What Might Changes in Psychometric Approaches to Statewide Testing Mean for NAEP?

David Thissen
University of North Carolina, Chapel Hill

Scott Norton
Council of Chief State School Officers

August 2013
Commissioned by the NAEP Validity Studies (NVS) Panel

This document is one of four reports by the NAEP Validity Studies Panel that explore the relationship between NAEP and the Common Core State Standards (CCSS) and consider how NAEP can work synergistically with the CCSS assessments to provide the nation with useful information about educational progress. The complete volume with all four reports can be found at www.air.org/common_core_NAEP.

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

This report was prepared for the National Center for Education Statistics under Contract No. ED-04-CO-0025/0012 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Acknowledgments

The authors wish to thank Peter Behuniak, Jim Chromy, Phil Daro, Richard Duran, Gerunda Hughes, Robert Linn, Ina Mullis, Karen Wixson, Frances Stancavage, and George Bohrnstedt for their helpful comments and suggestions. Any misstatements are, of course, our own.

Executive Summary

Development of the Common Core State Standards (CCSS), and the creation of the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC), changes the pattern of accountability testing. These changes raise the question: "How should NAEP's validity and utility be maintained?" The assessments planned by the consortia may be different enough from current state assessments to raise questions as to whether NAEP can continue to play its historic role as an independent monitor or "check" on the validity of state assessments.

It is also clear is that computer-based assessment is coming to K–12 education, and both consortia plan to include more varied item types than have been commonly used in the past.

In considering the future of NAEP and state assessments over the next few years, three scenarios seem possible:

(1) If most states use PARCC or Smarter Balanced assessments, NAEP would continue to have two roles: to monitor claims of improved achievement and to provide the "Rosetta Stone" (common metric) needed to compare performance across the consortia's boundaries.

(2) If the two consortia merge, there would be a nearly national test. In the near term, NAEP would remain useful by serving two of its traditional purposes: to monitor and to provide historical context.

(3) Even if the consortia do not continue indefinitely, their ideas are mostly likely the future of assessment. Questions about the validity of NAEP's results would arise if NAEP remained a paper-and-pencil assessment while statewide assessments were computerized.

NAEP as a Monitor

NAEP is widely regarded as a fair arbiter of results obtained from statewide assessments for the purpose of accountability. When statewide tests show improvement but NAEP results do not, questions are raised about the validity of the statewide test results.

For NAEP to continue to play this role, how similar must NAEP be to the new statewide tests? Statewide tests will soon be computer administered, with technology-enhanced item types. Should NAEP become a computerized test? Does it make any difference if the mode of administration of a test is paper-and-pencil or computerized?

Many studies examining mode effects in educational testing have reported inconsistent or mixed results. Comparability of results *can* often be maintained; however, computerization may have an effect on the results for some subgroups or subject areas.

Notable weaknesses in the literature on mode effects limit the extent to which it can be used to anticipate the effects that might be observed with NAEP. Most studies consider only a single point in time, and the literature is relatively silent on the question of whether gaps in scores among subpopulations may appear different. Examination of the pattern of results over time and among groups should be the foci of research on the effects of the computerization of NAEP.

Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests

Efforts to link the scales of other assessments to the NAEP scale have only been moderately successful, and a large number of cautions have been offered about their usefulness. However, if most states use one of two assessments, the situation changes: More data collection options are practical for linking NAEP to the consortia assessments. The consortia assessments are in their planning stages, so a window of opportunity exists during which they might be designed to incorporate linking data collection.

It is strongly suggested that the scales of the assessments from the two consortia be linked to NAEP. In August 2011, the National Center for Education Statistics (NCES) convened a group of experts on the future of NAEP, followed by a second summit of stakeholders in January 2012. The report from those meetings made the same suggestion.

Conclusion

Computerization of NAEP is inevitable and already planned by the National Assessment Governing Board. Computerized NAEP assessments may appear more similar to future statewide assessments. Comparability of results can usually be maintained as a test makes the transition from paper-and-pencil to computerized administration, but computerization may have an effect on results for some subgroups of the population. Computerization of NAEP is best approached in the same way as other changes to NAEP assessments have been approached: A bridge study should insure the comparability of results across the transition unless an *a priori* decision is made to "break trend" regardless.

Assessments developed by Smarter Balanced and PARCC may reduce the number of statewide tests to the low single digits, thus making linkage feasible. Associations between the results of disparate educational assessments tend to change over time, so any linkage between the NAEP scale and the consortia statewide tests will need to be maintained regularly. A singular opportunity exists in a short window of time—essentially right now—to design the data collection for linkage between the NAEP scale and the consortia assessments while the latter are under development.

NAEP has a long history of implementing gradual change so that results remain comparable from year to year, while, at the same time, the assessments remain relevant in the presence of continuing educational and curricular change. We expect that spirit of gradual incremental change will continue to guide NAEP in its adaptation to the introduction of the Common Core State Standards assessments.

CONTENTS

Introduction

The development of the Common Core State Standards (CCSS), and the creation of two consortia of states—the Smarter Balanced Assessment Consortium (Smarter Balanced)[6] and the Partnership for Assessment of Readiness for College and Careers (PARCC)[7]—to develop assessments based on those standards, promises to change the pattern of K–12 accountability testing in the U.S. These changes raise the question "How should NAEP's validity and utility be maintained in the context of the CCSS?"

Crucial aspects of this question have to do with the relationship between the CCSS and the NAEP content frameworks, which will be examined in other studies. However, it is also possible that changes in the approach to testing planned by the two assessment consortia may induce changes in the ways that existing assessments, such as NAEP, are perceived, or may change how NAEP needs to be scored and maintained to provide an accepted "check" on the validity of the new statewide assessments. Furthermore, a few states have indicated that they will not be joining either of the consortia, further complicating the job of NAEP as a monitor of states' educational achievements.

[6] http://www.k12.wa.us/SMARTER/default.aspx

[7] http://www.parcconline.org/

Background

Common Core State Standards Initiative

At the time of this writing, 45 states and the District of Columbia have officially adopted the CCSS.[8] However, adoption of the standards does not necessarily mean state content standards for K–12 mathematics and English language arts (ELA) will become identical across the states. According to documentation from the Common Core State Standards Initiative, "adoption" means that a "State adopts 100% of the common core K–12 standards in ELA and mathematics (word for word), with option of adding up to an additional 15% of standards on top of the core."[9] Thus, even at the level of standards, there is likely to remain some variation among CCSS states' curricula, and possibly their assessments, while additional between-state variation will arise from the states that have not (yet) adopted the CCSS.

Although both consortia plan assessments that are based on the CCSS, they plan tests that differ in a number of respects. This will split states into three clusters—the Smarter Balanced states, the PARCC states, and the small number of states that are members of neither group and will presumably continue to operate their own assessment programs. Membership of states in the two consortia is listed in Appendix A. All of the states that are members of one or both consortia have adopted the CCSS; none of the five states that are not members of either consortium have done so. Utah adopted the CCSS, but has since withdrawn from Smarter Balanced, while the small number of states that are currently in both consortia will presumably settle on one or the other by the time of operational testing in 2014–2015.[10]

Smarter Balanced Assessment Consortium (Smarter Balanced)[11]

Features of the Smarter Balanced assessments include "Summative Assessments" that are planned to be "Mandatory comprehensive accountability measures that include computer adaptive assessments and performance tasks, administered in the last 12 weeks of the school year in Grades 3–8 and high school for English language arts (ELA) and mathematics." These "capitalize on the strengths of computer adaptive testing, i.e., efficient and precise measurement across the full range of achievement and quick turnaround of results" and "produce composite content area scores, based on the computer-adaptive items and performance tasks." Smarter Balanced also plans "Interim Assessments" that are "Optional comprehensive and content-cluster measures that include computer-adaptive assessments and performance tasks, administered at locally determined intervals." These are to be

[8] However, an article in the May 8, 2012, issue of the *Wall Street Journal* indicates that up to five states are reconsidering their commitment.
[9] Slide presentation "Common Core State Standards Initiative, March 2010," downloaded from http://www.corestandards.org/about-the-standards.
[10] Membership lists for states adopting the Common Core State Standards were obtained from http://www.corestandards.org/in-the-states.
[11] Quoted material in this section is from http://www.k12.wa.us/SMARTER/pubdocs/SBACSummary2010.pdf

"Grounded in cognitive development theory about how learning progresses across grades and how college- and career-readiness emerge over time." System features include "coverage of the full range of ELA and mathematics standards and breadth of achievement levels by combining a variety of item types (i.e., selected-response, constructed response, and technology-enhanced) and performance tasks, which require application of knowledge and skills."[12]

Partnership for Assessment of Readiness for College and Careers (PARCC)[13]

PARCC lists six "priority purposes" for their assessments:

"1. Determine whether students are college- and career-ready or on track
2. Assess the full range of the Common Core State Standards, including standards that are difficult to measure
3. Measure the full range of student performance, including the performance of high- and low-performing students
4. Provide data during the academic year to inform instruction, interventions, and professional development
5. Provide data for accountability, including measures of growth
6. Incorporate innovative approaches throughout the system"

PARCC plans an assessment system with four components. "Each component will be computer-delivered and will leverage technology to incorporate innovations."

Two summative, required assessment components will be designed to:

- "Make 'college- and career-readiness' and 'on-track' determinations,
- Measure the full range of standards and full performance continuum, and
- Provide data for accountability uses, including measures of growth."

Two nonsummative, optional assessment components will be designed to "generate timely information for informing instruction, interventions, and professional development during the school year. An additional third nonsummative component will assess students' speaking and listening skills."

"PARCC will also leverage technology throughout the design and delivery of the assessment system. The overall assessment system design will include a mix of constructed response items, performance-based tasks, and computer-enhanced, computer-scored items. The PARCC assessments will be administered via computer, and a combination of automated scoring and human scoring will be employed."

[12] In late 2012, the Smarter Balanced assessment design was revised to include only one performance task in each subject—mathematics and English/language arts (Gewertz, 2012).
[13] Quoted material in this section is from http://www.parcconline.org/parcc-assessment-design

The PARCC assessments are not, however, currently planned to be computer adaptive, as is the case with Smarter Balanced assessments.

