Report to Ofqual

On the Reporting of Measurement Uncertainty and Reliability for U.S. Educational and Licensure Tests

Richard P. Phelps, April Zenisky, Ronald K. Hambleton, & Stephen G. Sireci

Ofqual/10/4759 **March 18, 2010**

Ofqual This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Table of Contents

Executive Summary

Ofqual seeks to determine the prevalence and character of measurement uncertainty reporting for high-stakes tests in the United States. The research questions might be phrased as: Is the reporting of measurement error (i.e., score imprecision) common or typical, or is it uncommon or atypical? And, if it is common or typical, how is it commonly or typically done?

We conducted Web searches (and followed up where needed with telephone calls) and contacted key researchers at relevant entities involved in reporting test results in the United States. We sought to learn:

> The prevalence among our sample respondents of the reporting of measurement uncertainty in high-stakes tests.
>
> The degree of ease or difficulty with which ordinary citizens may access such information.

The degree of transparency with measurement uncertainty issues varies. Transparency seems to be greater for education than for licensure tests, for mostly objective than for mostly essay tests, for larger programs than for smaller programs, and, perhaps ironically, the greater the role of test contractors and the smaller the role of state government.

With educational tests, many of the states highlight imprecision along with the student scores on the parent/student reports. (More states now are reporting score bands.) But all states prepare technical manuals, and just about all technical manuals are readily available to those who want them.

With licensure exams, the situation is mixed. Some provide information about uncertainly on the candidate report itself, and more reliability information in a yearly technical document. Others make available various technical reports and papers summarizing reliability information. Still others produce reports with substantial detail that are not released to the public.

Is the totality of uncertainty reported to all stakeholders in U.S. educational and licensure testing programs? No. It would be difficult for the average parent to find a full range of measurement uncertainty statistics for their children's tests, for example. But, then, the average parent would not be looking. And, that is why technical manuals are not found front and center on the home page of testing program Web sites. Documents that better respond to the typical consumer's needs are placed front and center, and the technical manuals are placed a few to several clicks behind. But, they are not hidden. There seems not to be any effort to hide information; the level of dissemination appears to respond well to the demand for it.

Background and Introduction

The U.S. "rule book" of test development and validation is the *Standards for Educational and Psychological Testing* (1999), published and periodically updated by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (A new edition is being written and is expected to be ready for publication in about two years.)

The *Standards* clearly specify that the degree of measurement uncertainty must be reported. Take, for example, Standards 2.1 and 2.2 (p.31):

> *Standard 2.1*. For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.
>
> *Standard 2.2*. The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

Granted, the *Standards* are rules made by professional associations and not public laws enacted by elected legislatures. In practice, however, U.S. courts consistently have deferred to them in their rulings, giving them *de facto* legal status (Buckendahl & Hunt, 2005). For this reason and because they themselves aspire to high technical standards, U.S. test publishers are careful to insure that their methods and practices are consistent with the Standards.

Moreover, under the federal No Child Left Behind (NCLB) Act of 2001, some state testing programs are mandated by law to "produce individual student interpretive, descriptive, and diagnostic reports consistent with clause (iii)[1] that allow parents, teachers, and principals to understand and address the specific academic needs of students …and that are provided to parents, teachers, and principals …in an understandable and uniform format, and to the extent practicable, in a language that parents can understand;" [section 1111(b)(3)(C)(xii)].

In addition, the *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001* (Peer Review Guidelines), developed by the U.S. Department of Education (2009) to help states meet the quality assessment guidelines mandated by NCLB, require states to provide information regarding measurement precision for any scores reported.

A high-stakes assessment in the United States not reporting measurement uncertainty to the public would be in violation of the *Standards*, and in the case of NCLB assessments, in violation of the *Peer Review Guidelines*. But, *how* it is reported can vary and how

[1] Clause iii reads: "[assessments shall] be used for purposes for which such assessments are valid and reliable, and be consistent with relevant, nationally recognized professional and technical standards;…" [section 1111(b)(3)(C)(iii)]

widely the reporting is disseminated is not specified by the *Standards* or the *Guidelines*.[2] Indeed, there seems to be a range of understanding as to what "reporting to the public" means exactly. The NCLB Act insists that parents—a public group—be provided information about their children's achievement consistent with technical standards of validity and reliability but (perhaps contradictorily) in a format and language "that parents can understand."[3]

In reality, there are multiple definitions of "public," not all of which portend public participation. For example, the *American Heritage Dictionary of the English Language* (1971) lists five for the adjectival version of "public." Two of them imply public participation, provision, or dissemination, and two of them do not. The latter two read as follows:

> "Of, concerning, or affecting the community or the people."
>
> "Connected with or acting on behalf of the people, community, or government, rather than private matters or interests."

