divergences on the meaning of assessment results evaluated to the extent feasible. Many methods exist for evaluating the invariance of score meaning across different test variations (e.g., languages or delivery media), examinee populations, and administrative conditions (see Chapter 7, Comparability When Assessing Individuals with Disabilities, and Chapter 8, Comparability in Multilingual and Multicultural Assessment Contexts). A justification for making the comparison in the presence of those divergences should be offered, including a logical rationale and a delineation of the empirical evidence supporting or challenging the comparison.

Technical advisory committee (TAC) guidance is essential in considering the comparison, empirically evaluating its tenability, and creating a justification built on logic and evidence. Of central importance is to start from the premise that results have to be reported and that score comparisons will inevitably be made. The task then becomes one of fashioning communications that responsibly describe results, offer defensible comparisons, and warn against unwarranted inferences.

To that end, we suggest working with the TAC to adjust the strength of the comparative claim as a function of (1) the extent to which the instruments, assessment conditions, student samples, and time between administrations diverge, and (2) the extent of the logical and empirical support available to back the claim. Claim statements can

be adjusted in terms of confidence level based on these two factors. A high-confidence claim would be one for which there is no or little divergence, or there is some divergence but good justification for the comparison given that divergence (e.g., scores have been equated). A lower-confidence claim might be very plausible given current education theory but have limited or no empirical backing. Claims of this type should be more tentatively stated. In all cases, the caveats that attend to the comparison should be clearly articulated and unjustified inferences identified as such (Toulmin, 1958).

Table 9-1 gives some examples of possible comparisons along with more and less defensible claims related to them. Note that the more defensible claims stick closely to the measures used and populations assessed; tenability decreases as claims take on greater levels of generality. For example, it would be reasonable to claim that females scored higher than males on the 2011 NAEP eighth grade writing assessment when composing online essays on demand to persuade, explain, or convey experience. It would also be reasonable to suggest that U.S. eighth grade females were better writers than males *in that context*. Less tenable would be the claim that females were better writers than males generally because, among other things, these 2011 NAEP results targeted a single grade, composition in a particular medium (on computer), writing on demand (which may differ from classroom composition), and particular writing purposes. More general still, and quite untenable, would be the claim that females received better writing instruction than males, a causal attribution that NAEP is not designed to support (NCES, n.d.).

TABLE 9-1 Example Comparisons and Claims of Varying Degrees of Defensibility

Comparison	Well-Supported Claim	Claim Requiring Additional Evidence	Claim Not Recommended
Performance of male and female students on eighth grade 2011 NAEP writing assessment	Female students scored higher than male students at the eighth grade level when composing online essays on demand to persuade, explain, or convey experience	Female students write better than male students *Comment*: This comparison requires evidence that the NAEP results extend to other grades, to writing on paper, and to other writing purposes than those assessed	Female students received better writing instruction than male students *Comment*: This comparison presumes a causal connection between the instruction received and the outcome measured, which NAEP was not designed to support

TABLE 9-1 Continued

Comparison	Well-Supported Claim	Claim Requiring Additional Evidence	Claim Not Recommended
Performance of students in the same school taking the fourth grade state reading assessment in 2018 and 2019	The percentage of fourth grade students reaching proficiency increased by 10 points from 2018 to 2019	Fourth grade reading instruction is having a positive effect *Comment*: This claim would be strengthened by evidence that the two assessed fourth grade populations were demographically comparable, similar percentages of eligible students tested, the test did not change in any material way across the 2 years, no pre-knowledge or other forms of cheating were evident, and no errors in scoring or analysis occurred	The reading skills of fourth graders improved *Comment*: We do not know that the fourth graders improved because the same group of students was not compared. This claim might be better stated as, "The reading skills of the 2019 fourth grade students were greater than those of the 2018 fourth graders"
Performance of two third grade students, each taking their home district's interim assessment	Both students received the same percentile score in mathematics and are estimated to be equally competent with respect to other third graders in the respective tests' norming samples	The students have similar levels of mathematics competency *Comment*: This claim would be strengthened by evidence that the two assessments were built to the same content standards, had similar types of items covering those standards to comparable degrees and levels of rigor, used similar student samples and methods in setting scales and norms, and were administered under analogous conditions	Their districts are equally effective in educating them *Comment*: This claim presumes that the districts' efforts are solely responsible for the students' achievement, goes well beyond mathematics, assumes that the districts have offered equal opportunities to learn, and is based on a single achievement indicator
Performance of 10th grade students on a new state achievement test compared to last year's results on the old test	This year's cohort had a lower percentage proficient	This year's test is harder *Comment*: This claim would be strengthened by evidence that the proficiency standards for the two tests were set in ways that allow meaningful comparison and there were no material changes in the 10th grade populations	The state's students are becoming less intellectually capable *Comment*: This claim conflates achievement of content standards with intellectual capability and presumes that differences between the two measurements are rooted in the populations measured rather than changes in the test

continued

TABLE 9-1 Continued

Comparison	Well-Supported Claim	Claim Requiring Additional Evidence	Claim Not Recommended
Performance of a school's fourth grade students on its English language arts (ELA) state test to an estimate of the national average for all fourth grade students taking their respective state ELA tests, when those tests are rescaled through NAEP	The school's fourth graders scored below the national average in ELA	The ELA achievement of the school's fourth graders is below the national average *Comment*: This claim would be strengthened by evidence that ELA content standards were similar enough across states to allow for creating a coherent common scale, the scaling was technically adequate, and assessment participation rates and accommodation policies were not divergent from the national average	Educational opportunity in the school is below the national average *Comment*: This claim presumes that an outcome, test performance, is equivalent to an input, opportunity

CONCLUSION

This chapter focused on interpreting test-score comparisons. The chapter began with the premise that comparisons are inevitable and, in fact, desirable because obtaining meaning from assessment results requires them. We noted that comparisons are strongest when the same measure is given under substantively the same conditions to comparable student samples at the same point in time, with departures serving to weaken comparisons. The more severe and numerous the departures, the less defensible the comparison is likely to be. In interpreting test-score comparisons one should articulate why the comparison is being made, ascertain if the comparison is appropriate given the technical methods used, present a rationale based on logic and evidence to support the comparison, and warn audiences against inappropriate inferences. Finally, tenable comparisons will usually be ones that stay reasonably close to the measures employed and populations tested. As comparative claims become more general, their reasonableness usually declines.

REFERENCES

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (Research Memorandum 03-05). Princeton, NJ: ETS.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B, Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 6*(9). Retrieved from http://files.eric.ed.gov/fulltext/EJ838621.pdf.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York: Routledge.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2). Retrieved from http://files.eric.ed.gov/fulltext/EJ843858.pdf.

NCES (National Center for Education Statistics). (n.d.). *Interpreting NAEP technology and engineering literacy results*. Retrieved from https://nces.ed.gov/nationsreportcard/tel/interpret_results.aspx#cautions.

NCME (National Council on Measurement in Education). (2019). *National Council on Measurement in Education position statement on the use of college admissions test scores as academic indicators in state accountability systems*. Retrieved from https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission_Statement_06-16-19.pdf.

OECD (Organisation for Economic Co-operation and Development). (2018). *PISA 2015 science test*. Retrieved from http://www.oecd.org/pisa/test/other-languages.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Wainer, H., Holland, P. W., Swinton, S., & Wang, M. H. (1985). On "State education statistics." *Journal of Educational Statistics, 10*, 293–325.