Summary

The extent to which even consortium-member states will have identical assessments is not clear at the time of this writing; it is possible the consortia assessments will be locally augmented, or otherwise modified. In addition, there will probably be some states that use unique assessments. Nevertheless, it is clear that computer-adaptive (Smarter Balanced) or computer-based (PARCC) assessment is coming soon to K–12 testing. In addition, both consortia appear to plan to take advantage of computer administration by including much more varied item types than have been the norm in large-scale assessment.[14] This represents a potentially dramatic shift in assessment; while some states currently administer online tests, they are typically paper-and-pencil tests that have been transferred to the computer. Finally, documentation from Smarter Balanced specifically mentions the idea that some items may reflect learning progressions.

[14] It now appears that both consortia will have to provide paper-and-pencil versions of the test as not all schools will be able to support computer-based assessments. Such paper-and-pencil alternatives will not be the same as the computerized versions with respect to any technology-enhanced item types that the consortia develop or use, so the paper-and-pencil versions would probably be relatively short-term solutions to specific challenges in the initial implementation, rather than continuing alternate forms.

Questions for the Future of NAEP

Correctly anticipating future events is always a challenge. At the September, 2011, meeting of the NAEP Validity Studies Panel (NVS Panel), Peter Behuniak suggested that three scenarios might be considered for the next few years:

1. There might be minimal change from current commitments—i.e., most states become aligned with one of the two consortia, and a few states associate with neither. After the consortia assessments become available in academic year 2014–2015, the majority of the states will use one of those two assessments, with a small number of states using unique, state-specific tests.

2. The two consortia could conceivably merge to become one. There is some basis for such speculation in recent history: As the current consortia were being formed, several smaller exploratory groups merged to become PARCC. If a merger happens, there would be one nearly national assessment, although a few states would likely continue using unique, state-specific tests.

3. The consortia might fragment, become much smaller, or go out of existence entirely after the current Race to the Top federal funding ends. Race to the Top funding is being provided for assessment development only, so new structures will have to be established for Smarter Balanced and PARCC to administer operational assessments. Because it is not clear at the time of this writing what the mechanism might be to provide continued financing for the consortia, prudence demands that this possibility be considered.

Scenario 1: Two Consortia and Nonconsortium States

In a future that has approximately half the states using the PARCC assessments, approximately half the states using the Smarter Balanced assessments, and a few states using unique tests, an appropriately configured NAEP would continue to have two obvious roles. The first role would be to monitor claims of improved achievement. Even when created at the level of consortia (instead of individual states), statewide assessments would be vulnerable to "teaching to the test" and the possible appearance of inflated achievement gains, which would be identified, as they have been in the past, when statewide assessment scores appeared to rise faster than NAEP scores. A second role of NAEP, as the only assessment administered in all states, could be to provide the "Rosetta Stone" needed to compare performance across the consortia boundary (i.e., between the PARCC states and the Smarter Balanced states), possibly including the nonconsortium states. Without some linkage, each year there could be a stack of statewide averages on the PARCC assessment, an unconnected stack of statewide averages on the Smarter Balanced assessment,[15] and results from a few states comparable to neither group. Suitable linking *may* make it possible to compare PARCC and Smarter Balanced results. Of the two possible linking designs, common-population linking appears unlikely, because it seems improbable that any local authority would administer assessments from both

[15] It is conceivable that the consortia could cross-link their assessments without NAEP as an intermediary; however, no plan for this has been announced.

PARCC and Smarter Balanced. A common-item linking design might be feasible, using NAEP to supply the common items; this is discussed in a subsequent section, "NAEP as Lingua Franca: Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests."

Scenario 2: Merged Consortia and Nonconsortium States

If the two consortia merge, the merger would produce a nearly national test. Setting aside for the moment the few nonparticipating states, a single merged consortium would displace NAEP from its unique role as the only national measure of achievement. In the very long term (i.e., decades), this development might render NAEP superfluous. However, in the nearer term, an appropriately configured NAEP would remain useful by serving two of its traditional purposes. NAEP would still perform a "monitor" function because the consortium's one nearly national test would still be vulnerable to "teaching to the test" and the possible appearance of inflated achievement gains. The latter would be identified, as they have been in the past, when statewide assessment scores appeared to rise faster than NAEP scores. In addition, NAEP would continue to provide historical context. It would take decades for a new assessment, even if it was national, to accrue the kind of trend data that NAEP possesses. Trend data have been important for policymakers for some time, and that would be expected to continue.

Scenario 3: No Consortia, but New Ideas Remain

Even if the consortia do not continue indefinitely, the ideas they plan to bring to large-scale assessment are most likely the ideas of the future. Specifically, the fact that both consortia, representing nearly all of the states, emphasize computerized assessment is a clear indicator that many statewide assessments may well use computerized administration within the next few years. In this scenario, NAEP's role as a monitor of fragmented statewide accountability systems could continue, but questions of the validity of NAEP's results would increasingly arise if NAEP remained an "old-fashioned" paper-and-pencil assessment while statewide assessments adopted computer administration and made use of technology-enhanced item types.

NAEP as a Monitor: Paper-and-Pencil NAEP in a World of Computerized Statewide Tests

NAEP is widely regarded as a fair arbiter of results obtained from statewide assessments for the purposes of accountability. When statewide tests show improvement but NAEP results do not, questions are raised about the validity of the statewide test results. Might the state results be the result of "teaching to the test" or "narrowing the curriculum" to obtain high scores?

How Similar Must NAEP Be?

The use of NAEP as a monitor depends on its acceptance as a widely respected measure of student achievement. NAEP's framework- and item-development processes and its data analysis procedures have been universally accepted as state of the art. For the less technically inclined, the paper-and-pencil format of NAEP is very similar to the paper-and-pencil format of most of the statewide assessments for which it serves a monitoring function.

However, this is about to change. Under any of the scenarios described above, within the next five (or very few more) years, statewide assessments will be computer administered, and, in many states, probably computer adaptive, with technology-enhanced item types. If NAEP remains as it has been, it will increasingly "look different."

If paper-and-pencil NAEP "looks different", and its results differ from computerized statewide tests with more varied item types, NAEP may cease to be accepted as the final arbiter, and NAEP results may be dismissed because "students were not as motivated on the old-fashioned paper test as they were on the attractive computerized test," or because "the old-fashioned paper test did not include the instructionally sensitive technology-enhanced item types that are on the computerized test."

Further, if linkages between the NAEP scale and those of the PARCC and Smarter Balanced assessments are proposed (see the next section on "NAEP as Lingua Franca: Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests"), it may, as a practical matter, be necessary for NAEP to become a computer-administered test to perform its part in the linkage.

Should NAEP become a computerized test? There are three classes of considerations involved in answering this question.

- The first class of considerations is practical: Would computer administration make NAEP more or less expensive? If the answer is that it would make NAEP more expensive, is the cost acceptable? Another kind of practical difference between computerized and paper-and-pencil administration involves accommodations: Some accommodations (e.g., large type, audio presentation, some kinds of translation) are easier or less expensive to provide with a computerized test than with paper and pencil. Such practical questions are beyond the scope of this essay (and our expertise).

- The second class of considerations involves the need for NAEP to be computerized in order to administer questions that appropriately measure aspects of the CCSS. If (other) groups examining the CCSS frameworks and consortia assessment plans conclude that there are some objectives that can only be measured with technology-enhanced item types, it may be necessary for NAEP to computerize in order to provide measurement of those aspects of knowledge or skills.

- The third class of considerations can be summed up by the question "Does it make any difference if a test is administered in a paper-and-pencil or computerized format?" There is evidence in the psychometric literature on this question.

What Is Currently Known About Mode of Administration Effects?

Over the past three decades, a number of assessments have been converted from paper-and-pencil administration to computer-based or computer-adaptive administration, beginning with the transition of the Armed Services Vocational Aptitude Battery (ASVAB) in the 1980s–1990s (Sands, McBride, & Waters, 1997). NAEP is among the programs that have computerized some assessments: The 2011 NAEP writing assessment was administered as a computer-based test for Grades 8 and 12, and a pilot study of a Grade 4 computer-based writing assessment was in the field in early 2012.[16] The 2009 NAEP science assessment included interactive computer tasks,[17] and the NAEP Technology and Engineering Literacy (TEL) assessment will be computer-administered when it appears in 2014.[18]

Does computer administration in and of itself affect the results of an assessment?

Many research studies have examined the comparability of results obtained with paper-and-pencil and computerized tests. Appendix B summarizes some of the conclusions that can be drawn from studies over the past 15 years (since 1997); earlier studies were excluded because they would have involved computer administration very different from what would be used now.

The conclusion of Appendix B is that:

> Many studies examining mode effects in educational testing have shown inconsistent or mixed effects. The research is clear in demonstrating that comparability of results *can* often be maintained overall as a test makes the transition from paper-and-pencil to computerized administration. For example, most of the studies suggest that the structure of the test is likely to remain unchanged in moving from paper-and-pencil to computer-based administration. However, the evidence is mixed on the effects of mode on score comparability; computerization may have an effect on the results for some subgroups of the population and these can vary further as a function of

[16] http://nces.ed.gov/nationsreportcard/writing/cba.asp
[17] http://nces.ed.gov/nationsreportcard/science/whatmeasure.asp
[18] http://nces.ed.gov/nationsreportcard/techliteracy/

> the subject area being assessed. Schroeders and Wilhelm (2011) perhaps best summarize what is required when moving to computerized assessment when they write "… equivalence research is required for specific instantiation unless generalizable knowledge about factors affecting equivalence is available" (pg. 1).

Characteristics of assessments that have been shown to raise the possibility of different scores from computerized and paper testing include essay responses, which may be graded more or less stringently depending on mode, and items with graphics or manipulatives, which may be made either easier or more difficult in translation to computerized delivery. Participant characteristics that may interact with the relative difficulty of computerized presentation have included gender (in some studies) and special education status. Probably the most salient (unintended) individual differences variable that may be related to the results obtained with computerized assessments is computer familiarity, which, while not a very well defined term, includes skills with a keyboard and probably some other aspects of the idiom used in the computer interface. However, these effects have been rare historically, and can likely be eliminated with careful assessment design and thoughtful instructions and preparation. Indeed, given the ubiquity of a range of computerized devices in everyday life, from personal computers through tablets and smart phones, it may soon be the case that the question would be whether paper-and-pencil testing accurately or authentically measures what children know and can do.