One can think of many examples of information that may "concern or affect" or be "connected with" the population as a whole but not widely available to them: e.g., classified intelligence documents; personal medical data in government data bases; and the text of proposals submitted for government grant funding prior to the award decision.

Some public information is legally required to be disseminated widely or made always easily available to the public. Some public information is not widely available, but is obtainable upon request. Some public information is available only with a Freedom of Information Act request or a court order. Still, there remains some public information that most members of the public cannot obtain. And, beyond the legal requirements exist typical patterns of practice.[4]

---

[2] For example, the "how" might be a descriptive statement in a footnote of a score report indicating reasons for score imprecision (e.g., the choice of items, or the role of guessing). Other possibilities include providing scores with a numerical band, or score bands shown graphically. Sometimes, however, the "how" might be addressed in an interpretive manual that accompanies score reports, or the "how" might be addressed in a technical manual.

[3] The following clause in the NCLB Act (clause iv) adds that states provide the U.S. Secretary of Education "evidence from the test publisher or other relevant sources that the assessments used are of adequate technical quality for each purpose required under this Act …and such evidence is made public by the Secretary upon request;" There is no elaboration on the meaning of "made public" in the Act.

[4] In addition, the American Psychological Association promotes the *Rights and Responsibilities of Test Takers: Guidelines and Expectations* (2010) that, like the *Standards*, are not law but, nonetheless, effectively have some of the force of law and influence Association members. The *Rights and responsibilities* state:

"Because test takers have the right to be tested with measures that meet professional standards that are appropriate for the test use and the test taker, given the manner in which the results will be used, testing professionals should:

"…Take steps to utilize measures that meet professional standards and are reliable, relevant, useful given the intended purpose and are fair for test takers from varying societal group.

"…Inform test takers, upon request, how much their scores might change, should they elect to take the test again. Such information would include variation in test performance due to measurement error (e.g., the appropriate standard errors of measurement) and changes in performance over time with or without intervention (e.g., additional training or treatment)."

Relevant entities

The public provision of education and the licensing of professionals both fall under the "reserved powers" clause in the Tenth Amendment to the U.S. Constitution—the last of the original group of amendments commonly known as the "Bill of Rights" (U.S. GPO, 1996). Neither is enumerated among the duties of the federal government, therefore they are primarily state responsibilities except in cases where another federally-guaranteed civil right may be affected.

U.S. states specify their responsibilities for education and licensure in their own constitutions and regulations. While states often copy constitutional and regulatory language from other states when they like what they read, they do not always and, even when they do, they rarely copy text word for word.

Thus, to know specific reporting practices in each state, one must consult each state. Not surprisingly, law and practice in test score reporting varies across the 50 U.S. states. A common practice is to report only the most easily digestible test result statistics in publications destined for wide dissemination, and more technical statistics—such as those specifying measurement imprecision—in "technical" and/or "test development" reports which are not widely disseminated. We can say with 100% accuracy, that every one of the states in the U.S. reports test score reliability for educational tests in technical manuals. It is also common (though not by all states) to communicate the concept of score imprecision to test takers with graphical score bands, with ranges (within which scores would fall if the testing were replicated), and sometimes statements appear at the bottom of score reports explaining the concept of measurement imprecision.[5] Whether in numbers, graphics, or words, and whether on score reports, in interpretive guidelines (sometimes, the concept is explained in an "interpretive guide for parents"), or in technical manuals, the concept of score imprecision is communicated. For tests with items scored subjectively, such as written answers, it is common, too, to report some measure of inter-rater reliability in a technical manual.

Practice also varies in the *extent* to which the source of measurement imprecision is reported. Often it is reported to students that score imprecision results from the sampling of test items used on a given day, the impact of guessing, how they might be feeling on a particular day, etc. Imprecision due to the subjectivity of scoring constructed-response items is almost always reported in technical manuals. Imprecision associated with the setting of performance standards (i.e., cut scores), too, is usually reported in technical manuals.

Reporting score imprecision is not typically as thorough among professional licensing organizations. Indeed, it is uncommon among the smaller professions to make technical manuals. The larger professions, however, typically comply with the *Standards* and provide reliability information in their technical manuals.

---

[5] See, for example, pp.13–15 in California's *Explaining 2009 STAR Program Summary Results to the Public: Assistance for School District and School Staff*
<http://www.cde.ca.gov/ta/tg/sr/documents/star09explpts.pdf>

Research questions

Ofqual seeks to determine the prevalence and character of measurement uncertainty reporting for high-stakes tests in the United States in qualitative rather than quantitative terms. The research questions might be phrased as:

> Is the reporting of measurement error (i.e., score imprecision) common or typical, or is it uncommon or atypical? And, if it is common or typical, how is it commonly or typically done?

Method

We conducted Web searches (and followed up where needed with telephone calls) and contacted key researchers at relevant entities involved in reporting test results in the United States. Because we are interested in prevalence, we sampled the largest among them, such as:

> Five of the largest commercial developers of high-stakes educational tests
> Five of the largest commercial developers of high-stakes licensure tests
> Ten of the largest professional associations that administer licensure tests
> The ten largest state education agencies
> Ten of the largest state professional licensing boards

We sought to learn:

> The prevalence among our sample respondents of the reporting of measurement uncertainty in high-stakes tests.
>
> The degree of ease or difficulty with which ordinary citizens may access such information.

As we searched and surveyed, we retrieved links to or copies of documents reporting measurement uncertainty. Our work was conducted in February and March of 2010.

U.S. Education Examinations

In this section we discuss the reporting of measurement uncertainty for education examinations and, in particular, four types: state tests required by the NCLB Act; high school (i.e., upper secondary) exit examinations; college (i.e., university) entrance examinations; and the federally-operated National Assessment of Educational Progress (NAEP). All but one of the ten largest U.S. states administer all four types, and Pennsylvania is considering developing the one type it currently does not (a high school exit exam). Note that for the NAEP assessments, there is no reporting of scores to or of

individuals, schools, or districts. The focus is on ethnic and gender groups, the states and the nation.

See Appendix A for selected details on the reporting of reliability issues, by type of test and state.

Tests used to meet federal NCLB Act requirements

The NCLB Act requires annual administrations of reading and mathematics tests across seven grades and of science across three grades. All but one of the ten largest U.S. states have contracted with private test developers for the more technical psychometric work. In North Carolina, psychometric experts at two state universities develop the tests.

In some states, it can be difficult to learn the identity of the contractual test development firm (but this is rare), whereas in other states their identity may be displayed quite prominently. The more transparent the identity and role of the test contractor, generally, the easier it is to find details on measurement uncertainty. Many of the states today are redesigning their score reports, adding score bands to student scores or some other indication of the measurement precision of scores and performance classifications. Interpretive guides caution against over-interpreting test scores because of measurement error. Technical manuals, too, are full of information related to the reliability of scores and student classifications (e.g., KR-20, coefficient alpha, stratified alpha, decision consistency, decision accuracy, kappa).

The attached spreadsheet compares reliability information provided by Illinois, California, Pennsylvania, and North Carolina. These states are arranged roughly in order of the complexity of that information: the technical reports of Illinois and California are very long, extremely thorough, and quite complex. Few without statistical training would understand them. The technical reports of Pennsylvania and North Carolina are thorough, but not exhaustively so like the other two, and written more accessibly.

Many testing firms and states also make a substantial effort to describe, or at least illustrate the size, of measurement uncertainty in their group and individual score reports. Two decades ago, most education test score reports provided point estimates only. Now, most also report scores in intervals, bands, or ranges.[6] The impetus for this trend can be attributed both to prodding from governments and from within the measurement profession itself.