For NAEP, the difference between computerization alone (making a computer-based test [CBT]) and adaptation (creating a computerized adaptive test [CAT]) should not be significant. NAEP is already "scored" (actually, aggregate summary statistics are computed) using an item response theory (IRT) model in the presence of planned missing data, due to the fact that each examinee responds only to the subset of items. Use of a CAT changes only the mechanism by which items are assigned to respondents. The assumption used in current NAEP IRT analysis—that the "missing" item responses are missing at random (MAR) (Rubin, 1976)—remains valid because in a CAT, the missingness mechanism depends only on observed data.

Two notable weaknesses in the literature on mode effects limit the extent to which it can be used to anticipate the effects that might be observed with NAEP. First, most studies consider only a single point in time, whereas NAEP is primarily an instrument to measure change. One might assume that a computerized test that appeared to measure the same constructs, in the same way, as an existing paper-and-pencil test at one point in time would also yield comparable trend results; however, there has been little, if any, empirical investigation of this question. A second weakness in the existing literature is that it is relatively silent on the question of whether gaps in scores among subpopulations may appear different, depending on whether computerized or paper-and-pencil tests are used. These two kinds of questions, on the pattern of results over time and between groups, should probably be the foci of research on the effects of the computerization of NAEP.

NAEP as Lingua Franca: Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests

For the past two decades there has been continuing interest in linking the scales of other assessments to the NAEP scale in order to obtain more value from expensive data collection efforts by producing linked results that can be compared with data from additional sources. These efforts have been successful to varying degrees, and a large number of cautions have been offered about their usefulness (Thissen, 2007; Linn, McLaughlin, & Thissen, 2009).

However, especially under scenario 1 (described above)—in which the states are divided roughly into halves using one of two assessments—the linking landscape changes in two ways. First, although only limited practical strategies exist for linking NAEP to 50 statewide tests, more data collection options are practical for linking NAEP to a universe of two consortium assessments. Second, the two consortia assessments are still in their planning stages and a window of opportunity exists during which they might be designed to incorporate linking data collection.

The strong suggestion made here is that the scales of the assessments from the two consortia should be linked to NAEP. In August 2011, the National Center for Education Statistics (NCES) convened a group of experts in assessment, measurement, and technology for a summit on the future of NAEP, and this was followed by a second summit of state and local stakeholders in January 2012. NCES then assembled a panel of experts from the first summit, chaired by Edward Haertel, to consider and further develop the ideas from the two discussions, and make recommendations that were summarized in a report to the Commissioner of NCES (NCES Initiative on the Future of NAEP, 2012). That report proposed "the development of mechanisms for flexible linking of NAEP to other scales. This would include reweighting of content within NAEP if necessary, so as to maximize alignment with any of a range of large-scale assessment programs, including the Smarter Balanced and PARCC summative assessments as well as PISA [Program for International Student Assessment], the Progress in International Reading Literacy Study (PIRLS), TIMSS [Trends in International Mathematics and Science Study], and others" (p. 40). To facilitate linkage, the panel placed high priority on "studies of NAEP design changes to facilitate linkages between NAEP and other large-scale assessment programs, including the summative assessments developed by the PARCC and Smarter Balanced consortia at grades 4, 8, and possibly 12" (p. 47).

A Manageable Design Based on a Great Deal of Cooperation

In the past, linkages among disparate assessments have rarely been symmetrical efforts in which the linking data are collected in the naturally occurring context for both tests. However, attempts to link the scales of the PARCC and/or Smarter Balanced assessments with NAEP may be different. With PARCC and Smarter Balanced still in the planning stages, it may be possible to design linking data collections that symmetrically embed NAEP blocks or items within the PARCC and/or Smarter Balanced assessments, and embed PARCC and/or Smarter Balanced items within operational administrations of NAEP. We note that the *Future of NAEP*

report (2012) suggested consideration of "main NAEP data collection with expanded slots for: (1) linking items; and (2) experimental item types" (p. 48) to facilitate such symmetrical linking.

If such symmetrical data were collected, questions about the effect of context on each assessment's item responses could be resolved empirically, and threats to the validity of linkage would be subject to data analysis. The strategy would, moreover, be amenable to many (technical) forms of linking. If both PARCC and Smarter Balanced participate along with NAEP, not only might the scales of both consortia be linked to the NAEP scale, but the PARCC and Smarter Balanced scales may be (implicitly) linked to each other. Thus, such linkage could serve to align the two "stacks" of statewide results, one for the PARCC states and the other for the Smarter Balanced states.

The question of what to do with the states that participate in neither PARCC nor Smarter Balanced would remain. However, those states would have NAEP results and possibly greater motivation to participate in one of the consortia because comparability of scores would add value to the products of both consortia.

Conclusion

Computerization of NAEP is inevitable. Indeed, recent discussion of assessment schedules by the National Assessment Governing Board suggest that all NAEP assessments (with the possible exception of the long-term trend assessment) may be computer-administered by 2019; some will be computerized earlier and some have already been computerized. There are several reasons for computerization. NAEP assessments may be computerized so that technology-enhanced item types can be delivered when required by the frameworks, as has already happened with the science assessment in 2009 and is planned for the TEL assessment in 2014. NAEP assessments may be computerized so that they appear more comparable with statewide assessments being developed by the consortia or to facilitate linking with those assessments. They may be computerized simply because computer administration has become more cost effective—this will ultimately happen for all assessments as the cost of computing equipment decreases and the costs of printing and physical distribution and scoring of paper response sheets grow. Finally, all assessments will gradually become computerized as computer use becomes ubiquitous for real-world tasks, both within and outside schools.

From the literature on the computerization of other assessments, it is clear that comparability of results can usually be maintained as a test makes the transition from paper-and-pencil to computerized administration. It is also clear that, sometimes, some aspect of computerization may have an effect on results for some subgroups of the population. This suggests that the computerization of NAEP is best approached in the way that all other changes made to NAEP assessments since the advent of the "new design" in 1983 have been approached: careful consideration should be given to the design of the computerized administration, and a bridge study should be carried out to ensure comparability of results across the transition (unless an *a priori* decision is made to "break trend" regardless).

At an unlikely extreme, it is *possible* that in some subject matter areas a computerized NAEP might be found to measure the relevant constructs sufficiently differently that choices would have to be made between "breaking trend" and using the new assessment, continuing with the paper-and-pencil measure for the sake of continuity, or creating another parallel NAEP, with the old paper-and-pencil measure running alongside a new computerized assessment (much as the NAEP's long-term trend assessment has run in parallel with the new design for the past three decades). Although this possibility is not likely (given accumulated experience with computerizing existing assessments), it is best to avoid *a priori* rejection of any possibility.

Assessments developed by Smarter Balanced and PARCC may reduce the number of statewide tests in Grades 4 and 8 from nearly 50 to the low single digits, starting in the 2014–15 academic year. Assuming this happens, it will change the ways in which NAEP can serve as a monitor of progress, as reflected by statewide tests. With such a small set of tests to work with, linkage may become feasible, permitting close quantitative comparison between NAEP results and those obtained with the consortia tests, and providing a mechanism to link the consortia tests' scales with

each other across the two groups of states. Historically, such linkage has been fraught with difficulties (Thissen, 2007; Linn, McLaughlin, & Thissen, 2009). However, linkage is better understood now than in previous decades, and there is agreement on the technical approaches required.

One result that is clear from the literature on linkage is that relations between the results of disparate educational assessments tend to change over time. This means that any linkage between the NAEP scale and the consortia statewide tests will need to be maintained regularly over the years of their use. However, we note that a singular opportunity exists in a short window of time—essentially right now—to design data collection for linkage between the NAEP scale and the consortia assessments while the latter are under development. At this time, central control remains possible, and cooperative agreements to collect suitable linking data may be more easily obtained than will be the case after the consortia tests branch and fork into two dozen statewide assessments. This opportunity is very attractive, and may spur computerization of some NAEP assessments so that parts of those assessments can be embedded by the consortia in item tryout or first operational administration, and vice-versa in NAEP in the 2014–15 time frame.

A useful side effect of embedded-block linkage of the new consortia tests with the NAEP scale during development may be that the process will help explain to policymakers any change that may arise in results reported by pre- and post-consortia statewide tests. The new tests, with associated new standards, may appear to suggest large changes in the proportions of students categorized as "proficient" in many states; such changes have, historically, been the reason that linkages have been found to change over time (Thissen, 2007; Linn, McLaughlin, & Thissen, 2009). Linkage of the results to some stable scale, like that of NAEP, could help consumers of the results distinguish between real change and artifactual "change" arising solely from new assessments or standards.

Looking ahead, we see that the only constant in educational assessment is change. NAEP has a long history of implementing gradual change so that results remain comparable from year to year, while, at the same time, the assessments remain relevant in the presence of continuing educational and curricular change. We expect that spirit of gradual incremental change will continue to guide NAEP's adaptation to the new environment of the second decade of the 21st century.

References

Gewertz, C. (2012). Test group rethinks questions. *Education Week*, *32*(13), 1–24.

Linn, R. L., McLaughlin, D., & Thissen, D. (2009). *Utility and validity of NAEP linking efforts*. A publication of the NAEP Validity Studies Panel. Washington, DC: American Institutes for Research.

NCES Initiative on the Future of NAEP. (2012). *NAEP: looking ahead—Leading assessment into the future*. A white paper commissioned by the National Center for Education Statistics. Washington, DC: U.S. Department of Education.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, *63*, 581–592.

Sands, W. A., McBride, J. R., & Waters, B. K. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*, 849–869.

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. 287–312). New York: Springer.