North Carolina offers examples in the interpretive brochures that accompany their score reports. Typical language reads like this (NCDPI, 2):

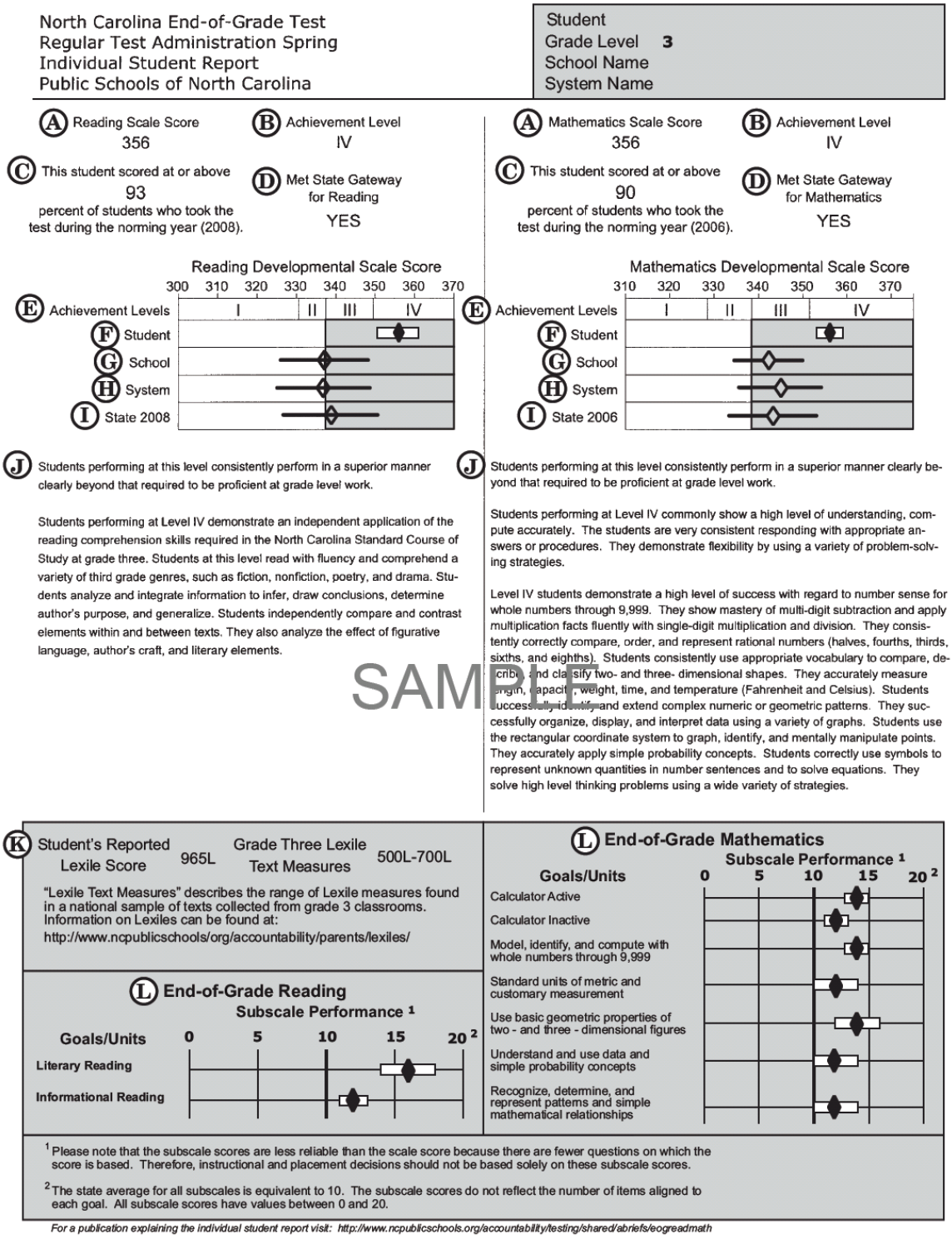
> "The closed diamond (♦) represents your child's performance at the individual goal level. The bar (▬) represents the standard error of measurement (SEM). The

[6] In their report to Ofqual, Bradshaw and Wheater (2009) describe U.S. score reporting with confidence bands in Washington State (p. 14) and by ACT and the College Board (p. 23).

SEM indicates how much your child's score is expected to vary if tested repeatedly with the same test, assuming that no additional instruction is given."

The graphics that accompany this text are shown in Figure 1.

Figure 1. North Carolina End-of-Grade Test student score report



Exit examinations

The attached spreadsheet shows a similar comparison for high school (i.e., upper secondary) exit examinations among California, Illinois, and North Carolina.

National Assessment of Educational Progress (NAEP)

The federally-mandated NAEP tests samples of students in grades 4, 8, and 11/12 in a matrix-sampling scheme. Extensive efforts are made to make NAEP data accessible to researchers and to policymakers. With the passage of the NCLB Act, the NAEP was bestowed an additional role as benchmark for the state tests used to meet the NCLB requirements. All states are now required to administer the NAEP to representative statewide samples.

Perhaps more than any other testing program in the United States, all aspects of the NAEP are transparent, its data are readily provided to researchers, and its results are made easier to interpret with numerous interpretive reports and on-line tools.

University entrance examinations

Most U.S. higher education institutions accept applicant test scores from both of the two entrance examination organizations, the College Board or ACT, developers of the SAT and ACT assessments, respectively.

Both organizations provide thorough technical reports and go to some lengths to explain scores and scoring to their student test-takers. Students may not be provided Kuder-Richardson 20 and coefficient alpha numbers, but they are informed of the uncertainties due to sampling and measurement and provided ranges for their score-point estimates in their score reports.

U.S. Professional Licensure Examinations

In this section we discuss the reporting of measurement uncertainty in the context of certification and licensure assessment. The focus here is on five fields: medicine, accounting, law, nursing, and teaching. These are among the biggest and most important certification and licensure assessments in the country. As with the K–12 assessment results, this analysis was conducted by reviewing materials from ten states: California, Texas, New York, Florida, Illinois, Ohio, Michigan, Pennsylvania, Georgia, and North Carolina.

See Appendix B for selected details on the reporting of reliability issues, organized by licensure test.

Nursing: NCLEX-RN and NCLEX-PN Examinations

The NCLEX-RN and NCLEX-PN exams are directed by the National Council of State Boards of Nursing (NCSBN), and are uniformly used in the ten states of interest as the exams used to license registered nurses (RNs) and practical nurses (PNs), respectively. Both the NCLEX-RN and NCLEX-PN are multiple-choice tests administered by computer, and are variable-length adaptive tests. In this case, the actual number of multiple-choice items varies depending on individual response patterns, as the test will end when a 95% confidence interval for estimation of ability is attained for either a pass or a fail decision *and* a minimum number of items are administered. Examinees will be presented with a minimum of 75 items on the NCLEX-RN and 85 items on the NCLEX-PN regardless of the location of the confidence bands to insure that candidates all receive a content valid examination.

By and large, the NCSBN makes a reasonable amount of information about the processes of test development and psychometrics available, and some technical details about measurement error can be found. At the same time, it must be recognized that the focus on testing until a high level of confidence is reached about the pass-fail status of candidates is a very direct way to handle the concern about measurement error. Thus, the computerized-adaptive testing algorithm ensures a high degree of measurement precision in the pass/fail decision. Reporting more information on measurement precision in this context is difficult, since the CAT algorithm results in candidates take different sets of test items, and so the reliability of scores from different test "forms" does not apply.

Accountancy: Uniform CPA Examination

The Uniform CPA Examination is the test used by all 50 states and other US jurisdictions to license individuals as Certified Public Accountants. The test is directed and developed by the American Institute of Certified Public Accountants (AICPA). The CPA Exam is a computer-delivered test that has both multiple-choice and constructed-response (simulations) components, and is administered using a multi-stage adaptive algorithm.

The CPA Exam provides a wide range of psychometric resources on its Web site, and specific details on many aspects of measurement uncertainty are diffused among various sources. In many cases, processes for detecting and minimizing measurement uncertainty are referenced, and quantitative results are typically found in the form of research studies that have been completed by AICPA staff and consultant psychometricians. As with the NCSBN exams, reporting measurement precision in a computerized-adaptive context is not straightforward, but concerns over measurement precision are built into the test development, administration, and scoring processes.[7] The candidate report does contain a statement about possible sources of measurement error, and in providing candidate diagnostic information, confidence bands are used to emphasize the role of measurement imprecision.