Appendix A. Membership in the PARCC and Smarter Balanced Consortia†

PARCC	Both	Smarter Balanced
Arizona*	North Dakota	Alaska
Arkansas*	Pennsylvania	California*
Colorado		Connecticut*
District of Columbia*		Delaware*
Florida*		Hawaii*
Georgia*		Idaho*
Illinois*		Iowa*
Indiana*		Kansas*
Kentucky		Maine*
Louisiana*		Michigan*
Maryland*		Missouri*
Massachusetts*		Montana*
Mississippi*		Nevada*
New Jersey*		New Hampshire*
New Mexico*		North Carolina*
New York*		Oregon*
Ohio*		South Carolina *
Oklahoma*		South Dakota*
Rhode Island*		Vermont*
Tennessee*		Washington*
		West Virginia*
		Wisconsin*
		Wyoming
	Neither	
	Alaska	
	Minnesota	
	Nebraska	
	Texas	
	Utah	
	Virginia	

* "Governing" states.

† Membership was compiled from the websites of the two consortia, http://www.parcconline.org/parcc-states for PARCC and http://www.k12.wa.us/SMARTER/States.aspx for Smarter Balanced on December 3, 2012. Membership in the two consortia has been somewhat fluid; these lists differ from the lists provided in the June 2010 Race to the Top applications.

Appendix B. Computer-Based Assessment: A Review of the Last 15 Years of Comparability Research

Sharyn Rosenberg, *American Institutes for Research*
Reanne Townsend, *American Institutes for Research*

As personal computers and other technologies become more advanced, and more prevalent among the U.S. population, it is becoming increasingly important to use these tools to improve and enhance educational assessment. The National Center for Education Statistics (NCES), which oversees the development of the National Assessment of Educational Progress (NAEP), recognizes this trend and plans to have NAEP fully computer-based by 2022. However, this transition cannot be made lightly; it is important to determine whether scores obtained from computer-based testing can be expected to be statistically comparable with those obtained from the previous paper-and-pencil based administrations, and whether meaningful comparisons can be made between the two modes. In other words, can trend be maintained in reporting?

In 1999, NCES commissioned two experimental studies—one for writing and one for mathematics—to examine potential mode effects when comparing paper-and-pencil based tests and computer-based administrations of NAEP. The writing online study, conducted in 2002 using nationally representative samples from Grade 8 main NAEP, found no differences in performance when comparing scores from the paper- and computer-based administrations overall or by subgroup, with one exception; students from urban schools performed significantly better on the paper test than the computerized test, with an effect size of 0.15 (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). The mathematics online study, conducted in 2001 using nationally representative samples from Grade 8, found that overall scores were 4 points lower for the computer-based administration than for the paper version; several item difficulty parameters varied substantially across the two modes, indicating that the mathematics test did have score differences by mode (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008).

The NCES experimental studies on mode effects are very informative, but they were performed on limited subjects and at a single grade, during a time period when computer use in schools for learning and assessment purposes was much less common. The purpose of this review is to examine research addressing the comparability of computer-based assessments and paper-and-pencil based tests as one way of informing expectations for a broader application of computerized NAEP.

The mode effects of computer-delivered tests and surveys have been the subject of investigation since the mid-1980s; however, the nature of interaction with computer-based technology has changed drastically since then. In light of this, and because of

the great breadth of literature available on the subject, this paper examines 65 comparability studies of academic assessments during the last 15 years (since 1997).[19]

Investigating Measurement Equivalence by Mode

In the literature, there is substantial variation in the approach taken to define and measure comparability between paper-and-pencil and computer-based testing. Of the 65 journal articles and conference presentations reviewed here, 21 included an investigation of measurement equivalence (Vandenberg & Lance, 2000) across modes; factor analyses, item response theory analyses, and/or differential item functioning (DIF) analyses were used to determine whether or not an assessment was measuring the same construct in the paper-and-pencil version as in the computer-based version. The remaining 44 studies purported to measure mode effects by analyzing whether there were mean differences between scores produced by paper-and-pencil and computer-based versions of the same assessment. Importantly, in the latter approach, potential differences between constructs across modes may be confounded with differences in mean scores.

The literature review found 21 studies that evaluated potential mode effects by measuring the extent to which an assessment measured the same construct in paper-and-pencil and computer-based formats; these are listed in Tables B1–B2. The results of most of these studies (14 out of 21) found no threats to measurement equivalence. (See Table B1.) Six studies found mixed results, and one study concluded that the assessment generally was not measuring the same construct across modes. (See Table B2.) In general, the more holistic confirmatory factor analysis approach found that paper-based and computer-based versions of the same assessment typically were measurement invariant, at least at the level of configural invariance (where patterns of free and fixed factor loadings were similar across modes) and metric invariance (where item factor loadings were similar across modes) (Horn & McArdle, 1992). The item-by-item approach employed by differential item functioning (DIF) analyses generally led to mixed results, ranging from no evidence of DIF (Taherbhai, Seo, & Bowman, 2012) to 38 percent of items flagged for DIF across modes (Gu, Drake, & Wolfe, 2006).

In many cases, there was no relationship between whether a study found measurement equivalence of constructs across modes and whether there were significant score differences by mode. Of the 14 studies that found measurement equivalence across modes, five concluded that there were no statistically significant score differences by mode either overall or by subpopulation (Karkee, Kim, & Fatica, 2010; Lottridge, Nicewander, & Mitzel, 2011; Randall, Sireci, Li, & Kaira, 2012; Schroeders & Wilhelm, 2011; Staples & Luzzo, 1999). Five studies concluded

[19] A librarian performed the literature search in ERIC by searching for experimental studies related to "mode effects," "comparability," "computer-based assessments," "paper-pencil assessments," and several other variations of these terms. Articles were also added by searching the reference lists of existing studies. Included studies were limited to education and certification exams administered to students (up to and including the college level). The review was limited to studies in which the same or equivalent students took paper-pencil and computerized versions of an assessment; simulation studies, literature reviews, and thought pieces were excluded.

that the computer-based assessment was associated with significantly lower scores than the paper-based version (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Pomplun, 2007; Pomplun & Custer, 2005; Rowan, 2010; Taberbhai, Seo, & Bowman, 2012), and one study found that the computer-based assessment was associated with significantly higher scores than the paper-based version (Pomplun, Frey, & Becker, 2002). The results of the remaining three studies (Choi, Kim, & Boo, 2003; Kim & Huynh, 2007; Kim & Huynh, 2008) were mixed.

Construct equivalence is a necessary condition for comparing mean scores across modes, and the majority of studies in the literature review did not include analyses of measurement equivalence. Given that the studies reviewed that focus on paper-based assessments were also administered by computer with minimal adaptation, and that 20 of the 21 measurement equivalence studies found full or partial measurement equivalence across modes, we extrapolate that the score differences of the remaining 44 studies likely can be analyzed by mode. Therefore, the remaining sections incorporate all 65 studies. The complete list of the studies (and capsule summaries of their findings) can be found in Tables B1–B7.

Investigating Score Differences by Mode

Of the 65 studies reviewed, 11 found consistent differences in scores between computer-based versions and paper-based versions of the same assessment; four studies found that the computer-based format was associated with higher scores than the paper-based format, and seven studies found that the computer-based format was associated with lower scores than the paper-based format. Nineteen studies found no significant score differences by mode, either overall or by subgroup. The majority of the studies reviewed (35 out of 65) found some score differences across mode, but the results varied by content area, ability, subgroup, and/or other dimensions of the assessment or students.

Despite the lack of consistent mode effects for all students in most of the research, the many studies that found significant mode effects under specific circumstances have important implications for NAEP. As NAEP transitions to computer-based testing, it is important to recognize that certain subjects or subpopulations may be more substantially affected than others by the change in delivery mode. For example, computer-based assessment introduces new possibilities for integrating testing accommodations into the main assessment, including some aspects of universal design that make certain features available to all students (Dolan, Hall, Banerjee, Chun, & Strangman, 2005; Lee, Osborne, & Carpenter, 2010). However, it is important to ensure that new features of the computer interface do not introduce construct-irrelevant variance. The literature uncovers several issues related to mode effects from computer-based administration that include aspects of the assessments and the participants, as well as interactions between the two.

Mode Effect and Assessment Characteristics

Although many innovative item formats are made possible through the use of computer-based assessment, most test developers with an interest in maintaining trend or investigating comparability with previous paper-based versions have simply

chosen to transfer more traditionally formatted items to computer-based administration. Unfortunately, the literature shows that this does not completely eliminate mode effects associated with assessment characteristics. Transferring a test from a paper-based version to the computer involves changes to item formats, but also has mode-specific implications related to the tools that students access to answer the questions and the perceptual differences by human scorers across modes.

Many of the studies with mixed results for score differences found that scores varied by mode for only a subset of the subject areas and/or grades tested, but there were few clear patterns among the results. For most subject areas, there were no consistent findings in terms of whether the computer-based version or paper-based version was more difficult or whether they were equivalent. For mode comparisons of mathematics tests, the majority of studies found either that the computer-based version was associated with significantly lower scores than the paper-based version, or that there was no significant difference across modes. Only one study (Kingsbury, 2002) found significantly higher mathematics scores for the computer-based condition than the paper-based condition, after controlling for students' initial performance, and the difference was small (about one point).

Gu, Drake, and Wolfe (2006) found that mathematics items that involved equalities/inequalities and variables were most likely to exhibit DIF by being more difficult on paper than on a computer as compared to other item types. Johnson and Green (2006) found that participants' scores were significantly lower on computer-based items that required scratch paper than on paper-based versions of the same items. Similarly, another study found that scores on items involving graphic and geometric manipulation were negatively affected by computer-based administration (Keng, McClarty, & Davis, 2008).

Other assessment characteristics that were found to affect comparability include item format and whether the computer-based test was linear (i.e., fixed form) or adaptive. Russell and Haney (1997) found no significant score differences by mode for multiple-choice items but significantly higher scores for the computer-based version of performance writing tasks and short-constructed response items compared with the paper-based version. In a meta-analysis, Kim (1999) found that computer-adaptive tests were associated with significantly lower scores than paper-based tests, while computerized tests that were not adaptive were associated with significantly higher scores than paper-based tests. In a separate meta-analysis, Wang, Jiao, Young, Brooks, and Olson (2008) found that effect sizes between computer-based tests and paper-based tests were significantly larger when the computerized version was adaptive than when it was linear.

In addition to comparability issues related to the assessment content, the mode of administration has been shown to affect perceptions of human scorers. Systematic differences in how paper-based and computerized assessments are scored also can lead to differences in student performance across the two modes. Several studies have examined the effect of composition mode on scores for written essays and constructed-response items. In general, these studies found that human scorers, on average, assigned higher scores to handwritten papers compared with typed essays, although typed essays

were longer, on average, and students generally preferred to work on computers. Researchers speculated that this difference was due to scorers being more lenient and forgiving of smaller errors when reading the handwritten essays (Russell & Tao, 2004; Way & Fitzpatrick, 2006). However, Russell and Tao (2004) found that this mode effect in scoring could be eliminated when scorers were made aware of the effect and given proper training. Further, Russell and Plati (2000a) conducted a study in which handwritten essays were later typed and provided to scorers blind to the original mode of composition. This study found that when scorers were blind to composition mode, essays originally written on computer were significantly longer and received significantly higher scores. Although the NAEP writing assessment transitioned to computer administration when it moved to a new framework in 2011, there are important implications for constructed-response items in other subject areas. Mode effects related to scoring will be particularly important for NAEP to examine given the large proportion of constructed-response items on NAEP assessments.

Mode Effects and Demographic Characteristics

Several studies also found that score differences between computer-based and paper-based tests varied by demographic characteristics, including gender, race, socioeconomic status, student ability, urbanicity, and SD/ELL (students with disabilities/English language learners) status.

Several studies have investigated mode effects by gender. A study by Gallagher, Bridgeman, and Calahan (2002) found that female performance was negatively affected by computers as the mode of test administration. In particular, the often-observed discrepancy between male and female performance on mathematics items grew significantly larger under the computer-administration condition. A similar effect was found in a study by Horne (2007) of a language arts and spelling test on which females performed significantly better than their male counterparts on a paper-based version; this score difference was eliminated in the computer-based version of the same assessment. However, several other studies (Bridgeman & Cooper, 1998; Clariana & Wallace, 2002; Fritts & Marszalek, 2010; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; MacCann, 2006: Randall, Sireci, Li, & Kaira, 2012) found no consistent mode effects as a function of gender.

Results from surveys in a Hong Kong-based study showed that, when given the choice, male participants preferred to take their tests using computers, while females tended to opt for paper-and-pencil administered assessments (Coniam, 2006). Fritts and Marszalek (2010) found no significant difference by gender on measures of test anxiety, regardless of whether the test was taken by paper-and-pencil or computer administration.

It is not clear whether differential mode effects by gender indicate a disadvantage for females taking tests on computers, or whether the computer mode increases motivation and engagement for males, thus eliminating some of the construct-irrelevant variance in paper-based tests.

Several studies have investigated whether mode effects are more pronounced for students with low socioeconomic status (SES). Pomplun and Custer (2005) and

Pomplun, Ritchie, and Custer (2006) found that students eligible for free or reduced-price lunch had greater gaps between scores from paper-based and computer-based versions of an elementary reading assessment. On a computing skills test of high school students in Australia, MacCann (2006) found that although there were no score differences by mode for high SES students, low SES students performed significantly better on the paper-based version than the computer-based version of the test. Although not a study of SES directly, Horkay, Bennett, Allen, Kaplan, and Yan (2006) found a significant interaction between mode and school location, a variable often correlated with SES. Students from urban fringe/large town locations performed significantly better on the paper-based version than the computer-based version of a writing test On a state 10th- grade science assessment, Randall, Sireci, Li, and Kaira (2012) found no consistent mode effect between students who were eligible for free or reduced-price lunch and those who were not eligible.

Another population that has been the focus of several mode effect investigations is SD/ELL students. Despite the use of universal design elements that incorporated some accommodations into the general assessment, several studies found that SD/ELL students performed significantly better on paper-based versions than computer-based versions of language arts (Russell & Plati, 2000b) and reading and mathematics (Taberbhai, Seo, & Bowman, 2012) tests. Wolfe and Manalo examined TOEFL Writing results from nearly 134,000 English language learners and found that participants with lower English language ability scored higher on the paper-based version, and students with higher English language ability scored higher on the computer-based version. Bridgeman and Cooper (1998) found no significant interactions between mode and ELL status for the GMAT. Conversely, Dolan et al. (2005) used a small sample of 10 SDs and found no significant mode effect overall; however, scores were significantly higher in the computer-based version as compared with the paper-based version for items with reading passages that were more than 100 words. Finally, Kim and Huynh (2010) performed differential bundle functioning analyses on a statewide, end-of-course English assessment and found that "Researching items" significantly favored the paper mode for students without disabilities and "Building Vocabulary items" significantly favored the computer mode for SDs. Although there is not a clear pattern in these results, what stands out is the complexity of how mode of testing administration can interact with both SD/ELL status and other factors. It is not clear whether SD/ELL students generally have less experience with computers, which could also account for performance differences.

Mode Effects and Computer Familiarity

Perhaps the most important student characteristic to consider when examining the mode effect of computer-based assessment administration is computer familiarity. Although many studies have examined the impact of computer familiarity on mode effects in assessment, the relationship between computer experience and performance on computer-based assessments remains unclear. Several studies show that higher levels of computer familiarity correlate with higher scores on computer-based assessments (Bennett, Braswell, Oranje, Sandene, Kaplan, &Yan, 2008; Bridgeman & Cooper, 1998; Chen, White, McCloskey, Soroui, & Chun, 2011; Horkay, Bennett, Allen, Kaplan, & Yan, 2006); one found mixed results (Goldberg &

Pedulla, 2002); others have found no significant effect on mode by computer familiarity (Clariana & Wallace, 2002; Higgins, Russell, & Hoffman, 2005).

One possible reason for this inconsistency in results is that computer familiarity is still a vaguely defined construct that has yet to be operationalized consistently across studies. Because there is no standard measure of computer familiarity, the construct being measured as "computer familiarity" is not necessarily consistent between studies. For example, Horkay et al. (2006) developed their own, study-specific survey to determine participants' levels of computer familiarity. Clariana and Wallace (2002) measured computer familiarity using four previously developed questions from the Distance Learning Profile (Clariana & Moller, 2000). Higgins, Russell, and Hoffmann (2005) broke the construct into three parts: computer fluidity and computer literacy, for both of which they created their own metric; and frequency of computer use, for which they used a survey adapted from a fifth-grade USEIT (Use, Support, and Evaluation of Instruction Technology) study survey, developed by Russell, Bebell, and O'Dwyer (2003).

One specific aspect of computer familiarity is keyboarding skills. Mode effects on students with low keyboarding skill levels have been of particular concern recently, as NAEP pilots its new Writing Computer Based Assessment (WCBA) at the fourth-grade level. Studies by Russell (1999) and Russell and Plati (2000a and 2002) have found that keyboarding skills significantly affect student performance on writing tasks, but apparently only at the lower skill levels; there appears to be a skill level "threshold", above which keyboarding skills seem to have no significant effect.

A similar effect to the "threshold" described in Russell's (1999) investigation of keyboarding skills is observed in other equivalency studies investigating computer experience and familiarity. It is possible that computer familiarity is much more predictive of a mode effect for certain subpopulations and may account for some of the differential mode effects observed for certain subgroups. However, the majority of studies addressing computer familiarity were not performed within the past five years, and it is likely that computer familiarity has greatly increased during this time.

The results from studies examining computer familiarity highlight the confounding role of demographics, making it particularly difficult to isolate and confirm the myriad factors involved in mode effects in computer-administered assessments. Although the extent to which familiarity with computers affects performance on computer-based assessments is still unclear, there is enough evidence to suggest that familiarity should be taken into account when moving to computer-based assessments, and steps should be taken to mitigate these effects as much as possible.

Conclusion

Many studies examining mode effects in educational testing have shown inconsistent or mixed effects. The research is clear in demonstrating that comparability of results *can* often be maintained overall as a test makes the transition from paper-and-pencil to computerized administration. For example, most of the studies suggest that the structure of the test is likely to remain unchanged in moving from paper-and-pencil to computer-based administration. However, the evidence is mixed on the effects of

mode on score comparability; computerization may have an effect on the results for some subgroups of the population and these can vary further as a function of the subject area being assessed. Schroeders and Wilhelm (2011) perhaps best summarize what is required when moving to computerized assessment when they write "… equivalence research is required for specific instantiation unless generalizable knowledge about factors affecting equivalence is available" (pg. 1).

This sentiment should also help guide and inform the move to computer-based assessment in NAEP. The computerization of an assessment should be treated as any other change one might make in NAEP: comparability of scores can be hoped for, but cannot be taken for granted. Research, including the use of bridge studies, is needed to evaluate the effects of moving assessments from paper-and-pencil to computer administration.

References for Appendix B

Anakwe, B. (2008). Comparison of student performance in paper-based versus computer-based testing. *Journal of Education for Business, 84*(1), 13–17.

Balizet, S., Treder, D., & Parshall, C. (1999, April). *The development of an audio computer-based classroom test of ESL listening skills.* Paper presented at the annual meeting of the American Educational Research Association, Quebec, Canada.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6, 9.*

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31*(1), 51–60.

Bridgeman, B., & Cooper, P. (1998, April). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*(1), 49–71.

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20,* 295–320.

Clariana, R. B., & Moller, L. (2000). *Distance learning profile instrument: predicting on-line course achievement.* Paper presented at the annual convention of the Association for Educational Communications and Technology, Denver, CO.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593–602.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL, 18*(2), 193–211.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation, 15*(3), 243–263.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning and Assessment, 3*, 7.

Escudier, M. P., Newton, T. J., Cox, M. J., Reynolds, P. A., & Odell, E. W. (2011). University students' attainment and perceptions of computer delivered assessment: A comparison between computer-based and traditional tests in a "high-stakes" examination. *Journal of Computer Assisted Learning, 27*(5), 440–447.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Journal Psychology of Education, 13*(3), 441–458.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal, 53*(4), 289–299.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133–147.

Goldberg, A. L., & Pedulla, J. (2002). Performance differences according to test mode and computer familiarity on a practice Graduate Record Exam. *Journal of Educational and Psychological Measurement, 62*(6), 1053–1067.

Gu, L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *Journal of Technology, Learning, and Assessment, 5*(4), 1–25.

Higgins, J., Patterson, M. B., Bozman, M., & Katz, M. (2010). Examining the feasibility and effect of transitioning GED tests to computer. *Journal of Technology, Learning, and Assessment, 10*, 2.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3,* 4.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on a computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 5,* 2.

Horne, J. (2007). Gender differences in computerised and conventional educational tests. *Journal of Computer Assisted Learning, 23*(1), 47–55.

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4,* 5.

Karkee, T., Kim, D. I., & Fatica, K. (2010). *Comparability study of online and paper and pencil tests using modified internally and externally matched criteria.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21*(3), 207–226.

Kim, D. H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *Journal of Technology, Learning, and Assessment, 6,* 4.

Kim, D. H., & Hyunh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*(4), 554–570.

Kim, D. H, & Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide English test for students with and without disabilities. *Educational Assessment, 15*(2), 107–121.

Kim, J. P. (1999). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Kingsbury, G. G. (2002). *An empirical comparison of achievement level estimates from adaptive tests and paper-and-pencil tests.* Presented during the annual proceedings of the American Educational Research Association, New Orleans, LA.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education, 2*, 22–37.

Lee, K. S., Osborne, R. E., & Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: Computerized vs. paper-pencil/regular vs. extended time. *Journal of Educational Computing Research, 42*(4), 443–458.

Liao, C. H., & Kuo, B. C. (2011). A web-based assessment for phonological awareness, rapid automized naming (RAN) and learning to read Chinese. *Turkish Online Journal of Educational Technology—TOJET, 10*(2), 31–42.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011). A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research, 46*(3), 544–566.

MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37*(1), 79–91.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24*(1), 29–39.

Minnick, J. R. (2009). *Motherboards cutting razor wire: Assessment and incarcerated youth*. Doctoral dissertation, Capella University.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *ALT-J: Research in Learning Technology, 18*(1), 29–47.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3,* 6.

Pommerich, M. (2002, April). *The effect of administration mode on test performance and score precision, and some factors contributing to mode differences*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2,* 6.

Pommerich, M., & Burden, T. (2000, April). *From simulation to application: Examinees react to computerized testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pomplun, M. (2007). A bifactor analysis for a mode-of-administration effect. *Applied Measurement in Education, 20*(2), 137–152.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K–3 reading tests. *Journal of Educational Computing Research, 32*(2), 153–166.

Pomplun, M., Frey, S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337–354.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127–143.

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice, 31*(4), 2–12.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning and Assessment, 6*, 3.

Rowan, B. E. (2010). *Comparability of paper-and-pencil and computer-based cognitive and non-cognitive measures in a low-stakes testing environment*. Doctoral dissertation, James Madison University.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*, 20.

Russell, M., Bebell, D., & O'Dwyer, L. (2003). *Use, support, and effect of instructional technology study: An overview of the USEIT study and the participating districts*. Boston: Technology and Assessment Study Collaborative.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing students' performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3), 1–20.

Russell, M., & Plati, T. (2000a). Effects of computer versus paper administrations of a state-mandated writing assessment. *The Teachers College Record: Technology and Assessment Study Collaborative, 103*, 1–34.

Russell, M., & Plati, T. (2000b). *Mode of administration effects on MCAS composition performance for grades four, eight, and ten*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer, and portable writing devices. *Current Issues in Education, 5*, 4.

Russell, M., & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research & Evaluation [Online], 9*, 10.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*(5), 849–869.

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Staples, J. G., & Luzzo, D. A. (1999). Measurement comparability of paper-and-pencil and multimedia vocational assessments. *ACT Research Report Series 99-1*. Iowa City, IA: American College Testing Program.

Taherbhai, H., Seo, D., & Bowman, T. (2012). Comparison of paper-pencil and online performances of students with learning disabilities. *British Educational Research Journal, 38*(1), 61–74.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K–12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Way, W. D., & Fitzpatrick, S. (2006). *Essay responses in online and paper administrations of the Texas Assessment of Knowledge and Skills (*Research Report 06-04). Pearson Educational Measurement.

Whiting, H., & Kline, T. J. B. (2006). Assessment of the equivalence of conventional versus computer administration of the Test of Workplace Essential Skills. *International Journal of Training and Development, 10*(4), 285–290.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology, 8*(1), 53–65.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education, 29*(4), 423–445.

Table B1. Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Bennett, Braswell, Oranje, Sandene, Kaplan, Yan	2008	National Assessment of Educational Progress (NAEP) PPT, 2001 Mathematics Online (MOL)	Independent *t*-test; item response theory analyses	1,970 Grade 8 students (nationally representative)	Computer facility predicted MOL performance (controlled for performance on paper-based test). Eighth-grade performance was significantly lower for those taking the computerized test, with an effect size of 0.15. At the item level, the difficulties for the computer test were generally greater and item discrimination differences estimates suggested minimal effects.
Choi, Kim, Boo	2003	Test of English Proficiency by Seoul National University (TEPS); listening comprehension, grammar, vocabulary, reading comprehension	Correlational analyses; analysis of variance (ANOVA); confirmatory factor analyses	971 university students in Korea	Statistically significant score differences were found among the listening comprehension, vocabulary, and reading comprehension subtests, but not for the grammar subtest. The factor structure for the four subtests was consistent across test administration modes. Correlations of subtests, disattenuated correlations, and confirmatory factor analyses support that the computer-based and paper-based subtests measure the same constructs.
Karkee, Kim, Fatica	2010	End-of-instruction social studies assessment	Item response theory and differential item functioning (DIF) analyses	50,000 participants	No statistically significant mode effect was found based on model fit, DIF, or student performance.
Kim, Huynh	2007	End-of-course assessments in algebra and biology	Counter-balanced repeated-measures ANOVA; item response theory analyses and comparison of information functions; confirmatory factor analyses	Students from 15 middle and high schools in a southeastern state (788 algebra students and 406 biology students); Black and Hispanic students were underrepresented.	No evidence was found to suggest that mode changed the constructs measured. Results suggest the comparability of computer-based and paper-based assessments at the item-, subtest- and whole test-levels. For algebra, scores were significantly higher for the paper-based assessment than the computer-based assessment, with an effect size of 0.17. For biology, there were no significant score differences by mode.

Table B1 (continued). Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Kim, Huynh	2008	NC end-of-course English test	Two-way repeated measures ANOVA; item response theory analyses; confirmatory factor analyses	439 middle- and high-school students; Black students were under-represented.	Students scored significantly higher on the paper-based assessment than the computer-based assessment overall with a small effect size. Results from the confirmatory factor analyses suggest the mode does not alter the test constructs. Analysis at the content domain level indicates that students perform worse in reading comprehension in a computer mode; however, there were no differences by mode in the other content domains.
Lottridge, Nicewander, Mitzel	2011	End-of-course algebra and English Assessments	Comparison of within-subjects design and propensity score matching; confirmatory factor analyses	3,628 students in Grades 8, 9	The study showed that the online and paper tests appeared to be measuring the same underlying constructs with the same level of reliability. The computer mode was slightly more difficult than the paper mode, but it is not clear whether the difference was statistically significant.
Pomplun	2007	Initial-Skills Analysis (part of the Basic Early Assessment of Reading)	Single-group counterbalanced design; bifactor model to test equivalence of paper-based and computer-based formats	About 2,000 students in K–3 across 12 states	Mean scores were significantly higher for the paper-based assessment compared with the computer-based assessment for all grades, with effect sizes ranging from .27 to .48. At each grade level, the model with the method factors included led to significant improvement in fit. There were some minor differences in the item factor loadings across formats. The authors concluded that score equivalence was found between the two modes but that the increased difficulty of the computerized version would require test equating to use results from the two modes interchangeably.

Table B1 (continued). Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Pomplun, Custer	2005	Initial-Skills Analysis (part of the Basic Early Assessment of Reading)	Single-group counterbalanced design; dependent *t*-tests, confirmatory factor analyses	About 2,000 students in K–3 across 12 states	Mean scores were significantly higher for the paper-based assessment compared with the computer-based assessment for all grades, with effect sizes ranging from .27 to .48. At three out of four grades, the test variance was significantly different across modes. Free/reduced-price lunch students had greater gaps between paper-based assessment and computer-based assessment scores, though it is not clear whether the differences are statistically significant. Confirmatory factor analyses found equivalence between the modes.
Pomplun, Frey, Becker	2002	Nelson-Denny reading test	Counter-balanced design; dependent *t*-tests; coefficient alpha; linear and equipercentile equating; predictive validity with grades	215 college students	Computer-based assessment generally produced higher scores compared with the paper-based assessment, though not all score differences were significant. The variance of the two forms was equivalent. The predictive validity of scores was comparable between the two modes.
Randall, Sireci, Li, Kaira	2012	State science assessment	Confirmatory factor analyses; Rasch item response theory DIF analyses	1,439 students (computer condition) and 10 random samples of 1,439 students drawn without replacement from 95,422 students (paper condition) in Grade 10	Confirmatory factor analyses found partial measurement invariance by mode, sex, and socioeconomic status (SES). DIF analyses found a few items with possible DIF. There were no consistent differences across modes, either overall or by sex or SES.

Table B1 (continued). Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Rowan	2010	Archival data from mandatory university assessments: Natural World, Ver. 9 (NW9): cognitive scientific knowledge and reasoning, computer-based and paper-based versions; Attitude Toward Learning (ATL): noncognitive, computer-based and paper-based versions.	Confirmatory factor analysis; Mantel-Haenszel DIF analyses	About 4,000 college students	The paper-based assessment and computerized versions of the test were found to be tau-equivalent. Mean differences between test administration modes were found to exist with higher scores on the paper version than the computer version, with an effect size of .26. The author noted that scores would need to be rescaled to be equivalent across the two modes. Three items exhibited C-level DIF across modes.
Schroeders, Wilhelm	2011	English Reading and Listening Comprehension (dichotomous items): English as a second language	Multigroup confirmatory factor analysis	442 German high school students, Grades 9, 10, English language learners (high ability)	Scores were measurement invariant across modes for both reading comprehension and listening comprehension.
Staples, Luzzo	1999	Unisex Edition of the American College Testing Inventory (UNIACT), Inventory of Work-Relevant Abilities (IWRA)	Scale correlations; coefficient alpha; exploratory factor analyses	1,022 students, Grades 9, 11	Factor loadings and internal consistency appeared similar across modes. There were no differences in mean scores by mode.
Taherbhai, Seo, Bowman	2012	Modified Maryland School Assessment (mod-MSA) in reading and mathematics	Analysis of covariance (ANCOVA); DIF	About 5,500 students with disabilities in Grades 7, 8	Students with disabilities who took the paper-based assessment performed significantly higher than the students with disabilities who took the computer-based assessment in reading and mathematics across grades, with effect sizes ranging from 0.06 to 0.12. No C-level DIF items were found.

Table B2. Seven studies that investigated measurement equivalence between modes of assessment, and found some lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Gu, Drake, Wolfe	2006	60 quantitative (mathematics) items, similar to GRE, Original items created using POWERPREP (ETS 1999)	*t*-test; differential item functioning analyses	165 first-year graduate students; high computer familiarity	No significant score differences were found between paper-based assessment and computer-based assessment groups; 38% of items were flagged for cross-medium DIF. Of the assessment characteristics examined, mathematical notation and content appeared to contribute most significantly to DIF across media.
Keng, McClarty, Davis	2008	Texas Assessment of Knowledge and Skills	*t*-tests; DIF analyses	Grades 8 and 11: 2,546 for mathematics; 3,680 for reading; 2,898 for social studies; statewide	Several items showed evidence of DIF. The paper-based assessment group significantly outperformed the computer-based assessment group on selected mathematics (e.g., Spatial Relationships and Geometric Relationships) and reading objectives (e.g., Basic Understanding, Applying Critical Thinking Skills) at Grades 8 and 11. No significant differences were found for social studies or science at Grade 11.
Kim, Huynh	2010	Statewide End-of-Course English Assessment	*t*-tests; confirmatory factor analyses; differential item/bundle functioning analyses; quasi-experimental design using propensity score matching	~15,000 participants, (~1,000 SD), Grade 9	There were some significant interactions between disability status and mode for some of the content areas, though the effect sizes were very small (less than 0.1). The confirmatory factor analyses found measurement equivalence by mode at the weak, strong, and strict levels. The DIF analyses found no items with C-level DIF. The differential bundle functioning analyses did find a significant result favoring the paper-based mode for Researching items for students without disabilities and Reading III —Building vocabulary items for students with disabilities.

Table B2 (continued). Seven studies that investigated measurement equivalence between modes of assessment, and found some lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Poggio, Glasnapp, Yang, Poggio	2005	Kansas Computerized Assessment (large-scale state test) and parallel paper-based version	Descriptive statistics; hierarchical linear modeling; item response theory analyses	2,861 students in 7th grade	No meaningful statistically significant difference was found in performance between computer-based assessment and paper-based assessment scores (less than 1 percentage point); 9 of the 204 items were flagged as having mode effects, but no common factors were identified to account for this.
Puhan, Boughton, Kim	2007	Praxis—reading, writing, and mathematics	Cohen's *d*; DIF analyses	About 7,000 participants entering teaching programs	Based on Cohen's *d*, results indicated no substantial difference between computer-based and paper-based scores. DIF analyses revealed all reading and mathematics items were comparable for both versions. DIF analyses indicated item-level differences exist across the paper-based and computer-based versions of the writing test, with the three items favoring examinees who took the paper-based version.
Schwarz, Rich, Podrabsky	2003	InView (norm-referenced aptitude test); Test of Adult Basic Education (TABE) (norm-referenced)	DIF)	1. Grades 4–9; 2. Adults in large-scale, matched samples	Several items in each assessment did exhibit cross-medium DIF. On the TABE, differences by mode were largest at the lower end of the ability distribution.
Way, Davis, Fitzpatrick	2006	Texas Assessment of Knowledge and Skills (TAKS)—Mathematics, Reading, Science and Social Studies	Random-groups equating; matched-samples comparability analysis	Students in Grades 8, 11	Mixed results across subjects, with the largest difference for TAKS 8th-grade reading.

Table B3. Eight studies that evaluated effects of assessment characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Johnson, Green	2006	Selected mathematics items from UK's Mathematics National Curriculum	ANOVA	104 students ages 10–11	No statistically significant differences between overall performance on paper and computer.
Kim	1999	Meta-analysis, various subjects, mostly mathematics and reading	Various	Age range: Grade 3–adult, (about 50% university students)	The type of computer-based assessment was the most important variable when evaluating the equivalence between computer-based and paper-based tests. For adaptive tests, mathematics, source, and sampling age were significant variables. For nonadaptive computer-based tests, the analysis did not find significant moderators. Computer-based testing was significantly more advantageous for the high school-aged population.
Kingsbury	2002	ALT and Measure of Academic Progress (MAP) state tests in Idaho—reading, mathematics, language use	ANCOVA	8,560 students in 4th and 5th grades	Language usage and mathematics scores were significantly higher for computer-based tests than paper-based tests after controlling for initial performance (by about 1 point); there was no significant difference for reading scores.
Russell, Haney	1997	NAEP items (multiple-choice and short constructed-response language arts, mathematics, science, and reading items); unspecified open-ended writing items	Independent *t*-tests	114 students in Grades 6–8	No difference in multiple-choice test results by mode of administration. For the performance writing tasks, scores were significantly higher for computer-based tests than paper-based tests, with an effect size of .94. When scores of open-ended items were used as a covariate, there was a significant mode effect for short constructed-response items in science and language arts.
Russell, Plati	2000a	Massachusetts Comprehensive Assessment System (MCAS) Language Arts	Independent *t*-tests; Welch's *t*-tests	Students in Grades 8 (144) and 10 (145)	Scores were significantly higher for computer-based tests than paper-based tests at both Grades 8 and 10, regardless of keyboarding skills.

Table B3 (continued). Eight studies that evaluated effects of assessment characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Russell, Tao	2004	MCAS Composition items	ANOVA	Grade 8, 60 responses	Composition scores produced on computer (typed) received significantly lower scores than on paper (handwritten). Study found that upon training scorers using both modes, especially noting problems with mode effect, the presentation effect was eliminated.
Wang, Jiao, Young, Brooks, Olson	2008	Various mathematics assessments	Meta-analysis of mean score differences by mode (11 studies with 42 independent effects)	K–12	Meta-analysis found that overall there was no difference between scores from paper-based testing and computer-based testing. Effect sizes across the studies did vary, however, as a function of study design, sample size, computer practice, and computer delivery algorithm.
Way, Fitzpatrick	2006	Texas Assessment of Knowledge and Skills—Writing	Rater agreement; logistic regression; ANCOVA	1,340 Grade 11 lower performing students	Computer-based essays were scored more stringently than those completed on paper (handwritten). There was a positive relationship between essay score and the use of computers for language arts classes in the school. The paper-based test had higher interrater reliability of essay scoring than the computer-based test.

Table B4. Eleven studies that evaluated effects of demographic characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Bridgeman, Cooper	1998	Graduate Management Admissions Test (GMAT), two 30-minute essay items	Within-subjects; ANOVA	3,470	Significantly higher paper-based test scores compared with computer-based test scores for people with relatively low word-processing experience. No significant differences between paper-based test scores and computer-based test scores based on gender, race/ethnicity, or ELL status. Mode effect was smallest for participants with the most computer experience. Found higher interrater reliability for word-processed essays. Found no interaction of score differences by gender, race/ethnicity, or ELL status.
Clariana, Wallace	2002	100-item teacher-made multiple-choice course tests for introductory university class on computer fundamentals; Distance Learning Profile (Clarianna & Moller, 2000)	ANOVA (posttest only)	105 freshman university students	Overall, the computer-based testing group scored significantly higher than the paper-based testing group. Gender, competitiveness, and computer familiarity were not significantly related to performance difference between modes. There was a significant interaction between the administration mode and content familiarity. Low-attaining students had similar performance in both modes, while high-attaining students performed better on the computer-based test than the paper-based test.
Coniam	2006	English Language Listening Test	Posttest survey on preferences	Grade 11, 12 students in Hong Kong	Significantly higher scores for Grade 11 computer-based assessment than paper-based assessment; no significant score differences for Grade 12. Survey found males preferred computer-based tests and females preferred paper-based tests.

Table B4 (continued). Eleven studies that evaluated effects of demographic characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Dolan, Hall, Banerjee, Chun, Strangman	2005	Released items from NAEP U.S. history and civics	Matched-samples *t* -tests	10 students with specific learning disabilities from Grades 11 and 12	There were no significant differences overall between scores in the two modes. Scores were significantly higher in the computer-based test condition for items with reading passages more than 100 words. Usability interviews indicated that participants preferred the computer-based test.
Fritts, Marszalek	2010	Measure of Academic Progress assessment (MAP), ALT	Regression analyses; *t* – tests	132 students (mean age: 13.36)	There was no difference between the two groups in the standardized mathematics score or standardized reading score. The computer-based test was found to produce less test anxiety than the linear paper-based test. No significant mode effect was found by gender.
Gallagher, Bridgeman, Calahan	2002	Graduate Record Examination (GRE), SAT I, GMAT, Praxis	Standardized mean differences; repeated-measures ANOVA; *t* – tests	Several hundred thousand high school and college students	Mode effects varied by gender, race/ethnicity, and gender by race/ethnicity interactions across the different tests.
Horkay, Bennett, Allen, Kaplan, Yan	2006	Main NAEP— Writing	Repeated-measures ANOVA	1,313 8th-grade students, nationally representative	No significant mean score differences between paper-based and computer-based modes. Computer familiarity significantly related to online writing test performance after controlling for paper writing skill. Subpopulation analysis indicated a significant interaction effect of delivery mode with school location (specifically, students from urban/large town locations performed significantly higher on paper as compared with computer).

Table B4 (continued). Eleven studies that evaluated effects of demographic characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Horne	2007	Lucid Assessment System for Schools (LASS) Secondary and LASS Junior (Language Arts, Spelling)	*t*-tests	242 students, ages 9–15	In the paper-based test, females scored significantly higher than males on the reading and spelling tests. In the computer-based test, there was no significant difference by gender.
MacCann	2006	Computing skills test	Regression analyses; repeated-measures ANOVA	14,248 volunteer students ages 15–16 (New South Wales, Australia)	There was no significant interaction between gender and mode of administration. There was a significant score difference by mode found for socioeconomic status (SES), where low-SES students performed better on the paper-based mode than the computer-based mode. There was no significant interaction between item format and mode of administration.
Pomplun, Ritchie, Custer	2006	Initial-Skills Analysis (part of the Basic Early Assessment of Reading)	Single-group counterbalanced design; omit rates by mode; regression analyses	2,000 students in Grades K–3, (23% free/ reduced-price lunch eligible, 78% white)	Mean scores were significantly higher for the paper-based test compared with the computer-based test for all grades, with effect sizes ranging from .27 to .48. More items were omitted in the paper form than the computer form, though the difference was significant for only two of the four grades. Deferring, omitting items, and free/reduced-price lunch status were significant predictors of computer-based test scores after controlling for paper-based test scores.
Russell, Plati	2000b	MCAS Language Arts	Independent *t*-tests; regression analyses	Students in Grades 4 (152), 8 (228), 10 (145)	Scores were significantly higher for computer-based test scores than paper-based test scores. At Grades 8 and 10, special education students had significantly higher midterm grades when performing composition items on paper. There was no significant difference for special education students in Grade 4 by mode.

Table B5. Five studies that evaluated effects of computer familiarity, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Chen, White, McCloskey, Soroui, Chun	2011	Functional Writing, items from 2008 National Assessment of Adult Literacy (NAAL)	Between-subjects; within-subjects; ANOVA and repeated *t*-tests	1,607 subjects, ages 16+	Scoring bias analysis: When handwritten essays were transcribed, there were no statistically or practically significant scoring differences between handwritten and transcribed computer responses to the three writing tasks. Regarding the effects of administration mode, the analyses showed a consistent advantage for the paper mode over computer mode for the overall tasks scores and individual scoring criteria. For the length of writing, there was no significant difference. Some significant effects were found in individual tasks by race/ethnicity, age, education, word-processor experiences, and employment status. None of these showed consistent effects across all three tasks.
Goldberg, Pedulla	2002	Practice GRE	Multivariate analysis of covariance (MANCOVA)	222 3rd- and 4th-year university students (28% male)	Positive main effect of computer familiarity on Analytical and Quantitative subtests (not on Verbal). Performance differences were statistically significant among test modes on each of the subtests: Analytical Verbal and Quantitative. There was a statistically significant interaction effect between test mode and computer familiarity on the Quantitative subtest performance.

Table B5 (continued). Five studies that evaluated effects of computer familiarity, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Higgins, Russell, Hoffmann	2005	Writing items from NAEP, Progress in International Reading Literacy Study (PIRLS), and New Hampshire State Assessment	Computer fluidity test, computer literacy test, computer use survey	219 participants, 4th grade	No differences in reading comprehension across testing modes (paper-based test, computer-based test with scrolling, computer-based test whole page); No statistically significant differences in reading comprehension based on computer fluidity (use of mouse and keyboard) and computer literacy; Computer anxiety levels did not significantly affect scores.
Russell	1999	MCAS, NAEP open-ended items in Language Arts, Science, and Mathematics	Independent *t*-tests; multiple regression	229 middle school students	The study found that computer-based testing led to higher scores in Science and lower scores in Mathematics subtests. In the English and Language Arts subtest, there was no overall effect, but there was a significant effect found by keyboarding skills.
Russell, Plati	2002	Writing items from MCAS	Independent *t*-tests; regression analyses	Grades 4, 8	Keyboarding skills were positively correlated with test scores in 4th grade; however, there appears to be a threshold above which keyboarding skills have no significant effect.

Table B6. Ten other studies that found score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Escudier, Newton, Cox, Reynolds, Odell	2011	Undergraduate dental school course assessments; attitude survey	Repeated-measures ANOVA; focus-group discussions	132 year 3 and 134 year 5 dental undergraduates	For year 3 students, there was a significant interaction between test order (whether the paper-based test or computer-based test was administered first) and performance. For year 5 students, computerized scores were higher than paper test scores regardless of the test order. The attitude survey revealed that participants felt the online test format did not disadvantage students, even in a high-stakes situation.
Fulcher	1999	English-Language Placement Test, 80 items: all multiple choice	Within-subjects; ANCOVA	57 university students	Computer-based test scores were higher than paper-based test scores. There is a possible practice/order effect because students took paper-based test first.
Kingston	2009	K–12 Assessments in Mathematics, Reading, English Language Arts, Social Studies, and Science	Meta-Analysis	K–12	The study found that computer-based assessment led to higher scores for English language arts and social studies, but lower scores for mathematics. No significant effect by grade level was found.

Table B6 (continued). Ten other studies that found score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Liao, Kuo	2011	Four Assessments on Chinese Language Ability: One-Minute Word Reading; Onset Detection; Rhyme Detection; Rapid Automatized Naming (RAN) (e.g., reading fluency). Paper-based assessment: In-person read-aloud of audio tasks. Computer-based assessment: Computer-delivered audio.	Hierarchical multiple regression	93 students, Grade 6	Results showed that the two modes for RAN are highly correlated, but not for Rhyme detection and onset detection. The results showed that conventional and Web-based versions were equally predictive of Chinese reading measures.
Pommerich	2002	Fixed-form tests in English, Mathematics, Reading, Science Reasoning	Two different computer interfaces were used; *t*-tests	Large scale (about 20,000)	Levels of comparability were inconsistent. A variety of factors appeared to be related to mode effects. Changes to computer interface seemed to have significant effect on cross-mode differences.
Pommerich	2004	English, Reading, and Science Reasoning assessments	Two different computer interfaces were used; *t*-tests	12,000 students from 61 schools in Grades 11, 12	Results varied by computer interface condition and subject area.
Pommerich, Burden	2000	20-minute content area tests in English, Mathematics, Reading, Science	Within-subjects, nonrandom assignment; *t*-test	36 students, Grades 11, 12	Assessment factors that were found to be the most likely to lead to construct-irrelevant effects were pages and line length, layout features, highlighting, and item characteristics.
Wang, Jiao, Young, Brooks, Olson	2007	Various mathematics assessments	Meta-analysis of mean score differences by mode (14 studies with 44 independent effects)	K–12	Meta-analysis found that overall there were few small differences between modes, with effect sizes ranging from -.28 to .08. There was a significant difference in the effect size by delivery algorithm (linear versus adaptive computer-based assessments). The paper-based test had larger variance than the computer-based test. No differences were found by grade or computer practice.

Table B6 (continued). Ten other studies that found score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Wolfe, Manalo	2004	TOEFL Writing	Generalized linear model (GLM)	133,906 English language learners ranging from 15 to 55	The paper-based test had higher essay scores than the computer-based test but mode explained only a small amount of variation (r^2=.01). Participants with lower English language ability scored slightly better on paper (interaction). Participants with higher English language ability scored slightly better on computer (interaction).

Table B7. Ten other studies that found no score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Anakwe	2008	University Accounting course assessments (3 courses)	*t*-tests	54 university students	No statistically significant score differences across modes in any of the three courses.
Balizet, Treder, Parshall	1999	Study-specific tests of Academic Listening Comprehension and Vocabulary; PPT: Audio-cassette, Computer-based test: Computer-delivered audio	*t*-tests; descriptive statistics	28 high-intermediate level English as a second language students	No significant score difference between the two administration modes.
Bodmann, Robinson	2004	Undergraduate Educational Psychology Course Assessments	Dependent *t*-tests	113 undergraduate students in an educational psychology class	Computer-based assessments were completed faster than paper-based assessments with no significant differences in scores.
Coniam	2009	2007 Hong Kong Certificate of Education Examination (HKCEE) Year 11 English Language Writing Paper (Hong Kong Public Exam)	Scoring modes compared: "Onscreen Marking" and "Paper-Based Marking" scoring methods; metric: inter-rater reliability (IRR); chi-square tests; *t*-tests	30 raters (scorers) in Hong Kong	Scores awarded by "Onscreen Marking" and "Paper-Based Marking" were comparable.
Higgins, Patterson, Bozman, Katz	2010	25 General Educational Development (GED) mathematics practice items	Regression analyses	216 participants	There was no significant difference between paper-based test scores and computer-based test scores after controlling for initial performance.
Mason, Patry, Bernstein	2001	Introductory psychology course assessments	One-way ANOVA	27 university students (mean age: 20.2)	There were no significant differences by mode.
Minnick	2009	Tests of Adult Basic Education (TABE)	*t*-tests	150 male prison inmates ages 14–18	There were no significant differences by mode.
Mogey, Paterson, Burk, Purcell	2010	Essay test, mock course exam	Responses were transcribed so that each response was scored in both modes; ANCOVA	70 first-year divinity school students (nonrandom: participants chose condition)	No significant differences (including length of essay, overall scores, and some qualitative measures designed to indicate essay quality) found by mode.

Table B7 (continued). Ten other studies that found no score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Whiting, Kline	2006	Test of Workplace Essential Skills (TOWES), Test of adult literacy skills, Subscales: Reading test, Document skills, Numeracy	Computer-based test scores and archived paper-based test scores matched based on years of education, age, gender; rank order equivalency; *t*-tests	73 undergraduate university students	Scores on all three subscales were equivalent based on their means and variances. In posttest survey, participants rated the computer-based test as easy to use.
Zandvliet, Farragher	1997	Three tests adapted from instructors' guide in an introductory college-level computer course.	*t*-tests	50 students in introductory computer classes	No significant mode effect on assessment scores was found.