Medicine: United States Medical Licensing Examination

The National Board of Medical Examiners (NBME) licenses medical doctors in all U.S. jurisdictions. The USMLE consists of three different tests, termed Step 1, Step 2, and Step 3, which are administered at different points in an individual's medical education. By way of overview, Step 1 assesses whether medical school students or graduates understand and can apply important concepts of the sciences basic to the practice of medicine. Step 2 has two components (Clinical Knowledge and Clinical Skills) and assesses whether medical school students or graduates can apply medical knowledge, skills and understanding of clinical science essential for provision of patient care under supervision. Lastly, Step 3 assesses whether medical school graduates can apply medical knowledge and understanding of biomedical and clinical science essential for the unsupervised practice of medicine. It is required that a candidate pass all of these tests in order to be licensed by the state in which that person would like to practice medicine.

By and large, a great deal of psychometric activity is known to take place relative to the USMLE, and quantitative information about the Step 1, 2, and 3 exams is available in the form of numerous published papers and invited presentations completed by NBME staff researchers and psychometricians, rather than in a single technical manual. Thus, locating some of the relevant measurement precision information for the USMLE involves some searching. Also, as with other licensure examinations such as the Uniform CPA Exam, the score reports provided to USMLE examinees provide diagnostic information with confidence bands to emphasize the role of measurement imprecision.

Teaching

Praxis. Developed and maintained by Educational Testing Service (ETS), the Praxis suite of tests is used for teacher certification in a range of jurisdictions. Of the ten states of interest, Praxis is used in four (Texas, Ohio, Pennsylvania, and North Carolina). The Praxis exams comprise Praxis I Pre-Professional Skills Tests, which are intended to measure basic skills in reading, writing and mathematics, and the Praxis II Subject

[7] Moreover, candidates who score near the passing threshold have their constructed-response items rescored, in recognition of the uncertainty of the original estimate.

Assessment tests, assess subject-specific content knowledge as well as general and subject-specific teaching skills. Praxis is administered across the country to about 50% of candidates applying for teaching licenses or certificates.

Teacher tests by NES/Pearson. The Evaluation Systems group of Pearson Educational Measurement is responsible for the remaining six states' teacher licensure examinations (California, New York, Florida, Illinois, Michigan, and Georgia). Typically, these states require prospective educators to take both a basic skills-type assessment and a subject- and level-specific test. Names of the testing programs in the six states follow.

California: California Basic Educational Skills Test (reading, math, and writing); California Subject Examinations for Teachers

New York: New York State Teacher Certification Examinations [Liberal Arts and Sciences Test (LAST) and Assessments of Teaching Skills-Written (ATS-W), Content Specialty Tests (CSTs)]

Florida: Florida Teacher Certification Examinations [FTCE General Knowledge Test (GK), FTCE Professional Education Test (PEd); FTCE Subject Area Examinations (SAE)]

Illinois: Illinois Certification Testing System (ICTS) [Basic Skills test, Assessment of Professional Teaching (APT) tests, Content-area tests]

Michigan: Michigan Test for Teacher Certification (MTTC) [subject-area tests, including the world language tests, and a Basic Skills test]

Georgia: Georgia Assessments for the Certification of Educators [Basic Skills Assessment, Content Assessments]

No technical documentation in the form of technical manuals or other reports appears to be easily accessible for any of the NES/Pearson teacher testing programs. Nonetheless, test development and technical reports are written for each state program, even though they are typically not disseminated. For the general public, various activities in the test development/quality process are referenced at the most general level, with the bulk of resources concerned with informing candidates of test content and job analyses.

Law

Multistate Bar Examination (MBE). Developed by the National Conference of Bar Examiners (NCBE), the MBE is a six-hour, 200-question multiple-choice examination covering contracts, torts, constitutional law, criminal law and procedure, evidence, and real property. The MBE is accepted by many jurisdictions in the United States as a general bar admissions test, though many (including all ten of interest here) require their own state tests in addition.

The MBE has a technical manual that is distributed to state boards of bar examiners and committees, but little psychometric information about the MBE is readily available to the public except as reports and presentations by NCBE staff. This documentation focuses on informing candidates about exam content and the job analyses used to develop it.

Multistate Professional Responsibility Examination (MPRE). Also produced by the NCBE, the MPRE is a 60-question, two-hour-and-five-minute, multiple-choice examination required for admission to the bars of all but four U.S. jurisdictions. Passing scores are established by each jurisdiction and currently vary between 75 and 86.

As with the MBE, technical documentation for the MPRE is disseminated to the state boards, with little psychometric information about the MPRE readily available to the public. Public documentation on the MPRE focuses on informing candidates of exam content.

Multistate Essay Examination (MEE). The MEE is a collection of 30-minute essay questions, also from the NCBE. Any jurisdiction may use it, but at present only Illinois among the ten largest states has adopted it.

As with the MBE and MPRE, the focus of the readily available documentation for the MEE is on exam content; technical documentation is generally only circulated to state boards of bar examiners.

Multistate Performance Test (MPT). The Multistate Performance Test (MPT) is also developed by the NCBE and comprises two 90-minute skills questions covering legal analysis, fact analysis, problem solving, resolution of ethical dilemmas, organization and management of a lawyering task, and communication. It is used in Texas, New York, Illinois, Ohio, and Georgia. As with the other NCBE assessments, while technical documentation exists, most of it centers on content, while technical details are reserved for the state boards.

Discussion and Conclusion

The degree of transparency with measurement uncertainty issues varies somewhat in the United States. Transparency seems to be greater for education than for licensure tests, for mostly objective than for mostly essay tests, for larger programs than for smaller programs, and, perhaps ironically, the greater the role of test contractors and the smaller the role of state government. Nonetheless, all the testing programs we examined adhere to the relevant provisions of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME).

With respect to the concept of score uncertainly or imprecision, it is dealt with differently by the agencies producing educational tests and licensure exams. With educational tests, many of the states highlight imprecision along with the student scores on the parent/student reports. It is our impression too that more of the states now are reporting score bands. But all states prepare technical manuals, and these manuals typically report KR-20s, coefficient alphas or stratified alphas (with tests of mixed item formats), standard errors of measurement of total test scores or conditional estimates of measurement error in the scores (for states using IRT in score reporting), and consistency and accuracy of performance classifications. When constructed response items are used, inter-rater reliability of scoring is nearly always reported. With licensure exams, the situation is more varied. All of the ten largest states report score reliabilities in their manuals, and just about all technical manuals are readily available to those who want the information.

The bits of measurement uncertainty information most often left unreported seem to be those related to item drift due to the inconsistent scoring of constructed responses on two or more occasions. This problem has more to do with validity than reliability, but it can affect decision consistency and accuracy.

With licensure exams, the situation regarding score uncertainty or imprecision is mixed. Some exams, such as the Uniform CPA Exam prepared by the AICPA, provide information about this uncertainly on the candidate report itself, and more reliability information in a yearly technical document. Other agencies, such as the NBME, provide various technical reports and papers summarizing reliability information. Finally, the National Bar produces reports on each test administration and reports score reliability and the standard error of measurement, but these reports are not released to the public.

Is the totality of uncertainty reported to all stakeholders in U.S. educational and licensure testing programs? No. But, neither does there seem to be any effort to hide information; the level of dissemination appears to respond well to the demand for it. It could fairly be argued that even few psychometricians would be interested to know the results of some of the most arcane statistical tests for measurement uncertainty short of a contractual obligation to know them. The typical parent is unlikely to request even a brief summary of measurement uncertainty findings.

Ironically, a sincere and successful effort to be transparent can sometimes make at least some of the information more opaque. The technical report for the California High School Exit Examination, for example, is over 800 pages long. It could hardly be more

thorough in its discussion and analysis of measurement uncertainty. But, how many California taxpayers will understand the difference between a coefficient alpha and a stratified alpha and why and where one would use one or the other? More to the point, how many California taxpayers would care to know the difference, or to read 800 pages of statistical jargon and data tables?

For the most part, the technical reports are written by psychometricians to be read by other psychometricians who may serve on technical advisory committees or as expert witnesses in a court trial. But that should in no way disparage them. The fact is: in most cases, the technical manuals are available, clever statisticians took the time to painstakingly assemble then, they can be incredibly thorough, and they are fairly easy to find for someone interested in finding them. They are not meant to explain erudite statistics to the masses, nor would the masses be much interested in the explanation.

It would, indeed, be difficult for the average parent to find a full range of measurement uncertainty statistics for their children's tests. But, then, the average parent would not be looking.[8] And, that is why technical manuals are not found front and center on the home page of testing program Web sites. Were they placed there, many would look at them, not understand them, and either feel perplexed or irritated.[9] So, documents that better respond to the typical consumer's needs are placed front and center, and the technical manuals are placed a few to several clicks behind. But, they are not hidden; in most cases, we had little trouble finding them.

In their report to Ofqual, Bradshaw and Wheater (2009) wrote "…the majority of …assessments do not report error or uncertainty" (p. 22) and "this report has found few examples of reporting of error or uncertainty…" (p. 25). They focused their efforts on score reports rather than on the full panoply of test report documents that would include technical and research reports. Indeed, we also have found score reports in current use that pay little to no attention to measurement uncertainty. Moreover, measurement imprecision statistics tend not to be found among the most accessible or prominent test results simply because there exists so much other important test information that is more popular.

[8] As one psychometrician familiar with the legal profession's testing programs explained to us: "They have technical manuals for the tests, which get distributed to state boards of bar examiners and committees, but they have not been distributed publicly. They are, like most such documents, fairly boring."

[9] We note in passing that in Massachusetts, the state commissioned the development of a "technical manual light" that was prepared for users of the scores to explain technical concepts in simple language. Such efforts, however, are not common.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.

American Heritage. (1971). *The American Heritage dictionary of the English language*. New York, NY: Author.

American Psychological Association. (2010). *Rights and responsibilities of test takers: Guidelines and expectations*. Washington, DC: Author. Available at: http://www.apa.org/science/programs/testing/rights.aspx

Bradshaw, J., & Wheater, R. (2009, December). *International survey of results reporting*. National Foundation for Educational Research & Ofqual.

Buckendahl, C.W., & Hunt, R. (2005). Whose rules? The relation between the 'rules' and 'law' of testing. In R.P. Phelps, (Ed.), *Defending standardized testing* (pp.147–158). Mahwah, NJ: Psychology Press.

Goodman, D.P., & Hambleton, R.K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220.

Hambleton, R.K. (2003). How can researchers and the news media work together to improve public understanding of educational assessment? Invited presentation at the annual meeting of the American Educational Research Association, Chicago, IL.

Hambleton, R.K., Sireci, S.G., & Zenisky, A.L. (2009). Making test score scales and reports more understandable and useful. Paper presentation at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hambleton, R.K., Sireci, S.G., & Smith, Z. (2009). Evaluating NAEP achievement levels in the context of international assessments. *Applied Measurement in Education, 22*, 376–393.

North Carolina Department of Public Instruction, Division of Accountability Services/ North Carolina Testing Program. (2009, March). *Understanding the individual student report for the North Carolina End-of-Grade Tests Grades 3, 4, and 5*. Raleigh, NC: Author.

U.S. Department of Education. (2009). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Retrieved March 17, 2010 from http://www2.ed.gov/policy/elsec/guid/saaprguidance.pdf.

U.S. Government Printing Office. (1996). *The Constitution of the United States: The Tenth Amendment—Reserved Powers*. Washington, DC: Author. Retrieved January 18, 2010 from http://www.gpoaccess.gov/constitution/html/amdt10.html

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2003). Performing at or above proficient: The reporting of NAEP results in the Internet age. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2008). Customizing the view: Evaluating the communication of national assessment results. Paper presented at the Sixth International Test Commission Conference, Liverpool, England.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2009). Evaluating the utility of NAEP reporting practices. *Applied Measurement in Education, 22*, 359–375.

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by The Office of Qualifications and Examinations Regulation in 2010.

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk