A Framework for Considering Device and Interface Features That May Affect Student Performance on the National Assessment of Educational Progress

Denny Way
College Board

Ellen Strain-Seymour
Pearson

March 2021
Commissioned by the NAEP Validity Studies (NVS) Panel

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Keena Arbuthnot
Louisiana State University

Peter Behuniak
Criterion Consulting, LLC

Jack Buckley
American Institutes for Research

James R. Chromy
Research Triangle Institute (retired)

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

James Pellegrino
University of Illinois at Chicago

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
College Board

Project Director:

Sami Kitmitto
American Institutes for Research

Project Officer:

Grady Wilburn
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS) Panel
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Email: skitmitto@air.org

CONTENTS

INTRODUCTION

An ongoing high priority for the National Assessment of Educational Progress (NAEP) is the ability to compare performance results from 1 year to another. A recent challenge to maintaining NAEP trends has arisen with the exploration of new testing methods and question types that reflect the growing use of technology in education. NAEP has introduced a variety of new question and task types in the assessments, including writing on computer tasks, interactive computer tasks, hybrid hands-on tasks, and scenario-based tasks from the Technology and Engineering Literacy (TEL) and reading assessments. To maintain trend in moving from paper-based to digital assessments, NAEP has utilized a multistep process involving bridge studies to establish links between administration modes and a gradual introduction of more innovative questions or tasks that make use of digital technologies.

The strategy of using bridge studies to link across paper and digitally based NAEP assessments has been relatively successful, but recent experiences with the NAEP digital writing assessment has revealed a further comparability challenge in the form of evidence that the device used in digital assessment (e.g., a laptop versus specific alternate digital devices) also may introduce unintended performance differences that threaten the validity of NAEP trends (National Center for Education Statistics, 2019). These recent findings are especially concerning in light of the practical necessity that the devices and interfaces used to deliver NAEP assessments will continue to evolve over time.

The NAEP Validity Studies (NVS) Panel has outlined potential studies that might be added to its research agenda to explore device effects and the impact of device familiarity on performance, including randomized trials of alternate devices, teacher surveys, expert panels, and cognitive labs. In addition, consideration is being given to how to maintain trend in the face of constantly changing technology, with options that might range from continual bridge studies each time the delivery device or administration interface changes to a reconceptualization of what "standardization" and "trend" mean. However, these various potential studies or policy actions are difficult to prioritize because there is no organizing framework within which to evaluate them. What is needed is an elucidation of significant causal variables to guide the studies.

The purpose of this white paper is to provide a framework for considering device and interface features that may affect student performance on digital NAEP assessments, and to prioritize the variables that should be examined in further validity studies. In building the framework, we propose to use writing as an entry point for several reasons. First, student engagement with a device is more intense with writing than in most other areas assessed in NAEP. In addition, motivation can affect how students engage with NAEP assessments, and the potential impact of motivation on writing performance is arguably larger than for other assessments because of the nature of the production task and the limited number of resulting data points. We hypothesize, therefore, that potential device/interface effects will be magnified in the context of writing. In addition, writing has some unique attributes. For example, experts have acknowledged that the constructs of digital writing and paper-based writing are not the same. Finally, there has been evidence of device/interface effects in recent NAEP digital writing assessments.

Although we focus initially on writing, we believe that aspects of device/interface effects possible within writing assessments will generalize to other NAEP subject areas. For example, the extent of device engagement expands with simulations and more complex item types, thereby making issues such as device familiarity, the precision of input devices, and motivation relevant to subjects beyond writing. As another example, screen real estate and focus shifting are issues within writing, as students need to access the prompt, gauge how much they have written as they respond, and refer back to previous paragraphs. However, these issues are not exclusive to writing. Reading is aided by backtracking, referencing earlier material, and comparing or juxtaposing different parts of a passage. Math may involve shifting focus between a calculator, a prompt, and answer choices. Social studies may involve maps and tables that occupy significant screen real estate. Thus, how much a test-taker can see at one time and how easily they can get to what is not currently visible has relevance across subjects. Finally, to the extent that tasks that involve productive writing are incorporated into other subjects (e.g., reading and social studies), device engagement impacts similar to those for writing may occur.

The structure of this paper is as follows. First, we summarize relevant research most directly related to score comparability across devices and modes of assessment. In the process of covering device/interface effects research done to date, we discuss the difficulty in relying entirely upon these findings to guide either device decisions or future research. We point to the inherent complexity of isolating variables in the face of the interplay between individual test-taker characteristics, the affordances of the device and peripherals used by the test-taker, the interactivity of the test delivery system, and the cognitive demands of a particular task. Nonetheless, within this section, special attention is given to particular studies that suggest fruitful directions in methodology.

Having established the need for organizing principles and appropriate methodologies for guiding future research in this area, next we introduce the foundation for our proposed framework. In this section, we present the initial contours of an analytical method for investigating assessment tasks, the cognitive processes involved, and the role played by the digital assessment environment, borrowing from a number of theoretical frameworks, and ultimately coalescing around an interdisciplinary definition of task modeling. In this section, we do not attempt to focus analysis on any singular task type. However, in the third major section of the paper, we begin to flesh out our methodology, beginning with an exploration of the elements of writing as a task with high cognitive and motor demands. We end this section with guidance on some of the elements of devices, peripherals, interfaces, and overall assessment design that may impact writing performance. In the fourth section, we address nonwriting subject areas, highlighting additional considerations for devices and interfaces that arise with digital assessment in science and mathematics in particular.

In the fifth section of the paper, we propose our framework for considering device and interface features that may affect student performance on NAEP. First, we identify the most salient device and interface variables and address each as an isolated feature (while still acknowledging the complex interactions among them). We classify the variables into one of three major categories (screens and input devices, test delivery system tools, and interface elements). For each variable, we provide summary information regarding considerations such as the range or measurability of the variable, research findings, potential impact, and predictions about the how considerations related to the variables will change in the future.

As part of this section, we apply the framework to NAEP assessments and point to implications as related to tradeoffs between usability and device familiarity.

In the final section of the paper, we discuss potential research focusing on device/interface familiarity given our understanding of NAEP's plans for evolving their digital delivery system and transitioning to new devices. Our discussion covers a range of options ranging from the development of additional questionnaire items to one or more complex bridge studies.

COMPARABILITY LITERATURE

Although some elements of NAEP's research agenda are shaped by unique characteristics of the NAEP assessments, many of the questions faced by NAEP in light of changing technology and new question types parallel those faced by other standardized testing programs, as reflected in the literature on comparability of scores across different testing conditions, modes, and devices.

Mode/Device Studies

For NAEP and other programs, the evolutionary use of technology has involved a series of transitional steps. The first steps involved the shift from paper to digital administration, first with continuity in terms of question types and next with differences in question types. The research on the comparability of digital versus paper testing in the context of this move identified some trends but also some ambiguity. An example of this ambiguity can be seen by contrasting a recent study identifying significant mode effects favoring paper performance on the Partnership for Assessment of Readiness for College and Careers (PARCC) assessments as they were rolled out in Massachusetts in 2016 and 2017 (Backes & Cowen, 2019) with recent comparability research on the ACT and SAT tests, which both indicated mode effects favoring *digital* performance (Li et al., 2016; Proctor et al., 2020). For NAEP, similar ambiguity can be seen in the seminal writing online study (Horkay et al., 2006) where results generally showed no significant mean score differences between paper and computer delivery, but also indicated that computer familiarity significantly predicted online writing test performance after controlling for paper writing skill. Furthermore, in a portion of the Horkay et al. study examining performance on school computers versus NAEP-supplied laptops, a significant interaction of gender with computer type was found, indicating that the difference in performance between computer types was not the same for male and female students.

Paper-based standardized testing has been in use for generations, and the procedures used with paper testing are well recognized and relatively stable. Some might argue, however, that recent shifts toward digital learning, instruction, and testing could render paper testing less familiar. Digital testing approaches continue to evolve across time, whereas paper testing is pretty much the same as it was 20 or 30 years ago. If one accepts the general notion that familiarity with paper testing is ubiquitous, conflicting results in paper versus computer mode comparability studies would seem to be related to factors associated with computer administration, such as the testing situations (e.g., purpose and use of tests), the online testing interfaces, and/or factors related to the test-takers themselves (e.g., motivation, familiarity with the computer or device). That the hundreds of paper versus online mode comparability studies over several decades have yielded few if any stable conclusions over time does not lend confidence that a new generation of research on comparability of test performance across different devices will prove more insightful.

This point is relevant in considering the role of the "bridge studies," which have been and remain the essential mechanism by which NAEP maintains comparable trends in assessment results across changes in constructs, measurement, sample, and/or administration conditions (National Center for Education Statistics, 2003). Bridge studies have been used to preserve trend when NAEP subjects transitioned from paper to computer administration (Jewsbury et al., 2020). Similar approaches have been used in other comparative assessment programs (e.g., TIMSS [Trends in International Mathematics and Science Study]; see Fishbein, 2018). The expectation is that performance differences revealed by bridge studies are generalizable and stable. However, if relationships between performance on computer versus paper, laptop versus device, or from one interface to another are not stable over time, the statistical adjustments applied in bridge studies may not serve their intended purposes. This becomes a heightened challenge for the second transitional step in the evolutionary use of technology for NAEP assessments, where inevitably the devices and interfaces used to administer NAEP will be changing over time.

The literature on device effects and comparability in the context of large-scale assessments is limited and mixed. Most studies do not come from peer-reviewed journal publications; a recent summary by DePascale and colleagues (2018) referred to "gray research," which the authors compiled from internal reports, technical papers, and project reports prepared by various organizations. Although such studies generally compare performance for students testing on tablet-like devices versus more traditional desktop or laptop computers, they are quite varied with respect to the assessments, devices, interfaces, subjects, and item types studied. Nonetheless, some of these studies may have implications for NAEP device considerations.

One published study relevant to writing by Davis and colleagues (2015) involved 826 students from Virginia and South Dakota at two grade levels who were randomly assigned to respond to a grade-level-appropriate essay prompt using a laptop, tablet, or tablet with an external keyboard. Results indicated no difference in the essay score or surface-level essay features across study conditions, although results were limited by possible motivation effects (as evident in relatively short essay lengths) and the use of relatively simple essay prompts (e.g., no reference materials were required to respond).

More generally, Davis and colleagues (2016) examined the comparability of test scores across tablets and computers for high school students in three commonly assessed content areas and for a variety of different item types. Results indicated no statistically significant differences across device type for any content area or item type. However, again, the study was conducted with no stakes involved for the participants and could have been limited by possible motivation effects.

An example of a device study with results that are difficult to interpret was based on data from the first operational PARCC administration (Steedle et al., 2016). This study compared performance on tablets and nontablet devices (i.e., desktop and laptop computers). Overall, the authors concluded that the study revealed "consistent evidence of comparability between testing on tablets and non-tablet devices" (p. 4). However, the methods section also indicated that data from one state were excluded from all analyses "because of highly atypical differences between the performance of students who tested on tablets and nontablet devices. When analyses included data from this state, extensive evidence of device effects was observed on nearly every assessment" (p. 9). Although inferences from this outcome are not clear, there is at least some suggestion that local administration factors can influence device effects.

Cognitive Laboratories

Because of the difficulties associated with large-scale research studies involving experimental conditions and adequate sample sizes, a sizable amount of research from the assessment community on mode and device comparability has been conducted using cognitive labs or usability studies (Davis et al., 2013; Pisacreta, 2013; Strain-Seymour et al., 2013; Yu et al., 2014). This more intimate look at how test-takers interact with assessment tasks sometimes followed an analytical trajectory of mode and device studies. Item-level analyses were used to determine if mode or device effects impacted item performance uniformly or if some items fared less well in the transition from paper to online administration or from computer to tablet. With some items earmarked as being more sensitive to mode or device effects than others, the next question was why. Could common item characteristics be identified? Theories circulated on whether items that scrolled, items that students tended to mark up on paper, and/or those that required consulting multiple data sources or using digital tools were more vulnerable to mode effects.

Cognitive labs have provided a method for testing out theories concerning mode-effect causality, discovering usability issues, and gathering thoughts regarding what content should be put into tutorials to better prepare test-takers for interacting with a test-delivery system. Cognitive labs also have been used to study possible implications of delivering tests on tablets. Many states, whether acting individually or as part of a consortium, looked to new devices, such as iPads and Chromebooks, as the solution for moving to online-only assessments that could include technology-enhanced items (TEIs), digitally enabled accommodations, or computer-adaptive formats. Such devices were becoming more prevalent in schools as a result of one-to-one device initiatives and state/federal educational technology funding (Blume, 2013). Making such devices a key component of assessment also was seen as another justification for device purchases, which would ideally enhance instruction throughout the school year. To clear these devices for use in high-stakes assessment, a number of topics were of interest within these cognitive labs: the use of touchscreens with TEIs; the impact of smaller screen sizes; familiarity with tablet conventions, such as finger-based zoom operations; and writing with full, compact, and virtual keyboards (Davis et al., 2013; Pisacreta, 2013; Strain-Seymour et al., 2013; Yu et al., 2014). Although some of the findings from these cognitive labs circulated in the form of gray research, other conclusions may have been highly specific to individual vendor's systems or item design processes, and less likely to be discussed at conferences and in white papers.

NAEP's extensive cognitive lab study of science scenario-based tasks (SBTs) by Duran and colleagues (2019) stands out among these studies for its methodology and intent to guide design considerations for construct-relevant use of visual and interactive features. Importantly, the study was informed by Mayer's (2009) principles for the use of multimedia in instructional tasks as well as consideration of the human computer interface research base (Watzman & Re, 2012). Results indicated that study participants were favorably disposed to the SBTs, and tended to comprehend the visual and interactive features of the tasks, but needed more time to respond than had been assumed in earlier pilots of the tasks. In addition, some features of the science SBTs were found to be problematic, and a range of design recommendations for future development of science SBTs and similarly complex multimedia tasks were offered. An important suggestion from the study was for integrating better articulated developmental guidelines for the visual and interactive features of the SBTs within the critical path of task development, along with appropriate quality control procedures. In general, the detail and depth of the research study underscored the complexity involved in the development of the SBTs and the challenge of being able to understand and predict a priori how students will interact with various task features.

The Need for a Framework

A number of conclusions can be drawn from this review of the literature. The first such conclusion is questioning the terminology "device effects" with its implied narrow focus on the tangible device, whether desktop, laptop, tablet, or some indeterminate blend of tablet and laptop. Certainly, the physical elements of devices and peripherals—such as screen size, screen reflection, input device, and keyboard size—were of interest within some of the studies described here and elsewhere in this paper. For instance, various cognitive labs and observational studies of tablet use with assessment materials drilled down to behavioral factors apparent in test-takers' physical interaction with a device: student choices regarding stylus usage, touchscreen usage, test-taker position in relation to tablets placed on a stand, styles of keyboard usage, and student reaction to light reflection on angled tablets (Davis et al., 2013; Strain-Seymour et al., 2013). In these cases, any performance impact—whether statistically or anecdotally indicated, or whether confirmed, suspected, or debated—could not be definitively tied to a device in isolation. Aspects of the device interacted directly with item type, content presentation, and the interface of the test delivery system, and thereby introduced difficulty or distraction not witnessed on another device or in a printed booklet. In other cases, if some disadvantage of a device were noted in a particular context, changes in tool design, interface, or content presentation offered potential solutions. Some of the examples include the following:

- Online tools, such as highlighters, answer eliminators, and other digital marking tools, to support thought processes similar to pencil mark-up of print booklets to decrease paper/online mode effects (Russell, 2018)
- The use of a paging interface to support relocating of text using spatial memory within passages in digital assessments (Higgins et al., 2005; Pommerich, 2004)
- Content guidelines of a minimum size of 44 by 44 pixels for draggable objects in interactive items delivered on touchscreens to avoid any negative impact stemming from occlusion from one's finger (Davis & Strain-Seymour, 2013a)
- A "halo" around a point on an interactive line, point, or function graph so that it can be positioned precisely despite being necessarily small in size (Davis & Strain-Seymour, 2013a)
- Interfaces for writing that provide insight into the amount of text written as well as remaining space in the case of word or line limitations (Way et al., 2008)
- Usability issues on the computer, which can be exacerbated on the tablet, particularly when partial solutions involve extending user feedback through rollover effects and cursor change-outs (Davis & Strain-Seymour, 2013a)
- Resizing and repositioning mechanisms for images, text, calculators, passages, formula charts, and other supplementary materials or tools to provide flexibility for smaller screens

Thus, we must acknowledge that our use of the term "device effects" is best understood as shorthand for the larger interplay of a number of variables. Elsewhere in this paper, we either rely on this shorthand or use the term "device/interface effects" to acknowledge the interaction of digital and physical factors alongside content design factors.

By recognizing that the results of the above-described comparability studies were shaped by a wide range of factors that extend beyond the question of tablet versus laptop, for instance, we also recognize the extreme specificity of many findings. This may, in part, explain some of the conflicting results of device studies, with each study working with different devices, item interactions, interfaces, test-taker populations, and motivation levels. Another factor inhibiting generalizability is the potential for findings to quickly become dated in light of rapid changes in technology (e.g., device models, screen/monitor clarity, operating systems, input devices, interactivity conventions, keyboard technologies). In addition, the adoption of technology in the classroom and student comfort with and mastery of those classroom technologies (not to mention personal technology use outside of the classroom) arguably increase with each passing year. This flux could be assumed to shorten the shelf life of any finding's generalizability to tomorrow's tests and test-takers. For instance, studies suggesting the utility of paper metaphors, such as paging interfaces for reading passages (Higgins et al., 2005; Pommerich, 2004) and page thumbnail views to provide a paper-like view into essay length (Way et al., 2008), assumed an existing classroom norm based on the ubiquity of paper for classroom reading and writing. Such an assumption may not hold true 15 years after those studies were conducted. Even reflections on the importance of device familiarity may not hold true as it becomes more common for students to use a variety of devices in the classroom and at home. Thus, a second unfortunate conclusion that follows from this review of the literature is that, collectively, this research does not provide concrete and extensive guidance for device/interface decisions moving forward.

Ambiguity in the research literature regarding the comparability of test performance between computers and devices or even across devices invites a widening of methodologies used to pursue comparability questions. One could argue that the assessment industry's comfort level with the various statistical methods for defining item/test characteristics and evaluating reliability is well established, unsurprisingly, in comparison to methodologies that were not homegrown for assessment purposes but instead crossed borders into assessment from adjacent disciplines. This leaves us with a two-part goal. First, in the absence of a well-researched and stable body of actionable knowledge to guide both assessment design/policy and future research in the area of device/interface effects, we intend to sketch out a framework that provides an initial step in this direction. By describing the range of variables to account for when investigating device/interface effects, summarizing research around those variables, and pointing to fruitful methodologies, we intend to provide a jumping-off point for further research. Second, we aim to continue the interdisciplinary trend of device/interface effects research to date by interweaving a range of approaches. It would be overly ambitious to suggest that we offer a full integration of disciplinary perspectives into a unified model for understanding the interaction of a test-taker with an assessment task delivered within a particular interface and on a specific device. Instead, our model is better described as a series of shifting perspectives to provide a multifaceted view to accompany more traditional statistical analyses of assessment results. We begin by describing the complementarity of several theoretical approaches in an attempt to explain analytical methods that can be used to explore both writing and nonwriting assessment tasks in the sections that follow.

INTERDISCIPLINARY FOUNDATIONS

When we consider how a test-taker engages with an item in a digital assessment, whether a simple arithmetic calculation or the creation of an extended essay, a process for understanding this engagement (and the role of the interface and the device) can be described as task modeling. We argue for a particular definition of task modeling found at the intersection of (1) an educational measurement's emphasis on construct validity situated within a sociocognitive framework as suggested by Robert Mislevy (2018); (2) methodologies common to human-computer interaction (HCI) research; and (3) cognitive load theory (CLT) as applied to instructional design. To the degree that all three of these theoretical constructs draw upon precepts from cognitive science, they align well with one another while bringing a larger arsenal of analytical tools necessary to disentangle the complex interactions between a device, the software interface of the test delivery system, a task's engagement of certain cognitive processes, and the interpretive value of the task response as evidence of knowledge or process mastery.

Cognitive Load Theory and Human-Computer Interaction

In exploring this three-way intersection of disciplines, we begin with the principles of cognitive load and the cross influences between HCI and CLT before tying these theoretical strands back to educational measurement and the realm of digital assessment. CLT dates back to the 1950s and has continued to expand its model of memory and cognitive processing, building on the core tenet that working memory has an inherently limited capacity when dealing with novel information (Cowan, 2010). Working memory is understood as a space for perceiving and manipulating that novel information and as a gatekeeper to long-term memory (Ericsson & Kintsch, 1995). Learning is then defined as the movement of information from working memory to long-term memory, which is not subject to the same capacity restrictions as working memory (Sweller, 2010).

Hollender and colleagues (2010) argue that, with the increased prominence of e-learning environments within K–12 and higher education, CLT has been heavily leveraged within HCI research directed at optimizing the effectiveness and usability of digital learning tools. Based on an analysis of the literature, Hollender et al. (2010) suggest that the concept from CLT that has had the most impact on HCI research is extraneous cognitive load: cognitive processing incurred by distracting or ineffective presentation/interactivity rather than by the content itself (Sweller, 2010). Three different types of cognitive load—intrinsic, extraneous, and germane—are conceived to be additive, with the potential for the total load to exceed the maximum capacity of working memory. For instance, instruction is ineffective when the intrinsic cognitive load—the cognitive demands inherent to the learning task itself—is too great for a learner who cannot borrow from existing knowledge within long-term memory to work through a highly complex task. Another learner faced with the same task but with greater mastery of a subject area may need to mobilize less working memory as they connect this novel information to existing knowledge schemata held in long-term memory.

When extraneous cognitive load taxes working memory due to how a learning task is delivered, a reduced portion of a learner's working memory is available for active engagement with a task's intrinsic cognitive load. To provide examples, extraneous load could come from an unlabeled graphic that must be synthesized with prose for adequate understanding or from an unfamiliar way of representing an equation. Within an e-learning or assessment environment, extraneous cognitive load could be introduced when the student grapples with an interface that fails to meet the usability standards of learnability and efficiency (Nielsen, 1993). HCI practices can be applied to identify and minimize sources of extraneous cognitive load by improving the software's overall usability.

HCI research traditionally engages with a number of types of tasks and a wide range of users, bringing ergonomics and human factors to bear on the problem of interface design. Similarly, CLT, as a principle for learning efficiency, may be applied to a wide range of instructional environments, such as classroom instruction, project-based learning, textbooks, workbooks, simulations, game-based learning, augmented reality, and so on. The disciplines intertwine when the classroom task, whether instructional or assessment-based, is delivered digitally, such as on a computer or tablet. Although HCI has leveraged CLT's theoretical precepts, CLT studies have expanded into new areas by borrowing and expanding upon empirical methods traditional to HCI. For instance, eye-tracking technology, introduced more than a century ago to analyze reading, became a mainstay of usability studies starting in the 1980s. In some ways, however, eye-tracking technology came full circle within CLT when married with pupillary response measurement as a way to quantify cognitive load and provide evidence of higher-order cognitive processes during reading and other cognitive tasks. Although eye-tracking synchronized with screen-activity capture could indicate which elements drew in the user's gaze (also used in marketing research) and how the user moved through the visual/textual material, further evidence of unseen cognitive activity was sought. The use of pupillometric data filled that gap. Studies of computer-based tasks involving memorization, mathematical calculation, listening comprehension, reading, and translation suggested that increased pupil dilation was a reliable indicator of higher cognitive demand (Kruger et al., 2013). More recent studies have focused on microsaccade magnitude as an indication of cognitive load: decreased control of the magnitude of these miniature, involuntary eye movements while focused on a singular location and performing nonvisual cognitive tasks was highly correlated with other measures of cognitive load (Krejtz et al., 2018).

Although many of these research studies have been designed to compare cognitive load in specific conditions, such as comprehension of a video with and without subtitles, these studies collectively have produced a clearer understanding of certain tasks in the abstract, which then provide clues as to how to design tools and interfaces to support such tasks in ways the minimize cognitive load. For instance, eye-tracking studies of reading have produced a general theory of the mechanics of reading. Small, rapid eye movements (saccades) involve jumping forward about seven to nine characters interspersed with pauses (fixations) of 20 to 40 microseconds. Longer pauses can be indicative of higher processing loads, such as when inference making is required for comprehension (Just & Carpenter, 1980). Backward eye movements, known as regressive saccades, are common—whether traveling small distances (potentially due to oculomotor error) or larger distances as a result of comprehension issues (Inhoff et al., 2019; Eskenazi & Folk, 2017). Such a theory of reading then provides a foundation for pursuing questions of differences in spatial clues derived from paging versus scrolling text, the impact of how much text is visible at one time (such as with variable screen size), and the utility of certain tools for aiding on-screen reading.

The Applicability of CLT to Assessment

In transitioning to concerns more specific to assessment, we might inquire about the degree to which CLT and HCI research findings can be considered generalizable to testing environments. It is true that one goal pursued within this rich area of intersection between HCI and CLT has been the development of instructional design best practices. This focus on learning efficiency is obviously distinct from the measurement goals of assessment. Nonetheless, the processes described by CLT pertain to cognitive tasks generally speaking and have been used to explore reading, writing, problem solving, deductive reasoning, and mathematical calculations— many of the same essential activities performed by students in assessment environments.

At the heart of both assessment and learning is a performative role by a student working with content that combines the new and the known. With active learning and the role of rehearsal within instruction, students may be asked to use existing knowledge to solidify their understanding of patterns by applying a known approach across numerous variations, such as mathematical problem sets. A test-taker may confront novel information in the form of a reading passage or a simulation and be asked to illustrate their ability to use existing knowledge and process skills to produce meaning from this content. In either case, the shared goal of assessment and instruction design is to optimize the conditions for the student's performance of prior knowledge while encountering new information or novel variations.

The concepts of element interactivity and germane cognitive load, the third type of cognitive load introduced but unexplored above, are relevant to further exploration of prior knowledge's role in both assessment and instruction. When discussing element interactivity, an element is assumed to be at the most granular level of learning or assessment; an element could be a fact, process, or skill. Low element interactivity exists when the learning of an element or task completion using this element can be easily isolated from other elements. For instance, in foreign language instruction, a vocabulary word with a singular meaning and direct translatability between English and the language being taught may have low element interactivity in comparison to other concepts such as grammar or verb conjugation. Within instructional design, recognizing tasks with high intrinsic cognitive load due to extensive element interactivity is of value, particularly if the cognitive load can be lowered by teaching those elements successively rather than simultaneously (Sweller, 2010). A similar observation holds true for assessment: When designing an assessment task with high element interactivity, greater measurement precision can be achieved if the design includes methods to evaluate mastery of the component elements and not just the totality.

At a higher level than elements are schemata: cognitive frameworks that help to organize existing knowledge and interpret novel information (Kalyuga, 2010). CLT, drawing on schema theory, explores how lower-order schemata are integrated with higher-order schemata within the learning process. Germane cognitive load, the third and final type of cognitive load assumed to contribute to the total load of a task, occurs with the mental activity required to connect an element to other learned elements and, more broadly, to existing schemata, thereby solidifying or expanding the learner's schemata (Debue & van de Leemput, 2014). To return to the example of foreign language instruction, the overall cognitive load involved in vocabulary acquisition may increase with the involvement of germane cognitive load in the case that a new term is integrated into learner schemata related to word roots, language-specific syllable structures and stress patterns, and perhaps even comparative linguistics concepts. The development, exercise, and automation of these schemata is a more robust and desirable form of learning than disparate fact accumulation. Parallels also can be found in assessment. Although evidence of success with tasks involving low element interactivity is certainly pursued, measurement of a test-taker's ability to manipulate new data using higher-order cognitive skills and previously acquired information-processing knowledge structures is highly valued.

Returning to the idea of optimizing task performance as a shared goal between instruction and assessment, one optimization strategy involves reducing extraneous cognitive load. When the tools used for a task are not second nature, and when the student's proficiency with them does not approach automaticity, some extraneous cognitive load can be anticipated. Within assessment, the variability of extraneous cognitive load is worth noting. For instance, if device, tool, or interface unfamiliarity introduces extraneous cognitive load, and if familiarity varies across test-takers, then score differences cannot be entirely attributable to construct knowledge. Messick (1989) references this type of systemic threat to assessment validity, regardless of its source, as construct-irrelevant variance.

Another strategy for optimizing task performance involves the appropriate triggering of prior knowledge, whether at the element or schema level. Pursuing this idea rather simplistically, we might point to triggering mechanisms within an instructional environment: a teacher harkens back to yesterday's lesson, a textbook references the prior chapter, or a hyperlink defines a previously introduced term. However, when the goal is to measure—not impart—this knowledge, and when a standardized test is a bracketed experience that stands apart from the social fabric of a student's day-to-day learning environment, methods for eliciting an appropriate performance of prior knowledge within a task response may not be simple nor well understood.

Resource Activation and Construct Validity

Robert Mislevy in his 2018 work, *Sociocognitive Foundations of Education Measurement*, explores learning experiences and assessment instruments as situated within a sociocognitive framework and investigates the impact of this situatedness on the elicitation of prior knowledge. He begins with the nature of knowledge, referencing information structures and practices in a somewhat traditional fashion as he touches upon prior research in the areas of second-language acquisition, scientific models for thinking about motion, procedures for subtracting mixed numbers, and strategies for troubleshooting computer networks. Ultimately, however, he opts to use the terms "patterns" and "resources" as opposed to relying on a decontextualized notion of stored knowledge. Mislevy argues that individuals become attuned to linguistic, cultural, and substantive patterns, or LCS patterns, as ways of interacting socially and within the physical world. As the conditions for use of these LCS patterns are internalized, individuals develop resources that are mobilized in certain situations to perform these patterns.

An assessment is an opportunity for the performance of such LCS patterns. Mislevy (2018) writes about assessment design as follows: "An item writer crafts encapsulated situations around LCS patterns, so its features tend to activate resources for understanding and acting accordingly in persons who are attuned to them" (p. 393). To the degree that these patterns may involve language, social interaction, cultural meaning, and perceptual/motor engagement with one's surroundings, conditions surrounding their formation and activation may be subject to many permutations. In this formulation, complexity around reliability and validity are introduced by the need for the sociocultural context within which these patterns were learned to parallel those of the assessment for the appropriate resources to be activated. Important for our purposes is this supposition that resource activation may involve variables ranging from something as simple as familiar vocabulary to something as intangible as the triggering of stored schemata for processing new data, and that it can extend to elements of the physical world, such as the tools and interfaces that form the stage for the performance of LCS patterns.

In his 2018 book and throughout his work, Mislevy is concerned with the evidentiary value of test scores, which relies on both the idea that similar test scores on the same assessment across different groups of test-takers or different years has a similar meaning (reliability) and that the test is measuring what we think it is measuring (validity). In a 2017 book chapter, Michael Kane collaborates with Mislevy in an exploration of validation evidence based on process models. This method involves developing a fine-grained procedural model and an inventory of required knowledge for successful task performance. The task within an assessment instrument is designed with "features and directives [that] activate the targeted cognitive processing, at least in proficient test takers" (p. 14). In pursuit of evidence of validity, the characteristics of the responses, patterns across responses, and, if available, other data (response times, eye-tracking data, captured screen activity, and verbal protocol transcripts) are scrutinized to see if successful solutions tend to align with the granular elements of the proposed procedural model (Kane & Mislevy, 2017).

Kane and Mislevy's process models bring another dimension to our proposed task modeling framework, whereby procedures associated with construct mastery are defined, assessment tasks are designed to evoke the kind of performance described by the model, and a wide range of data can be used to confirm or confound the proposed interpretation of scores. In fact, each of our disciplinary forays has contributed additional perspectives for analyzing the design of assessment tasks, the test-taker's performance that is evoked, and the meaning that can be attributed to a scored response. Although CLT provides mechanisms to analyze the cognitive processes that are recruited by a given task and by the tools provided for performing the task, HCI further breaks down the elements of the performance environment to understand the factors of perception, motor skills, ergonomics, interpretation of the user interface, and engagement with the physical input devices. Aligning these avenues of investigation within the specific concerns of assessment, such as construct validity and replicability, provides the final dimension for a rigorous analytical method.

FRAMEWORK CONSIDERATIONS: WRITING

As described in the prior section, we propose a methodology for investigating device and interface effects rooted in an understanding of the cognitive and sensorimotor demands of a task. Because we have chosen to focus on NAEP writing as an entry point to considering device/interface effects, we begin with cognitive theories of writing and how the demands of writing operate within the context of an assessment.

A Cognitive Model of Writing

Deane (2011) present a sociocognitive framework that attempts to connect writing instruction and assessment with social and cognitive theories of writing. Synthesizing across a number of cognitive models of writing, Deane distinguishes among several forms of representation that play critical roles in the cognition of composition:

- Social and rhetorical elements (rhetorical transactions between writer and audience)
- Conceptual elements (representations of knowledge and reasoning)
- Textual elements (representations of document structure)
- Verbal elements (linguistic representations of sentences and the propositions they encode)
- Lexical/orthographic elements (representations of how verbal units are instantiated in specific media, such as written text)

For each of these forms of representation, there are corresponding skills. In particular, orthographic skills include the ability to convert words and sentences into printed matter; that is, the cognitive abilities to produce words and sentences in written form. Although not a focus of most cognitive models of writing, these skills have an obvious role in writing production and writing assessment. Deane and colleagues (2011) refer to these as "print model" cognitive processing skills that for writing include spelling, word recall, knowledge of conventions, and the motor skills supporting handwriting and typing. Print model skills also are posited for reading; some are shared with writing (Berninger et al., 1994) and others are unique, such as decoding, orthographic conventions, word recognition, and knowledge about how printed text is parsed to approximate speech equivalents. For writing assessment (and more broadly, assessments where some level of reading is required to access assessment tasks), print model skills represent endpoints within which other cognitive skills (social/rhetorical, conceptual, textual, and verbal) operate. If a particular digital device impacts performance on a writing assessment, a reasonable explanation is that the device is interfering with the lexical or orthographic elements of the test-taker's written response. If the assessment task involves both reading and writing, a particular digital device could impact both types of print model skills.

Text Production, Typing Skills, and Cognitive Load

An aspect of text production that takes place simultaneously with these cognitive processes is the engagement of the test-taker's perceptual and motor capabilities, interacting with a device and peripherals to type, read, and edit. Prior research has identified typing/keyboarding skills as a factor correlated with writing performance (Goldberg et al., 2003; Graham et al., 2012). We will return to this correlation and an analyses of some of these studies in the context of revision and the role of overall computer skills. But to first explore this idea from a cognitive load perspective, we draw upon the idea of automaticity regarding tool use. Sweller and colleagues (1998) defines automaticity as "having acquired sufficient familiarity that a task can be performed accurately and fluidly without conscious effort" (p. 258). More specific definitions of fluidity and fluency exist for typing. According to de Smet and colleagues (2018), "…fluent writing means that writers make a transition from one keystroke event to the next one without exceeding an individual interkey interval threshold that is necessary to realize a motoric transition from one key to another" (p. 417).

Although the implication is that the cognitive demand of typing may be negligible for a fluent typist whose keyboard mastery has reached the level of automaticity, many students fall short of that mark. Various studies focused on cognitive load have shown evidence of a cognitive cost due to the degree of attention that must be dedicated to motor execution on the part of the nonfluent typist. Alvès and colleagues (2007) cite about a dozen such studies, including a 1994 study by Bourdin and Fayol where they were able to generalize this effect to other types of "untrained" writing. Bourdin and Fayol were able to compromise recall while writing when adults were instructed to write using only cursive capital letters. Graham (1999) likens difficulties with text production skills to trying to write with a Chinese typewriter (the most challenging typewriter in the world to use since it includes 5,850 characters). The writer spends so much effort searching for the characters needed to produce words that they lose track of the ideas and plans they intended to express. Graham's example was in the context of children with learning disabilities and is somewhat of an exaggeration but does illustrate the importance of minimizing the extraneous cognitive load associated with using a digital device and interface to produce assessment responses.

Text Production and Device Form Factors

Fine motor skills related to typing are shaped in part by muscle memory, which consolidates the motor tasks involved in typing into memory through repetition. It follows that muscle memory will be most applicable to a keyboard that is familiar or has similar characteristics to a frequently used keyboard. Quantifying the impact of an unfamiliar keyboard in terms of typing speed and accuracy may be difficult, as a number of differences between keyboards are possible. Nonetheless, some keyboard attributes that may differentiate keyboards and that have been correlated with differences in typing speed and accuracy include key size, key travel distance, and tactile/haptic feedback.

Prohibiting use of on-screen (or virtual) keyboards for longer writing sections within high-stakes assessments, as is done by NAEP, is fairly standard practice for a number of reasons. Despite the use of haptic feedback on some virtual keyboards, learned touch-typing skills generally have limited transferability to such keyboards. The keyboard may obstruct on-screen content and require that a test-taker frequently switch between hiding and activating the keyboard when alternating between writing and reading previously written text. The key size on virtual keyboards tends to be small, and the differences between the ideal device positions for typing versus viewing can lead to awkwardness if not outright ergonomic issues, as writing involves alternating between typing and viewing. In a study of virtual keyboards, Kim and colleagues (2014) found a 60% reduction in typing productivity and decreased self-reported comfort, as well as ergonomic findings suggestive of an increase in shoulder muscle activity.

Despite this cautionary approach regarding virtual keyboards, some evidence exists regarding novice typists, "hunt-and-peckers," who benefit from the visual proximity of the on-screen keyboard to the text being typed as opposed to the greater distance in eye gaze shift between a monitor and a physical keyboard. Although capitalization often needs to be triggered on and off through multiple keystrokes (i.e., "sticky caps") on virtual keyboards, younger students, particularly those with shorter finger reach, may find this preferable to challenging two-fingered maneuvers (shift plus another key) on standard keyboards. In addition, for these younger students who look at the keyboard while typing, there is the benefit of visual feedback with the key appearance changing to reflect the capital letters unlike the more subtle appearance change for some keyboards' caps lock functionality (Davis & Strain-Seymour, 2013b). Despite these possible advantages for keyboard neophytes, typing instruction tends to precede or accompany digital writing curricula and online writing assessments. Thus, assumptions can be made (and have been made by many state assessment programs) regarding the desirability of physical keyboards for students who have learned touch typing.

Key size is a factor when choosing a keyboard. Guidelines for the horizontal and vertical distance between keys suggest an ideal range between 18 and 20 mm. This distance accounts for an inactive space surrounding the key of around 3 mm and a key size of 16 to 18 mm in width to be considered full-sized (Gunawardena, 2013). The key sizes on compact and virtual keyboards, on the other hand, may be as small as 9.5 by 9.5 mm.

The most comprehensive research regarding the impact of key size on typing performance was conducted with males with large fingers (based on a middle finger length of 8.7 cm or greater or a finger breadth of 2.3 cm or greater). In two separate studies using large-fingered subjects, Pereira and colleagues found that some compact keyboards contain key sizes falling beneath the threshold for performance effects on typing. The first study (Pereira et al., 2013) found that keys with a horizontal width of 16 mm or less were associated with reduced productivity and usability ratings, while the second study (Pereira et al., 2014) found a similar effect with keys with a vertical length of 15.5 mm or less. Although research with virtual keyboards cannot necessarily be assumed to translate to physical keyboards, typing speeds 15% slower were noted with 13 by 13 mm key sizes on virtual keyboards used by a combination of male and female adult subjects (Kim et al., 2013). Unfortunately, no conclusive research conducted with children exists to understand the degree to which key-size recommendations for adults hold true for smaller hand sizes. In the absence of further information, we might surmise that keyboards with key sizes between 15 to 18 mm to be optimal and those with smaller key sizes to be eyed with caution if significantly different from those used in the classroom or for students with larger hands.

Compact keyboards may not only have smaller keys; they also may be thinner than conventional keyboards, which typically means reduced key travel (i.e., the distance that a key needs to be pressed down before the keystroke is recognized). Key travel distances can range from 0.0 mm (such as for a virtual keyboard) to 6.0 mm, with typical distances for conventional keyboards ranging between 2 and 4 mm. With reduced travel distance, typically less force is required for each keystroke. However, no significant differences have been found in fatigue or physical risk factors across a range of key travel distances associated with portable keyboards (Sisley et al., 2017). Another study compared keyboards with four different key travel distances, finding higher words per minute with 1.6 mm and 2.0 mm keyboards than with 0.0 mm and 0.4 mm keyboards, and decreased accuracy along with lower subjective usability ratings with the 0.0 mm key distance (Hoyle et al., 2013). A similar study by Chaparro and colleagues (2014) compared a thinner keyboard with a low key travel distance and pressure-sensitive keys with two more standard keyboards with mechanical keys, and found typing speeds to be about 10 words per minute slower in addition to increased typing errors with the thinner keyboard.

Key size and key travel are only two of several possible differences that can differentiate keyboards. Other differences can include key layout, such as the inclusion of a separate 10-key number pad, auditory feedback (including the ability to turn off such feedback), ergonomic designs, backlighting, variable key label size/contrast, color coding of the keys, and the level of integration with the device itself for keys such as escape, control, and function keys. Of particular interest is the inclusion of arrow, page-up, and page-down keys to support a variety of personal preferences, habits, and learned approaches that may exist in terms of cursor movement and control over scrollable contents. In addition, the amount of prior use a keyboard has been subjected to also may impact responsiveness and clarity of key labels, among other things. Portable, low-cost keyboards purchased for the classroom also may show the effects of wear and tear to a greater extent than full-sized keyboards that stay put in computer labs rather than getting packed up at the end of each class.

The Coordinated Activities of Writing and Reading

Although appropriate keyboards help support bursts of fluid typing by experienced typists, the writing process is far more complex than the linear transcription of thought into on-screen text. In an elucidation of the complexity of writing tasks similar to Deane's described above, Flower and Hayes (1981) suggest the interleaved cognitive subprocesses of planning, transcription, and review as a part of their cognitive process theory of writing. Although some procedural views of writing posit formal breaks between prewriting, writing, and revising, Flower and Hayes suggest that actual practice involves constant movement between these modes. This theory is supported by eye-tracking and keystroke logging studies of digital writing, such as that by de Smet and colleagues (2018). De Smet et al. present conclusions from their own research as well as prior studies regarding the amount of reading that occurs with text production. They cite evidence that more text reevaluation conducted through reading occurs during the writing process as opposed to within a separate stage dedicated to proofreading. Eye tracking shows that a writer may even be reading previously generated text concurrent with typing new text (de Smet et al., 2018).

In Hayes's revisiting of the earlier Flower and Hayes model for the cognitive processes of writing, he emphasizes the role of reading as an effort intensely coordinated with the larger writing effort (Hayes, 1996). He distinguishes between local reading of the text produced so far (referred to as TPSF), such as to check spelling and grammar within a single sentence, and more global reading that can serve multiple purposes: evaluation, refreshing one's memory as to what has already been written, and generation of new ideas that flow from the rereading process. Another notable factor regarding reading in the context of writing is that some evidence suggests that reading to identify and fix issues involves a higher cognitive load than reading for understanding (Piolat et al., 2004).

Reading and Device Form Factors

With the significant role of reading within the writing process, various device form factors and interface design features that can impact on-screen reading are in play. Interfaces for reading passages are often subject to a number of design decisions related to font, contrast, scrolling versus paging, and imitation of book, magazine, or web page layouts. Some test delivery interfaces provide options to toggle between a view that dedicates more screen space to a passage and a view that divides available space between the passage and questions about the passage. This careful attention to passage design is justified by the reading comprehension construct. Anything that imposes extraneous cognitive load on the reading process may compromise validity.

An assessment system designer who does not explicitly acknowledge reading as a critical aspect of writing may not pay the same amount of attention to optimizing an open text area for reading purposes. Text legibility and an interface that maximizes the amount of TPSF that is visible at one time are both valuable. For those assessments that do involve writing to sources, an interface that supports the ability to easily switch between reading source material and TPSF, as well as the ability to view both at once, provides the test-taker with appropriate flexibility.

Another factor impacting how much of TPSF is visible at one time is screen size. The existing research on screen size has not been solely focused on writing but confirms the assertion that, assuming legibility, the most impactful factor appears to be the amount of content that is visible rather than the size of the content, as long as zoom options are available and known to users (Davis et al., 2013). Bridgeman and colleagues (2003) noted a particular impact on reading: Verbal scores were lower by nearly a quarter of a standard deviation when less reading material was visible on screen without requiring the user to scroll. Although Bridgeman et al.'s findings relate to the display of reading passages and possibly the difficulty of recall when working with reading comprehension items, one possible conclusion relates to an out-of-sight-out-of-mind phenomenon: Test-takers are aware of off-screen content, but the extra effort of scrolling appears to be an inhibiting factor. It is possible that a similar effect would apply to reading TPSF and that fewer errors might be recognized via casual glance with more off-screen content. For this reason, the recommendation from prior studies (Davis et al., 2013; Keng et al., 2011) to not dip below a screen size of 10" seems prudent to apply to writing assessments.

Revision: The Impact of Devices, Interfaces, and Computer Skills

Reading one's TPSF outside of the immediately-visible active text production area, in what Hayes calls global reading, may be for the purposes of further text production, such as picking back up on a train of thought or remembering what was signposted in an earlier part of the essay. However, it is often for the purposes of revision, whether or not that was the original intent when the test-taker began the process of reading some portion of TPSF. In addition to rereading, revising, and proofreading in earlier portions of the essay, revisions also occur within the active area of text production, such as using the backspace key or character-by-character movements of the cursor using an arrow key, making changes in a word or sentence as it is being typed. Thus, revision can be interwoven with either reading or text production.

In terms of the intrinsic cognitive load of revision, differences exist based on the type of edits. According to Piolat and colleagues (2004), correction of a spelling error involves low element interactivity, and thus a lower cognitive effort, when compared with syntactical errors that require engagement with a full sentence or transitions between sentences. Piolat et al. also compare correction of surface errors (grammar, spelling, capitalization, punctuation, and word choice) to coherence issues that a writer may be trying to resolve with more global revisions: "…for global revisions, writers build ill-defined representations of the problems that require them activating high-demanding reflection processes. Consequently, revision of coherence errors requires a greater amount of resources than revision of surface errors" (p. 3).

Although many of these observations about revision behaviors hold true for pen-and-paper writing as well as digital writing, revision patterns differ across modes. NAEP and others have acknowledged these differences, considering digital writing to be a separate construct rather than trying to resolve the issue of comparability across modes. Students in cognitive labs and follow-up interviews describe the digital writing process within an assessment as requiring different steps and strategies in comparison with "blue book" writing (Way et al., 2008). They describe an abbreviated prewriting stage, sometimes forgoing prewriting entirely for a stream-of-consciousness approach, with more far more attention paid to as-you-go revisions as well as wholesale reorganization by rearranging text. Thus, digital writing is often referenced as a fragmentary process with frequent shifts between text production and revision, with various evidence suggesting that more revision occurs than in handwritten writing (Arms, 1983; Van Waes & Schellens, 2003).

A question regarding digital writing as a separate construct from paper-based writing is the degree to which basic word-processing skills should be considered a part of the construct and, if so, whether this includes the use of tools designed to improve writing, such as spelling/grammar checkers and right-click or search-based digital thesauruses. A test delivery system's writing interface should reflect the outcomes of these decisions. To the degree that this interface is most likely an abbreviated text editor, in comparison to Microsoft Word or Google Docs, an assumption is being made that test-takers' text editing skills are transferable to this environment. This may not be an outlandish assumption with the ubiquity of text editors in educational and personal technologies and with well-established common conventions for text editors. Examples of common conventions include text selection used with formatting buttons, keyboard shortcuts to apply styling such as italics and bold, and an "undo" button with a counter-clockwise curved arrow. Nonetheless, cognitive labs and usability studies to verify an interface, the availability of tutorials or sample tests using the same interface, and tutorials are often standard practices for large-scale assessment programs with writing components.

Revision practices within a writing task arguably engage computer skills as well as fine motor skills more than any other assessment task. Test-takers may be making use of keyboard shortcuts and right-click contextual menus. They may be exercising precise control over text selection and cursor placement. Hold-and-drag operations may be used to select and move text from one location to another. Pressing down on an arrow key briefly or repeatedly, or holding it down to prompt faster cursor movement are learned techniques to be deployed while engaged in revision operations. In addition, spatial and computer literacy skills may be engaged to understand the size and position of a scrollbar (even as the amount of text changes) to navigate back to a particular part of TPSF.

Not only does the deployment of these skills require that all of these actions be supported by the test delivery system's writing interface (and possibly enabled in multiple ways to support a variety of habits, such as with cut/copy/paste), but the tablet/computer, keyboard, and input device can introduce differences. These may be differences from what a student is accustomed to and/or differences between students using different devices. Text interaction and scrolling operations can be controlled in part by a device's graphical user interface (GUI) conventions, which means that the test delivery system's writing interface may need to account for these differences, with quality control and usability testing on a variety of devices. Similarly, the location of buttons to move the cursor and backspace/delete keys, and the accessibility of special characters can vary by keyboard.

The precision of an input device and the test-taker's ability to use that input device are critical to cursor placement and text selection/deletion. In cognitive labs, students expressed frustration with touchscreens when editing text, sometimes switching from finger to stylus and not always locating or being aware of arrows keys to control cursor position on an unfamiliar keyboard (Davis & Strain-Seymour, 2013b). A mouse, in addition to often having a scroll wheel, has input precision that is only limited by human vision, in comparison to finger-based touch input, which requires target sizes of at least 40 pixels according to standard usability conventions. Styluses offer greater precision but involve some occlusion, although less than fingers. Trackpads and track-points (also called nubs and other less appropriate names) may offer some level of precision but are not suitable for test-takers unaccustomed to them.

The size of the device screen or monitor also may have an impact on motor movements associated with revision practices. If the text is smaller on a smaller device, then cursor placement may be that much more difficult if the input device lacks precision. In addition, when text is dragged from one location to another on a small screen, there is a greater likelihood, in comparison to larger screen, that the new location for the text is not currently visible, thereby making this text repositioning a potentially complex operation.

Negotiation of these various factors—transferability of text editing skills to the assessment writing interface, comfort with the device, input device precision, motor skills, screen size, and so on—all play a role simultaneous to the cognitive processes dedicated to the revision process and to the overall linguistic and orthographic maneuvers inherent in writing as described above. Nonetheless, research on the impact of devices, level of computer proficiency, and keyboarding skills on writing performance are mixed.

In regard to computer/keyboarding skills, Barkaoui's (2013) study of the impact of typing skill on TOEFL [Test of English as a Foreign Language] writing tasks begins with a review of the conflicting research in this area. He concludes that a common limitation in these studies pertains to how computer proficiency and keyboarding skills are measured. Self-reported measures of computer use and familiarity based on interviews and questionnaires tend to overestimate computer ability (Barkaoui citing Larres et al., 2003). NAEP research has attempted to avoid this weakness by including hands-on exercises to capture more objective indicators of typing speed (words per minute), accuracy (typing error avoidance), and editing skills (text correction, insertion, deletion, and repositioning) (Horkay et al., 2006). Barkaoui similarly uses typing tests to determine keyboarding proficiency, but his approach is less comprehensive than that of Horkay et al. in terms of evaluating computer-based text manipulation used within editing, perhaps in part because the assessment writing interface used by Barkaoui was limited in terms of capabilities emulating a full word-processing environment.

Nonetheless, Barkaoui's findings are ultimately illuminating, even if we must acknowledge differences between English language proficiency testing and K–12 writing tests. He finds a weak but significant correlation between keyboarding skills and writing performance after controlling for the effects of overall English language proficiency. He notes that keyboarding skills have a greater impact on one part of the TOEFL (the independent task) than on the other (the integrated task): "The independent task seems to have required more writing (and typing) and to be more cognitively demanding as it requires the generation, planning, organization, and typing of more content compared to the integrated task which involves summarizing ideas from the reading and listening" (p. 13). This suggests that the cognitive cost of nonfluent keyboarding, or potentially other sources of extraneous cognitive load, may not be measurable outside of the extensive demands of longer writing tasks involving planning and revising. As we revisit device studies related to writing, such as that by Davis et al. (2015), where no device effects were found with short essays, we might apply Barkaoui's findings and wonder whether the results would have been the same with longer essays.

Motivation and Device Performance

Review of the literature on mode and device comparability, as well as consideration of theoretical aspects of cognitive load and HCIs, suggests the potential for construct-irrelevant factors that could affect NAEP performance. One additional factor that potentially contributes construct-irrelevant variance to NAEP performance is student motivation. NAEP was originally designed to have no stakes for students, parents, or teachers, partially out of the fear that it would otherwise be the first step in an evolution toward a federally mandated national curriculum (Bracey, 1996). A significant change to NAEP design was proposed by Messick and colleagues (1983) and served as the methodological basis for today's NAEP as well as significantly raising the profile of NAEP in the measurement community. One question raised about NAEP under the new design concerned whether student performance was generally underestimated because of NAEP's "low-stakes" nature (Kiplinger & Linn, 1995; O'Neil et al., 1995). Research suggested that although NAEP did not seriously underestimate student performance compared with assessments with more moderate stakes (e.g., state testing programs), one might characterize NAEP scores as representing what students will demonstrate with minimal effort. In a qualitative study examining why students skipped some Grade 8 reading and civics items, Jakwerth and colleagues (1999) found that students' engagement was a factor, and that motivation was a problem for children attending schools that served the most disadvantaged students. In a more recent study, Braun and colleagues (2011) found that Grade 12 students who were given either fixed or contingent incentives did significantly better on NAEP than those without an incentive.

In general, although assessment practitioners have suggested that NAEP performance is likely affected by lower student motivation, especially at higher grade levels, little attention has been given to the possibility that motivation, in and of itself, might have trend aspects that contribute to NAEP scores over time. Although NAEP student questionnaires include items that probe student motivation and effort, recently (e.g., between 2015 and 2017) these questions have been revised and, as a result, it will be more difficult to interpret responses to these questions as related to the transition to digital NAEP assessments.

In providing a framework for understanding cognition and affect in writing, Hayes (1996) includes two major components: the task environment and the individual. Part of the individual component focuses on motivation and affect, specifically as related to goals, predispositions, beliefs/attitudes, and what Hayes refers to as "cost/benefit estimates." These individual characteristics clearly come into play during a writing assessment, in particular as students encounter the particular testing situation posed by NAEP writing. Hayes concludes that changes in writing media can influence the cognitive processes involved in carrying out writing tasks, in part because cost-benefit estimates of effort and strategy are impacted by the familiarity and comfort with the writing interface.

In his recent book on the sociocognitive foundations of educational measurement, Mislevy (2018) emphasizes the interactions of linguistic, cultural, and substantive (LCS) patterns with assessment situations. He also introduces the terms "emic" and "etic," respectively, to distinguish between meaning as construed by the individual (e.g., the test-taker) and meaning construed externally (e.g., by the test-maker's framework of the testing situation and performance). This distinction can lead to contrasts when one considers the difficulty of assessment tasks. From an emic point of view, Mislevy says, "A task is difficult for a student if she is not able to activate resources for LCS patterns needed to perceive a situation, understand it, and act effectively as seen from the assessor's perspective" (Mislevy, 2018, p. 83). More revealing in the context of NAEP is a footnote offered by Mislevy at the end of this sentence, where he expands some on this statement:

> Recognizing an assessment is low stakes, deciding not to engage with it, and gaining an hour to think about other things is an example of a student perceiving a situation and acting effectively to maximize his own objectives. It is rational, intelligent, and an effective use of resources. It draws on understandings of practices and LCS patterns. It is just not aligned very well with the objectives of the assessor (Mislevy, 2018, p. 100).

Applying an emic perspective to NAEP, it seems worth asking whether test-taker motivation is an issue that should be reconsidered at this point in the evolution of NAEP:

- There have been increased criticisms of standardized tests over the past 20 years, with increasingly cited themes such as "there is too much testing in the schools" and "standardized tests provide little or no value to teachers, students or schools." Students and teachers hear these criticisms all of the time.
- NAEP is the ultimate "drop in from the sky" assessment (to use Mislevy's term); There is no classroom preparation or performance feedback given to students or teachers. Students and teachers know this, and it could be that this knowledge has a different impact today than it did in the past.

- Although some test-takers may find NAEP's digital formats more engaging than paper, particularly for interactive science items and scenario-based tasks, other test-takers' motivation could suffer due to the unfamiliar test format and required information retention while progressing through a multistep task or scenario.
- NAEP is delivering digital assessments using Microsoft Surface Pro, which recent data have suggested are only used in about 1% of classrooms for instructional purposes (EdWeek, 2017).

Current NAEP administration procedures include preassessment activities to encourage participation, with the following goals as listed in the current NAEP School Coordinator Manual:

1. Plan activities to notify participating students of the importance of doing their best on NAEP.
2. Plan activities to notify teachers and other staff to encourage students to arrive on time and to do their best on NAEP.
3. Document activities the school has completed or will be doing to notify students and school staff about the importance of NAEP.

These activities are estimated to take about 1 hour, and the coordinator manual includes links to videos that provide information about NAEP as well as a nonpublic site where additional strategies to encourage student participation and engagement can be accessed. The coordinator's manual also provides step-by-step planning instructions for the day of assessment, which includes considerations for selecting locations for tablet testing. However, there can be a range of specific NAEP preparation activities that are enacted across schools.

FRAMEWORK CONSIDERATIONS: NONWRITING SUBJECTS

To the degree that all subjects may involve some reading and open-response items, many of the same device/interface issues discussed in relation to reading and writing tasks may be applicable. However, variables unique to social studies, math, and science assessments have relevance for device selection and interface design. Such variables include an increased reliance on interpretation of visual stimuli, tools (e.g., calculators, rulers) and nontextual annotation and expression.

Visual Stimuli

Although image interpretation is not unheard of within English language arts (ELA) assessments, math, science, and social studies generally involve a range of skills that require visual interpretation of nontextual stimuli. Device screen size, magnification controls, screen clarity/contrast, and antiglare characteristics may all impact the testing experience in the case of large and/or detailed stimuli. For example, in a social studies test, a smaller screen or the need to magnify may force a test-taker to make choices about which subsets of available content can be viewed simultaneously: an area of interest on a map, the map key, a source line for a map, and the response area for an assessment task that asks the test-taker to evaluate the historical value or accuracy of the map. Visual renderings of primary documents, images of political cartoons printed in newspapers with low resolution, and many map shadings to distinguish geographical features may require careful visual attention unhindered by glare, limited contrast ratios, or low refresh rates. Controlling for all variables—device wear-and-tear, ambient lighting, brightness settings on the device, and test-taker awareness of ways to adjust brightness on a device—quickly gets complicated, even without considering individual students' visual acuity. For NAEP, the factors of task design, limited device usage, and setting checks may be the easiest to control. One area for attention with a phased replacement of devices is variable device age amidst the overall set of NAEP devices. Although loss in screen luminance over time should be negligible in a low-use device, newer generations of tablets and computers tend to have more powerful graphics cards and a higher number of nits (a measure of luminance with one nit equivalent to one candela per square meter). For instance, some generations of Microsoft Surface Pro have brighter screens than prior versions, typically with more nits than many Chromebook models (350–375 nits) but fewer than on most iPads (450–500 nits). Nonetheless, assuming that a testing location is not bombarded with bright sunlight and that standard accessibility guidelines for image contrast are followed, the range of most Microsoft Surface Pro devices (low 400s) should be sufficient.

The factor of detailed visual stimuli requiring magnification for some students, even if not specifically identified as having a visual impairment, is typically managed through item design. However, the need to see details may still arise in map labels; the smaller intervals on rulers or illustrations of scientific equipment; small symbols, such as degrees and chemical compound notation using sub-/superscript; and small but meaningful font and character differences within math equations (e.g., for length versus the number 1 or a superscript 8 versus a 9). Student familiarity with a device can lead to the ability to quickly magnify a detail without losing the flow of one's thoughts (i.e., without adding extraneous cognitive load). As device or operating system-specific magnification functionality can be hidden— controlled by gesture or keyboard shortcut— a replacement for device familiarity is a magnification tool provided by the test delivery interface that is always accessible, highly usable, and introduced via tutorials. Usability in this case involves the ability to purposefully (not accidentally) trigger the magnification, control the level of magnification, move the magnifier or object of magnification, and turn off magnification. As seen in other examples, the absence of automaticity of some actions that comes from extensive device familiarity can be partially offset by the assessment interface, when tools with equivalent functionality are provided and adequately introduced to test-takers.

Calculators

A unique factor in math (and some science) assessments is calculator use. Calculators can be separate devices or integrated into the test delivery platform. Physical calculators can offer the advantage of familiarity when they are the same devices as used in the instructional environment. Like typing, knowing where to find the proper key, using visual and muscle memory, while remaining focused on the task is optimal for complex items that may tax cognitive resources. Other advantages of physical calculators include tactile feedback and the ability to work with a calculator in proximity to scratch paper.

Many state and consortia-based testing programs have adopted the use of calculators incorporated into the testing software. For more complex calculators, such as graphing and scientific calculators, the ability to incorporate Texas Instruments (TI) emulators into testing platforms, often without additional cost to assessment vendors, leverages the familiarity stemming from the TI-84 calculator's ubiquity in upper level math classes. The advantages of such calculators embedded into the testing software include improved security (no hints or formulas stored in memory), lower costs, and simpler logistics for schools as the need for calculator distribution, memory clearing, and battery hoarding disappears with test-embedded calculators. For assessment designers, the ability to include a calculator with some items but not others allows for greater nuance. Items measuring estimation skills or requiring algebraic computation can appear without a calculator, while other items can focus on process and analysis with the availability of a calculator, leading to less impact from careless arithmetic calculations. One drawback, however, is that student consumers of alternate calculator brands, possibly lower cost models, may end up disadvantaged on testing day when grappling with an unfamiliar calculator.

Alongside the benefits of familiar, embedded calculators are some issues to be managed. Limited screen space and resolution translates to a trade-off between the usability of a larger calculator (bigger keys as mouse/stylus-click targets and more visible labels and displays of calculated results) and the disadvantages of content occlusion. Fitts's law predicts that the time required to move to a target area is a function of the ratio between the distance to the target and the target's width (Fitts, 1954), and can be useful in considering calculator placement as well as key size. Although applicable to many types of goal-oriented human movement, Fitts's law is often referenced in HCI circles when discussing usability related to click or touch targets (MacKenzie, 1992). Usability is quantified as time in this case—a user action that takes longer is more difficult than one that is nearly instantaneous. Cursor or hand travel across long distances takes time, and acting with precision to hit a small target, such as a calculator button, takes time. When a test-taker is interacting with test content and an embedded calculator, the ability to move the calculator within the test area is important to limit travel distance and avoid blocking the test-taker's view of the item. The requirement of a test-taker to hold one or more numbers in short-term memory while using a calculator that blocks the content is a classic example of unnecessarily increasing cognitive load. A calculator that can be toggled back and forth between showing the entire calculator and just the output screen also can add convenience in limited real estate situations. Fitts's law also suggests that a physical calculator may optimize performance by minimizing time if a test-taker is doing large portions of work on scratch paper. Conversely, an embedded calculator may be preferable when a test-taker, based on item characteristics or preference, does not use scratch paper.

As of 2017, TI quoted the number of high-stakes assessments using a TI calculator at 60 (McFarland, 2017). Although no competitor has significantly threatened TI's 80% share of the international calculator market over the past 20 years, the secondary school calculator market has shifted slightly over past couple of years with 40 million users of the Desmos calculator in schools. Desmos founder, Eli Luberoff, intended to lessen the impact of economic disparity by offering online calculators at zero cost to students and schools while charging assessment and instructional system vendors (Crockett, 2019). With this small market shift, some assessment systems used for high-stakes testing now offer a calculator choice, either providing both options to the test-taker or configuring the system to use either Desmos or TI based on a state customer's preference.

Although feature-by-feature comparisons of the calculators are beyond the focus of this study, a comparison of graphing calculators—the calculator versions that require the most space to be highly usable—highlights some trade-offs. The original TI-84 emulator, as a 2D equivalent of the physical calculator, makes visible all keys and selection possibilities at once, such that a visual scan to locate a function or symbol across a contained, although crowded, space is possible. The Desmos calculator, on the other hand, with its design to work on computers and mobile devices of all shapes and sizes, uses space more flexibly. The user can control whether the keyboard is visible, and the screen space used by the calculator can be modified—not simply proportionately scaled but stretched or squished across either dimension to fit into available space, with the responsive design responsible for optimizing the display within that new footprint. A test-taker may use this feature to optimize what portion of the item content is viewable simultaneously with calculator use or to enlarge one's view of either the calculator keys or the output display. However, many selection options are not visible without active engagement on the user's part to open several palettes to locate a desired function. A casual glance is less likely to lead to the accidental discovery of a symbol or function that triggers a "that might help" discovery moment for the test-taker. Various arguments could be made regarding the comparative number of functions supported by each calculator or how quickly the nuances of one calculator's functionality can be learned. Nonetheless, although not confirmed by research, one might theorize that the most impactful variable is test-taker familiarity with one calculator model versus the other. Regardless of make/model, a final consideration for large-footprint graphing calculators viewed on small device screens and made available through a test delivery system is the ability to close and reopen the calculator (such as to view the content beneath) without triggering a calculator refresh that loses the displayed calculation/graph.

Math Markings: Annotation

Discussions of online marking tools applied to assessment often focus on annotations, note-taking, and answer-option elimination. In Adler and van Doren's (1972) discussion of active reading strategies, the authors differentiate annotations as markings tied directly to the text (e.g., marginalia, sticky tabs, underlining, highlighting) and note-taking, where those notes exist independent of the text and do not necessarily require the original text to have meaning. The implementation of such functionality within digital testing environments and online instructional environments is often text-bound. Highlighting tools allow the test-taker to visually differentiate text, sometimes even with the availability of different colors. A notepad allows a test-taker to type in notes that are stored with an item or passage and access those notes whenever that item or passage is displayed. Answer elimination tends to be text strikethrough or some type of marking that covers the general area of the answer option text. Such features of a test delivery system may provide utility for test-takers across a range of subjects. Nonetheless, in math tasks, the ability to use text-bound tools to support active engagement with the content may fall short when a large portion of that content is nontextual. Understanding the nature of math marks—those tied to and those independent of one or more visual stimuli—and how they support spatial thinking, mathematical reasoning, and other process skills is essential.

One researcher focused on interfaces that support dynamic information processing, Sharon Oviatt, uses the term "gesture" to describe nontextual expressions that traverse space in a nonlinear way and that may be combined with linguistic expression. These gestures may leave marks or be fleeting, and they may be analogue or computer based. According to Oviatt (2006), "The physical activity of manual or pen-based gesturing is believed to play a particularly important role in organizing and facilitating people's spatial information processing, reducing cognitive load on tasks involving geometry, maps, and similar areas" (p. 874). Such gestures may be annotations (i.e., markings operating within the 2D space of a visual stimulus within a math item). For instance, in a paper-based test or assignment, a student working on a problem related to the reflection and subsequent rotation of a figure might support their thinking by making marks to indicate the position of a figure during each stage of a transformation. Similarly, a student identifying parallel lines, complementary angles, and right angles in a figure to determine the measurement of a mystery angle might draw "feathers," hash marks, arcs, and boxes while problem-solving. Cognitive load is reduced by allowing a student to break a larger problem into its component parts, solve one at a time, and rely on the markings instead of maintaining all parts of the solution in working memory.

If available, freeform marking tools—ones that are not text-bound—within the digital testing environment can provide a rough equivalent of such paper-based marking activities, while avoiding one major drawback of using scratch paper for such tasks: the difficulty of redrawing the visual stimuli quickly and accurately. The usability of freeform marking tools is a matter of both interface design and ergonomics, as we acknowledge that on-screen drawing can be awkward using a mouse on an upright screen. One could argue that computers and tablets that can be used with a keyboard, mouse, touch, and stylus, and that can easily be repositioned to optimize how a user works with an input mechanism, offer test-takers ample flexibility to adjust input device and screen position as needed. Such flexibility can be valuable on end-of-year tests that cover a wide range of task types. For instance, typing a short answer to explain a procedure calls for keyboard usage and an upright or angled screen. Marking up a figure as described above may be best done with a stylus on a horizontal screen, positioned like a graphics tablet, or a screen with enough support from behind to counteract the pressure of the stylus. What skills are required for test-takers to deftly switch input methodologies and screen positions to match the nature of a task? One could argue that they include comfort with the device, prior experience with multiple input methodologies, and metacognition to recognize how to optimize one's own assessment performance. These may not be skills that are honed within the classroom, if the focus on a narrow range of concepts and a limited set of problem types at a time is typical. A teacher may instruct students to engage with a physical or device-enabled graphing calculator during a work session focused on exponential functions. Paper and pencil, on the other hand, may be the chosen tool for work on an algebra problem set. In this way, the instructional environment may not provide adequate preparation for an assessment that combines a range of task types delivered via a device that offers multiple input modalities.

Math Markings: Equations and Symbols

Another type of math marking involves mathematical equations and expressions, whether to support the procedural steps to arrive at a final answer or to express the final response. In the case of show-your-work items, the steps and answer are one in the same. For show-your-answer items or responses requiring mathematical notation unsupported by a standard keyboard, an equation editor embedded within the item may be used. HCI researchers working on math interfaces, such as Anthony and colleagues (2005), have noted, "Mathematics notations have evolved to aid visual thinking and yet text-based interfaces relying on keyboard-and-mouse input do not take advantage of the natural two-dimensional aspects of math" (p. 1184). In their study of undergraduates asked to handwrite a series of equations or create them with Microsoft Equation Editor, keyboarding equations took nearly three times as long as handwriting and had twice the number of errors (with an error defined as an error left in place or recognized by the subject and corrected) (p. 1186). In this study, variable familiarity was noted—fewer than 5% of participants reported knowing Microsoft Equation Editor "very well," and more than two-thirds had no experience with the software) —but its impact was not analyzed.

Oviatt's 2006 study of handwriting versus computer-based equation editors differs from Anthony et al.'s in its involvement of problem solving (rather than transcription), which permits analysis of the role of cognitive load. Oviatt's approach includes multiple interfaces representing a continuum of similarity to pencil and paper, with keyboard/mouse inputs being the least similar and "digital inking" with a paper-like interface being the most similar. The problems given to the participants "required complex problem solving using linguistic, symbolic, numeric, and diagrammatic representational systems, as well as translation among them" (p. 875). The study's results show a correlation between declining student performance and the interfaces' dissimilarity to pencil and paper, with pencil and paper invariably matching students' existing work practice. For lower performing students, the high cognitive load of unfamiliar interfaces led to a disruption of their performance that was disproportionately greater than with higher performers. This was evident in all performance indices: speed, attentional focus, metacognitive control, correctness of problem solutions, and memory (p. 876).

Although Oviatt's conclusions about the extraneous cognitive load of equation editing interfaces may seem disheartening for proponents of digital assessments interested in machine- or artificial intelligence (AI)-based ways of evaluating robust student input on math tests, two points stand out in terms of fruitful directions for the future. First, Oviatt's phrasing—"the study evaluated whether student performance would deteriorate as interfaces departed more from students' existing work practice" (p. 875)—suggests that the evolution of instructional tools used in math classrooms may change “work practice” and that our assessment instruments should mirror those work practices. Secondly, her use of three interfaces beyond pencil and paper is indicative of the innovation occurring in the space of digital marking tools, innovation that has continued in the years since her study. Advancements in the areas of stylus-based input, graphics tablets, multimodal interfaces, and device arrangements that allow for more seamless movement between inputs are ongoing. Even more notable are advancements in the conversion of handwritten equations into multiple formats (e.g., text, MathML, and LaTeX). Examples include Microsoft Math Recognizer, Ink Equations, MyScript (integrated with Desmos and utilized by Khan Academy since 2015), EquatIO, Mathoix, Photomath, WebMath, and MathType’s handwriting recognition.

Dynamic recognition of handwritten mathematical expression has proven to be complex. Any system must segment an expression into its component parts and recognize discrete symbols organized in a nonlinear format. Structural analysis is deployed to understand relationships between symbols in even the simplest of equations (e.g., begin/end parentheses, numbers above and below fraction bars, superscript). No system is flawless, with each having its own quirks and conventions to be mastered by the user. Functionality for a user to correct inaccurate recognition is a requirement and takes on many forms, from erase and start over to “lasso” the incorrect bit for correction to the popping up of several likely possibilities for a user to choose between. Some of the most advanced technologies in this space become increasingly accurate, adapting to a user’s handwriting over time. With these technologies integrated into commonly used applications and deployed within interactive tutoring systems, we can hope that accuracy will improve and that student familiarity with such tools will grow, making integration into future digital assessments more likely.

Tools, Animation, and Interactivity

The transition from a physical calculator to a digital equivalent does not substantively change the way of interacting with a calculator: a mouse, finger, or stylus is used with a digital calculator while only finger or stylus (or some other type of pointer) would be appropriate for use with a physical calculator. The entirety of a physical calculator’s meaning (minus “where does the battery go?” or “how do I charge this?”) is perceptible in a 2D equivalent. To what degree does this apply to other math tools? Protractors and rulers are like calculators in that their communicative value is adequately captured in two dimensions. However, digitally mediated interactions, such as rotation, are not as immediately intuitive as a calculator button click. Test-takers must have knowledge, through prior exposure, tutorial, or experimentation, of grabbing a corner to rotate versus grabbing the center to drag and place. Difficulties with rotating or pushing the ruler off-screen or a limited rotation area when confined to the visible screen also must be easily avoidable.

Resourcefulness and device comfort may play a role when test-takers opt to use various tools or functions in combination with one another in a way that is not typical with these tools’ physical equivalents. For instance, on a smaller screen, difficulty with precise placement and legibility of tool markings can be addressed by using magnification in combination with the protractor or ruler. When operating system or device magnification is used and when there are differences among devices among the testing population, differences could exist in terms of how easily a tool can be repositioned under magnification.

Not all math tools are so easily translated into the 2D space of digital assessments. For instance, the ease of transitioning between a physical and a digital compass may depend on what type of compass is used in the classroom: a safety compass or the type with the pointy bit, the operation of which is a 3D affair. Some instruction and mastery of the interface is required in either case, as placement, radial adjustments, and rotation with and without "inking" are all required actions for proper compass usage in a digital assessment with such items.

Three-dimensional to 2D tool transitions in science tend to extend beyond a small selection of tools found in mathematics, so there is limited utility to exploring usability concerns and device implications for any individual piece of equipment as realized in a simulation. In addition, NAEP's past use of hands-on tasks (HOTs) limited the need for such translations: Procedural knowledge, equipment handling, and other lab techniques are more faithfully measured in the context of having a test-taker identify metals through magnetic properties or design an electrical circuit using a science kit provided by the assessment facilitator. NAEP's interactive computer tasks (ICTs), on the other hand, provide the opportunity to extend beyond the safety needs, cost implications, and space/time constraints inherent in HOTs (the factors that put future use of HOTs in question). In the case of the ELA and math tasks referenced above, drawing a rough dividing line between task design and interface design has been straightforward. For instance, an item evincing good task design might contain a conceptually and visually clear and relevant image. The ability to enlarge that image, mark on it, or view it with low or high contrast are matters of interface design—capabilities provided by the test delivery system. With ICTs, the demarcation between item and interface is less clear: The task has its own interactivity that must be highly usable, the way the task operates must work well with aspects of the test delivery system's interface, and, as always, content and interactivity must be attuned to work with operating system-enabled functionality and device characteristics.

NAEP has grappled with many aspects of effective interactive task design through research, surveys, observational studies, and task evaluation using Richard Mayer's multimedia design guidelines (Duran et al., 2019). Application of such guidelines, the ability to design simulations and science scenarios with a singular screen size in mind, and the usability testing (and/or cognitive labs) of the task's interactivity on a single device contain some of the complexity. Those factors could remain relatively static moving forward, or future delivery strategies may involve school equipment or use of multiple NAEP-provided devices, such as with the phasing in of Chromebooks. With either of these options, device diversity may require more usability testing on a range of devices. Differences in responsiveness, motion clarity, and audio/video handling across devices require greater attention with ICTs than with less action-packed tasks. In addition, ICTs tend to be designed with greater attention to effective utilization of available screen space than other tasks. With a variety of screen sizes, the integration of responsive design to accommodate the accepted range of screen sizes may be necessary. Last, any future transition of the NAEP platform to cloud-based delivery in lieu of assets preloaded onto devices will have the greatest impact on ICTs with the larger asset sizes of animations, videos, and detailed imagery.

THE PROPOSED FRAMEWORK FOR CONSIDERING DEVICE AND INTERFACE FEATURES

The purpose of this paper is to offer a framework for considering device and interface features that may affect student performance on NAEP. Through our discussion of these considerations, we find that these features or variables are multifaceted, with complex interactions between them. In response, we take a two-pronged approach. First, we identify the most salient device and interface variables. We attempt to "flatten" the complex interactions across variables, organize the variables, and treat each like an isolated feature. In Table 1 that follows, we sort the variables into one of three categories (screens and input devices, test delivery system tools, and interface elements). For each variable, we provide summary information regarding the following:

- The range or measurability of the variable
- Summary of existing research findings
- Level of potential impact
- Subject area or task type impact
- Recommended approaches
- Likelihood that considerations related to the variables will change in the future
- Summary of existing research findings (with a classification of the strength of the evidence: S = strong, M = moderate, W = weak)

If the first step in establishing our framework is disentangling variables and isolating them for the purposes of considering issue-by-issue solutions, then the second step in this two-pronged approach involves reintroducing some of the complexity for a more nuanced perspective. In the section that follows, we explore the multidimensionality of the framework and implications for research directions.

Table 1. A Proposed Framework for Considering Device and Interface Features

Screens and Input Devices						
Variable	**Range/Measurability**	**Summary of Research***	**Impact**	**Interactions and Impact Area**	**Recommendations**	**Forecast for Change**
Screen Size	Easily measured; tablets/laptops range from ~8" to 17".	Suggests screen size > 10" is minimally adequate for tests with NAEP characteristics. **S**	Minor; most devices will meet research-based minimums.	Greater for complex Technology-Enhanced Items (TEIs), tasks involving extensive written input, and tasks requiring access to multiple sources.	Minimum screen size should be established even under "bring your own device."	Devices unlikely to fall below 10" in school environments.
Screen Resolution	Often correlated with screen size but adjustable within a supported range; impacts how much content will render; laptop standard is 1920 by 1080.	Discomfort and performance effects below 1024 by 768. **S**	Minor if designers plan for standard tablet resolution; most devices with appropriate capability for testing have adequate resolution.	Same as above, with special attention to visually rich stimuli, such as detailed maps.	Test items should be evaluated for suitability at the low end of the spectrum (1024 by 768); include the possibility of on-screen tools such as calculators.	Likely that resolutions on small devices will increase or be enhanced by other image clarity functionality; unlikely to see devices with less resolution in future classrooms.
Screen Glare and Adjustability	Built-in antiglare features; adjustability of screen angle measurable in degrees, with minute adjustments possible.	Anecdotal evidence from observational studies suggests issues with tablet stands that limit adjustability or are unstable in some positions. **W**	Easily managed through device guidelines; devices like Chromebooks have been more popular than tablets for classroom purchases.	Possible impact across item types but most risk for text-intensive items, such as reading selections and writing tasks.	Glare-causing lighting conditions are variable and difficult to control, and student heights vary, so screen angle adjustability is a must.	Screen designs likely to improve, diminishing the likelihood of glare or discomfort from screen angle/position.
Styluses	Pressure sensitivity, ability to use a stylus as an eraser, ability to use a stylus for the equivalent of a right click, and hover detection can make styluses more useful.	Observational studies show students experiment with styluses when offered them but generally opt to use more traditional input mechanisms. **S**	Minor, as devices have other input mechanisms with equal or greater precision that can be leveraged based on student choice.	Narrow range of tasks that might benefit from stylus: editing mark-up and drawing (e.g., geometric figures, computer science flowcharts, hash marks).	Offering styluses to users who are unfamiliar with them may be distracting. Students should be allowed to use them if used regularly in the classroom.	Styluses and device support for styluses are increasingly sophisticated. May evolve so that their use feels natural with more functionality than mice.

*Entries under "Summary of Research" include a strength of research evidence designation: **S** = strong, **M** = moderate, **W** = weak.

Table 1. A Proposed Framework for Considering Device and Interface Features (Continued)

Screens and Input Devices						
Variable	**Range/Measurability**	**Summary of Research***	**Impact**	**Interactions and Impact Area**	**Recommendations**	**Likelihood of Change**
Touchscreens	Differences exist in touchscreen technologies: number of simultaneous touches detected, touch resolution, and responsiveness.	Suitability of tasks for a touchscreen and appropriate interface/content design considerations are more critical than the quality of touchscreen technology. **M**	Possibly significant; may introduce construct-irrelevant variance due to finger occlusion, imprecise input, or loss of information provided by system on hover.	Text editing, online calculator use, highlighting text, drawing, scrolling, and interaction with objects smaller than 44 by 44 pixels are more difficult than with a mouse.	Touch input for multiple-choice and drag-and-drop items is acceptable if design is with a touchscreen in mind. Not appropriate as the sole input for more robust assessment tasks.	Likely to evolve, although built-in limits to the precision possible with fingertip input exist.
Keyboard	Full, compact, and virtual. Size of keys, distance between keys, key placement, availability of certain keys, travel distance, resistance, and tactile or haptic feedback can vary with keyboard size and type.	Typing speed may decrease and error rate may increase when using virtual keyboards or the most compact of compact keyboards. **S**	Possibly significant in extensive writing tasks. Limited impact in tests with minimal typing.	Cognitive demands of essay-type responses are best met when typing is automatic and natural due to learned typing skills and a usable keyboard.	Physical keyboards recommended when writing is involved; optimal key size 18-20 mm. Keyboards smaller than 13" may leave off keys and have undesirable characteristics.	New typing styles and alternatives to QWERTY have failed to dislodge traditional styles of typing dating back to 1874, but innovation is possible.
Mouse and Trackpad	Device-integrated input devices (e.g., trackpad or nub) and mice may vary in terms of sensitivity, latency, ergonomics, and precision as well as specialized features such as scroll wheels and programmable buttons.	Mouse found to be more effective than other input devices for object manipulation and text entry. Anecdotal evidence regarding the significance of familiarity. **M**	Minimal impact when high-quality, traditional mice are used and when the input devices used for testing resemble those used in the classroom.	Possible impact for small object manipulation (placing points on a graph) and text editing, which involves cursor placement.	Leveraging familiarity is recommended. Precision of mouse input may be more critical for some item types. Possible to allow use of same mice used in classroom.	Mouse styles are evolving (e.g., more ergonomic designs) but traditional mouse has staying power among new models and attempts to innovate with touchscreens, nubs, and styluses.

*Entries under "Summary of Research" include a strength of research evidence designation: **S** = strong, **M** = moderate, **W** = weak.

Table 1. A Proposed Framework for Considering Device and Interface Features (Continued)

Test Delivery System Tools						
Variable	**Range/Measurability**	**Summary of Research***	**Impact**	**Interactions and Impact Area**	**Recommendations**	**Likelihood of Change**
Digital Markup Tools	The presence and usability of online mark-up tools, such as highlighters, drawing tools (freeform and straight line), and compass, which support thought processes and are utilized for test-taking strategies and possibly for response input.	Usability for some tools achieved by mirroring physical equivalents. For others (e.g., compass, straight-line drawing tool), the most usable digital version may not resemble the physical equivalent. Eraser tool usability often problematic. **M**	Relevant within high cognitive load tasks; can reduce load on working memory.	Reading and geometry tasks may such benefit the most from tools. The design of the tools integrated into the test platform will interact with input devices. Drawing, underlining, or highlighting with finger vs. stylus will differ.	Cognitive labs should be used to ensure that digital markup tools are usable; emulate common learning tools; and include the best range of options (colors, behavior, opacity, thickness). Also ease in erasing or temporarily hiding.	Although the tools themselves may not change radically, students may use such tools increasingly in digitally delivered content. They may become more usable by virtue of familiarity.
Measurement Tools	Presence and usability of tools, such as protractors, rulers, and digital equivalents of tools used for measuring mass, volume, pH, and so on.	Detailed, tool-specific research unavailable, but measurement tasks within online tests are common. **W**	Impact may be significant at the item level, when a task requires measurement but usability issues with the tool use exist.	Science and math. 2D tools are easily translated but ease of ruler or protractor positioning is critical. Use with zoom tool is beneficial.	Usability testing needed with special attention to 3D and tactile measurement equivalents (e.g., graduated cylinder, abacus, triple beam balance, and so on).	Students expected to have greater exposure to science simulations and digital tools; more ease in working with 2D equivalents of 3D instruments.
Calculators	Resemblance between online and physical calculators, size and visibility of buttons, and ability to move the calculator as well as easily hide and restore.	Calculator developers (Texas Instruments [TI], Desmos) involved in maintaining equivalence between test and classroom calculator tools. Anecdotal evidence around usability requirements. **W**	Negative impact has been minimized by use of name-brand calculator simulators offered within many test delivery systems.	Higher level math tasks using graphing calculators are more likely to be impacted. Interactions with small screen size and resolution are due to size and detail of calculators.	Hide/restore, resize, and reposition are key features. Minimum screen sizes and resolution must account for information-rich tasks that require a graphing calculator.	Change is possible as illustrated by Desmos's breaking of TI's near monopoly. Free online calculators are replacing costly physical calculators.

*Entries under "Summary of Research" include a strength of research evidence designation: **S** = strong, **M** = moderate, **W** = weak.

Table 1. A Proposed Framework for Considering Device and Interface Features (Continued)

Test Delivery System Tools						
Variable	**Range/Measurability**	**Summary of Research***	**Impact**	**Interactions and Impact Area**	**Recommendations**	**Likelihood of Change**
Editing Tools	Cut, copy, and paste functions are enabled through multiple means (e.g., keyboard shortcuts, right-click contextual menus, and buttons with familiar icons and informative rollover text).	Literature suggests that editing is a critical part of writing tasks and should be supported through familiar means. **S**	May have significant impact with the increased role of revising within digital writing. Transference of fluidity with cut, copy, and paste is key.	Complex writing tasks impacted. Some interaction with device and input mechanisms in terms of cursor insertion and placement of delete, insert, and backspace keys.	Interaction between multiple variables requires careful attention when designing interfaces and choosing devices and peripherals. This is particularly true for text editing.	Writing behaviors are changing rapidly with the use of a variety of devices in school and at home.
Viewing Flexibility	Flexibility in what is viewable at one time and how intuitive such functionality is. Ability to zoom and adjust position of different content sources and tools (e.g., calculator).	Seeing more content at once can be beneficial to reading and writing tasks. Observational evidence regarding occlusion by immovable calculator. **M**	Variable impact based on task type.	Tasks involving reading, writing, and multiple information sources.	Expandable reading and writing areas. Obvious way to move between information sources.	Strong digital assessment solutions for notecard arranging and "paper shuffling" are emerging and will be available to students.
Reading Interfaces	Markup tools; how much content can be seen at once; ease in navigating to off-screen content, such as in a scrolling or paging interface; ability to change the view to match the task; specialized passage tools such as in NAEP's eReader.	Literature suggests that mode effects have been addressed across time by improving interfaces and tools. Comparability with paper less of a goal as computer-based reading becomes more the norm. **M**	Minimal in tasks without extensive reading requirements but more significant in tasks involving engagement with long passages.	Reading tests and social studies tests with a large reading load. Tasks that may require consulting multiple information sources to respond to an item. See also digital markup tools.	Maximize how much text is viewable at once while maintaining acceptable font sizes and line lengths. Pagination may increase ability to relocate text using visual memory but may have decreased future relevance.	Transition between paginated and scrolling interfaces for presenting passages as web-based reading becomes more common. Move to responsive design may impact line length so maximums should be set.

*Entries under "Summary of Research" include a strength of research evidence designation: **S** = strong, **M** = moderate, **W** = weak.

Table 1. A Proposed Framework for Considering Device and Interface Features (Continued)

Interface Elements						
Variable	**Range/Measurability**	**Summary of Research***	**Impact**	**Interactions and Impact Area**	**Recommendations**	**Likelihood of Change**
Writing Interfaces: Other Elements	How much of a student's writing is visible at once; availability of spell-check, auto-correct, and word suggestion; indications of how much a test-taker has written.	Literature suggests that digital writing involves eye movement across the written text, which guides revision and assists in planning. **S**	Relevant to writing tasks and other subjects that include lengthy written responses.	Interface may interact with task and construct. Weaknesses in writing interfaces are less likely to be an issue with shorter responses.	See as much as possible. Cues as to how much is written. Decisions on inclusion of spell-check, auto-correct, and word suggestion are dependent upon how the construct is defined.	Writing behaviors are increasingly digitally mediated with a variety of interfaces that appear in different contexts, such as online forms and social media.
Scratch Paper and Digital Notepad	Provided automatically or by request; amount of space available; item-/passage-specific; expiring or persisting; ability to draw and write.	Assessment programs vary in scratch paper policies. May reduce load on working memory. **W**	May support prewriting for writing tasks and complex mathematical operations (including freeform drawing).	Interaction with individual preferences likely. Writing and math tasks impacted.	Scratch paper may allow for more authentic approaches when used in the classroom.	Ability to easily use online notepads and drawing areas instead of scratch paper may develop across time as well as voice annotation.
Interface/Task Design–Technology Enhanced Items	Overall usability; familiar interface conventions; authenticity; ability to transfer 3D skills and knowledge to 2D digital rendering.	Visual and interactive features critical to the development of construct-relevant tasks; human-computer interaction principles. **S**	Interface can inhibit or facilitate engagement with the assessment task.	Element interactivity and germane cognitive load are important in design.	Consistency of design across tasks; focus on tutorial design and maximizing interface familiarity.	Assessment possibilities will continue to evolve and expand with improvements in technology and digital assessment design.
Device/Interface Interaction	Usability testing across devices can reveal small issues that are device-, operating system-, or input mechanism-specific.	Anecdotal evidence and gray research reference system-specific changes made to address such issues. **W**	Level of impact can vary.	As discussed above, text editing is an area where input devices, display devices, operating system, and interface operate in a coordinated way.	Usability and quality control testing on a range of devices and use of responsive design; design with touchscreens in mind if such devices are allowed.	May continue to be an area for attention, particularly with the mix of old and new devices that often occurs in the classroom.

*Entries under "Summary of Research" include a strength of research evidence designation: **S** = strong, **M** = moderate, **W** = weak.

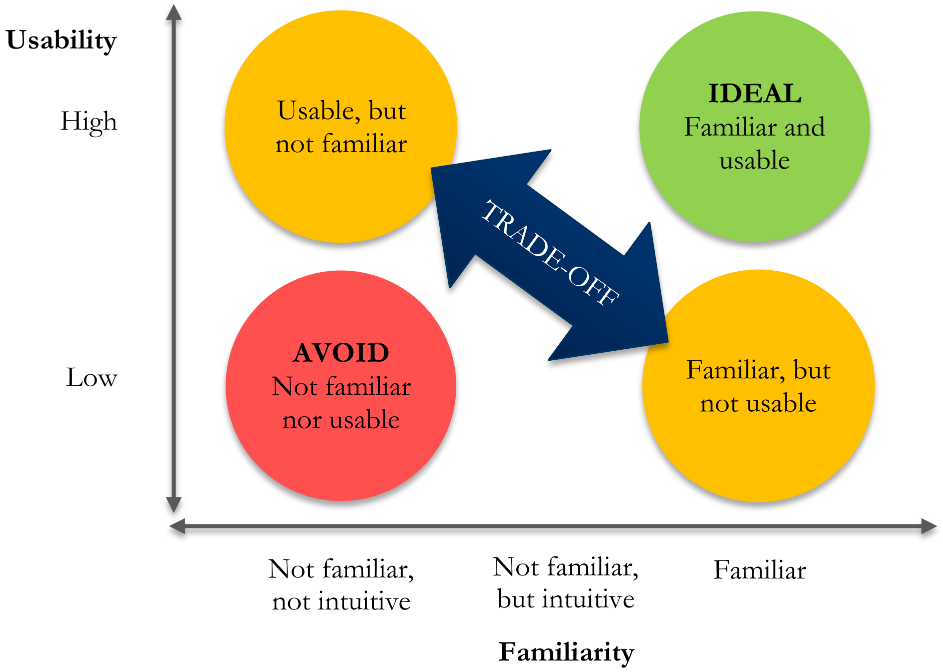
Usability and Familiarity

In constructing the matrix of considerations for device/interface effects, our primary focus was on usability. Usability is undoubtedly a good thing and an important factor to consider, but usability findings and guidelines are often based on a "generic user." That is not to say that user groups who access an application with different goals and sensibilities are not recognized within usability engineering. To the contrary, user personas are often developed based on user research to keep these different user types in mind while design decisions are made. However, as we seek to flesh out the matrix by identifying diverse behavioral characteristics and the differential impact of certain variables within assessment, context and the classroom environment begin to play a central role. As we explore context, usability takes its place as just one contributor to construct validity. An equally important potential contributor to device/interface effects is tool familiarity. How familiar the test-taker is with the tools available within the device and interface acts as a critical triggering mechanism for the test-taker, communicating the opportunity for exhibiting one's skill and knowledge. Thus, the degree to which an assessment can faithfully represent student abilities may be compromised if the tools are unfamiliar.

In a perfect world, usability and device/interface familiarity are not in tension with one another. The most usable interface mechanisms and devices are in common use, and thus are familiar to all students. Realistically speaking, however, when locating an assessment task on a usability continuum and on a familiarity continuum, we find those continuums at best may not be coextensive and at worst may pull in different directions and force difficult decisions on the part of assessment designers. Using keyboard design as an example, one could draw upon existing research to specify preferred key layout and ideal ranges for key sizes, strike distances, haptic/tactile feedback, and required key pressure. These specifications would provide a guide for choosing a keyboard that is highly usable for most test-takers. However, if a student had learned to type and completes all academic writing using a keyboard that is well outside of these usability guidelines, then would he or she perform better on a writing test with a familiar keyboard or a more usable keyboard?

The question to privilege usability or familiarity must be asked more expansively to encompass familiarity with the device, its operating system, other peripherals, and the test-taking interface. Figure 1 depicts the interaction between usability and familiarity along two axes. A student might be expected to perform their best when elements of the device and the interface are highly usable and when the student is familiar with these elements (upper right portion of the figure). Movement away from this zone of optimized performance could happen along either or both axes. A well-designed interface can be described as intuitive, but if it is unfamiliar to a test-taker, some experimentation is inevitably involved in intuiting how an interface feature might work and confirming that assumption. Learning the interface and aspects of working with the device may compete for cognitive resources that would otherwise be applied to the construct. On the other hand, performance-inhibiting usability problems might be familiar to a student. Familiarity may not eliminate their impact, although compensatory strategies may become internalized. (As an example, anyone who grew up with first-generation word processing programs that had the usability problem of occasional crashes may remember how natural "frequent save" operations using keyboard shortcuts became as a safeguard.) Most impactful would be the situation characterized by both unfamiliarity and poor usability: a new interface is not easily learned, and the student has not had time to develop compensatory mechanisms for dealing with certain usability failings.

Figure 1. Interaction Between Usability and Familiarity


Usability
High
Low
Usable, but not familiar
IDEAL Familiar and usable
TRADE-OFF
AVOID Not familiar nor usable
Familiar, but not usable
Not familiar, not intuitive
Not familiar, but intuitive
Familiar
Familiarity

Tool familiarity can be broken down further. Table 2 includes some considerations related to these subcategories of tool familiarity.

Table 2. Tool Familiarity Subcategories

Variable	Considerations
Device Familiarity	Is this exact device, or a device of the same model, used in the classroom? If not identical, is the device used in the classroom similar? What is the level of usage in the classroom? Does home usage of a similar device have an impact?
Peripheral Familiarity	Are similar or identical peripherals (e.g., keyboards, mice, styluses) provided in the classroom?
Interface Familiarity	• *Learning opportunities.* Are tutorials and practice tests provided? To what degree do students choose to use them? • *Prior test-taking experience.* Do students encounter this system routinely? • *Common interface characteristics.* Are the user experience conventions used in the test delivery system common? Do calculators and markup tools resemble what is used in the classroom?
Computer-Based Tasks	Does the learning environment provide opportunities to learn and exercise skills using a computer? When equivalent learning tasks are done without digital tools, how easily can students translate appropriate knowledge and skills to a digitally delivered environment?

Framework Multidimensionality

In exploring the framework as multidimensional, interface/device familiarity is the most notable force that can act alongside or be in tension with usability, as described above. We might ask if other variables are at work within this interplay of factors. In the area of negative interaction between factors, are there other variables that act as potential amplifiers of construct-irrelevant noise? Or, on the positive side, are there other controllable factors that can help compensate for those that are less under assessment developers' direct control?

One such factor deserving of at least a passing mention is motivation. As previously discussed in this paper, because of the lower stakes nature of NAEP, test-takers may be less likely to put considerable effort into a NAEP test compared with a high-stakes state assessment or a test used in college admissions or for assigning course credit. Other than observations that some younger students regard assessment content delivered on popular devices (e.g., iPads) as "fun" and speculation that video, simulations, and TEI functionality may increase engagement, device/interface choices are not generally considered to be primary factors in determining test-taker motivation levels. Nonetheless, we must consider that if less motivated students experience usability issues or are not familiar with the device/interface, impact on performance may be magnified. Table 3 includes certain variables that may impact motivation and some associated considerations.[1]

Table 3. Variables That May Impact Motivation

Variable	Considerations
Individual investment in consequences	Is the assessment associated with consequences that are likely to motivate a test-taker? Will the student receive the results? Will the results be considered for a course grade, course credit, graduation, or acceptance into an academic program?
School/classroom environment	Do schools and teachers introduce the test with certain encouragement around students performing their best? Are other efforts made to encourage student engagement with the test? Do the administration conditions encourage students to rush through the assessment in order to partake in a preferred activity?
Larger sociocultural factors	Do differences exist from year to year regarding public opinion of assessment? Could student motivation be impacted by larger negative views of assessment? Do students view assessments like NAEP differently than they have in the past?

Applying the Framework to NAEP

One use of the framework involves approaching it as a knowledge base. As an issue-by-issue guide to existing research, the framework can be used as a checklist. For instance, an assessment developer creating guidelines regarding eligible devices and peripherals for test-taking may wish to consult with the framework for pertinent variables to address. Similarly, when working through a new assessment design and/or test delivery interface design, the framework can provide a reality check to verify that key issues have been considered.

For NAEP validity studies, the goals for the framework, however, involved ambitions beyond utility as a checklist. For instance, one intention was to reveal gaps in existing research and provide a basis for prioritizing some research agendas over others. If either (1) an analysis of NAEP's items, interface, or the device/peripheral selection suggested a threat to the validity of NAEP, or (2) some variables were revealed to be more impactful than others, then that would naturally lead to conclusions about fruitful research to pursue.

[1] Speaking more generally about NAEP, Dan Koretz argued that changes in motivation are not likely to explain performance trends in recent years because NAEP performance has been so stagnant across time (Supovitz, 2020).

To what degree has the process of developing this framework revealed a natural course of action in terms of research for NAEP? In terms of #1 above, reviewing the usability-focused portion of the framework in relation to NAEP assessments and performing what could be described as a heuristic usability evaluation on a selection of NAEP assessments and items on a Microsoft Surface Pro have led to some good news, or potentially bad news depending on one's point of view. NAEP's design, device, and administration protocol decisions appear sound. Using the framework as a checklist leads to the conclusion that no "smoking gun" is apparent. In other words, no overwhelming design flaw or neglected consideration immediately explains the disparity in writing results across devices found in recent research. NAEP's adherence to best practices with no evident failure to account for a critical aspect of the student test-taking experience is positive, but one or more smoking-gun hypotheses would have been far more convenient in terms of suggesting a series of research studies.

As described below, the multidimensionality and interplay between variables explored in the framework are particularly relevant to NAEP, and, as it turns out, the multidimensionality complicates an attempt to easily rank variables according to their impact in terms of device/interface effects. In this way, the framework fails to deliver on #2 listed above, but it is here where the framework does point to possible research areas of emphasis.

RESEARCH DIRECTIONS FOR NAEP: DEVICE/INTERFACE FAMILIARITY

Returning to the idea that the development of this framework should involve a discovery process that both highlights how the dimensions within the framework can compete with one another and points to research gaps, we posit that this potential tension is not well understood. The relative benefit of usability versus familiarity as depicted in Figure 1 constitutes a research gap that seems particularly relevant to NAEP. More than most assessment developers, NAEP can fine-tune the usability of the current eNAEP test delivery platform through specifying the exact device, peripherals, and screen size that will be in use. Because of the policy of supplying the devices used for administration (currently Microsoft Surface Pros), NAEP has an advantage over most assessment developers that must account for a wide range of variability in content layout on different screen sizes/resolutions and that must evaluate their test delivery platforms with many devices and operating systems, and hope for at least a baseline of usability across devices. However, if familiarity is as important as usability, then NAEP's current policy of provisioning devices could put the program at a disadvantage in comparison to a test delivery strategy that leverages school-provided equipment that also is in regular use by students throughout the year.

Based on information provided by NCES, NAEP's future device plans involve two strands of development. First, there is an intention to migrate over time from Microsoft Surface Pros to Google Chromebooks as the device provided for NAEP administrations. Second, a next-generation version of the eNAEP delivery system is under development. The system will support cloud-based assessment and is being developed in a way that will allow it to be device-agnostic. The system has already been adapted to support Chromebooks and is being designed to provide a common display regardless of device, so as to minimize the variation in user experience across devices. Most of the eNAEP system changes are at the back end and will not be noticeable to users. The next-generation version is targeted for field study in 2022 and full implementation in 2023.

The motivation for these transitions is partly related to updating the inventory of devices used for NAEP assessments. The program already supports the simultaneous use of three different Surface Pro iterations with no evidence of differences in system performance across versions.

Because of the volume of devices involved, the transition to Chromebooks will be accomplished in waves and will be dependent on funding, which is currently uncertain. Under the best-case scenario, Chromebooks could be introduced over three phases, replacing about one-third of the Surface Pros in the field during each wave. For example, transitions might begin with the 2023 assessments but might not be completed until the 2027 assessments.[2]

Because NAEP's approach to delivery system development should support a similar level of usability as Chromebooks begin replacing Surface Pros, the more salient concern for NAEP as it relates to device effects may be whether differences in device familiarity of students with Chromebooks versus Surface Pros will impact performance. The case for worrying about

[2] NAEP assumes only trend assessments in 2024 and no assessments in 2026.

familiarity is heightened by the fact that currently Chromebooks are by far the dominant device used in the schools, and the use of Surface Pros is relatively rare (EdWeek, 2017).

A second phase of research and development under consideration is to investigate the potential use of school-provided equipment for NAEP delivery. There is not yet a target date for this strand, as schedules and priority will depend upon internal policy decisions and the continued evolution of devices in the schools. This phase would anticipate a potential hybrid model where both assessment using NAEP-supplied devices and assessment using school-provided devices might occur simultaneously. This phase would require focused school-based studies, building on the recent school-based equipment proof-of-concept study (NAEP Alliance, 2018) as well as developing the supporting logistics. If this phase of research and development is prioritized, a more explicit examination of trade-offs between usability and familiarity may be even more important to undertake.

Within this backdrop of NAEPs device transition plans, some ideas about research and data collection to further understanding about the impact of device and interface familiarity on NAEP performance are discussed below. The discussion addresses considerations for bridge studies that might be undertaken as Chromebooks are transitioned into the field as well as auxiliary studies that might support future plans.

Research Options to Support the Transition of Device Types

Traditionally, when NAEP has implemented instrument or administration changes, bridge studies have been used to permit comparisons to be made across years. In general, NAEP bridge studies involve randomly equivalent samples receiving the old and new administration formats. This common population-linking design is employed because using items in common between administration formats as a basis for linking is an untenable assumption. Bridge studies generally support point-in-time transitions in NAEP administration procedures. For example, in 1988, the administration of NAEP assessments was changed to assess each student in only one subject area (Johnson & Zwick, 1990). In 2004, the NAEP long-term trend assessments were modified to reorganize and standardize the assessment booklet design to a more common structure used in other NAEP assessments (Perie et al., 2005). More recently, the introduction of digitally based assessment (DBAs) in mathematics and reading in 2017 involved an extensive bridge study to evaluate the effect of the mode of administration on performance, and permitted comparisons of the 2017 results to later assessments administered digitally as well as to the earlier assessments administered on paper (Jewsbury et al., 2020).

With respect to the transition of NAEP digital devices from Surface Pros to Chromebooks, a natural strategy would be to implement bridge studies. However, in this case considerations for bridge studies become complicated. The first complication is the expectation that the device transition will take place over a number of years. Thus, a bridge study design for a given content area might have to support adjustments over more than one NAEP administration. Suppose a first wave of Chromebooks is initially deployed in 2023 and replaces approximately one-third of the existing Surface Pro devices. Although a bridge study similar to the study done in 2017 could be done for reading and mathematics, the study also would have to support a mixed-device assessment in 2025 when perhaps another one-third of Surface Pro devices were replaced with Chromebooks. This might be feasible but would likely have design implications. Bridge studies for other content areas (e.g., science, civics, U. S. history, and digital writing) also might be necessary, and the bridge studies would involve unique features depending on the year these content areas are assessed and the state of the device transition effort.

The 2017 bridge study to support the digital transition of reading and mathematics was extensive. It required drawing a full sample for the DBA similar to past paper assessments and an additional sample for the paper-based administration (PBA), increasing overall sample sizes by approximately 27%. In addition, the student sampling process consisted of two parts. The first part was the within-school student sample selection, and the second part was the assignment to the assessment mode of the selected students. Participating schools therefore had to support the administration of both DBAs and PBAs, which added to scheduling and logistics burdens. A decision to incur the costs and administrative challenges of a similar bridge study to support device transition is not one to be made without careful consideration.

Prior to the introduction of the reading and mathematics DBAs, NAEP had done field trials in 2015, which clearly indicated that digital versions of both mathematics and reading items were more difficult than the original paper format items. Thus, research evidence clearly indicated the need for the bridge study conducted in 2017. That situation can be contrasted with the current situation regarding expectations about performance on NAEP reading and mathematics DBAs administered using Chromebooks rather than Surface Pros. Currently, there is no evidence to suggest that—for these subject areas—administering NAEP assessments on Chromebooks will result in any performance differences compared with administering them on Surface Pros. Moreover, the next-generation eNAEP system under development is being designed to provide a common display regardless of the device used, which should minimize any possible effects that might be attributed to the use of different operating systems in the different devices.

In this paper, we have called out the tensions between usability and device familiarity, and we have hypothesized situations where trade-offs among these variables could impact student test performance. One might argue that a transition to Chromebooks for NAEP delivery might be introducing a device more familiar to students, as Chromebooks are used far more extensively in schools than Surface Pros (or any other device, for that matter). On the other hand, there are a variety of Chromebook manufacturers and although Chromebooks by definition share a common operating system, they can vary considerably in terms of the screen and input device characteristics outlined in our proposed framework. NAEP can address these potential effects by purchasing Chromebooks that have screen and input device characteristics that are as similar to Surface Pros as possible.

NAEP is therefore in a strong position to minimize any potential device and interface features associated with a shift from Surface Pros to Chromebooks that might affect student performance. Still, recent experiences with the NAEP digital writing assessment make it risky to simply assume that the inevitable device transitions NAEP will be making in the coming years will have no effect on assessment performance. What follows are some research options that NAEP may wish to consider over the course of device transitions.

Option 1: Additional Questionnaire Items. A starting place for gathering data related to device transition considerations would be through additional questionnaire items. NAEP surveys to date have not directly asked students about specific device familiarity but rather asked more general questions about computer and device use. Below, are two examples of questions that might be added to the NAEP student questionnaire:

How familiar to you is this device being used to administer your test?

A. Very familiar
B. Somewhat familiar
C. Not so familiar
D. Not at all familiar

How easy was it for you to take your test with this device?

A. Very easy
B. Somewhat easy
C. Not so easy
D. Not at all easy

Other possible questions might ask about the way items in particular content areas were presented or about specific features of the system interface or test delivery tools.

More pointed questionnaire items about device familiarity and ease of use could be added to NAEP assessments relatively quickly, perhaps as soon as 2022. These additional questionnaire data would not necessarily address specific research hypotheses but rather would establish baseline trend data that could be tracked through the device transition period. In addition, for students responding negatively to these questions, associative patterns might be searched for in process data. For example, for a student indicating that the device made it difficult for them to take the test, could we identify in the process data evidence that the student was indeed struggling with the interface? Unproductive mouse clicks? Amount of time spent on an interactive item before interacting with it when not correlated with reading load? Typing speed within writing tests?

Option 2: Cognitive Labs. A second research option would involve cognitive laboratories designed to contrast the Surface Pros and Chromebooks to be used for NAEP administrations. A NAEP mini-test could be assembled using items that involve the most engagement in terms of device/peripheral use: writing, science simulations, field test science SBTs, and other reading and math items with less typical interactivity. The revised eNAEP platform would be used to deliver these tests on the NAEP-provided devices (e.g., Surface Pros or Chromebooks). Content could either be crossed, so that students experienced some items on Surface Pros and others on Chromebooks, or students could be randomly assigned to use one or the other device to experience all of the items included. Ideally, students would be recruited for participation based on varying experiences with Chromebooks versus Surface Pros. If this was not practical (because of the low incidence of Surface Pro users), targeting students with varying experience with using devices in school and at home might provide the basis for contrasting experiences with the two devices.

In such a cognitive lab, an observational rubric would be used to describe the ease with which the subjects engage with the interactivity. The amount of time spent on an exploratory activity and establishing preference should be noted by the observer, and contrasted across the two devices. Obviously, differential impacts on performance would be difficult to infer, but a cognitive lab might reveal whether or not different types of students engaged in device/interface learning behaviors differently across the two device types. Findings could lend confidence to the assumption that the two device types are interchangeable or reinforce the need for subsequent quantitative studies intended to understand the extend of the performance differences by device. In addition to any usability difficulties being noted (and ideally captured via screen-capture software), a survey would be given to students to rate ease of use and the similarity of the Surface Pros and/or Chromebooks to devices they use at home and/or in the classroom. An ideal time for conducting the cognitive labs might be late 2021 or early 2022.

Option 3: Limited Pilot Study. It might be possible for cognitive labs to provide enough confidence about the comparability of performance between Surface Pros and Chromebooks, but, if not, it might be worth conducting a limited pilot study in 2022 to compare performance on existing Surface Pros and newly acquired Chromebooks. Although there is no other NAEP testing in 2022 to hook into, conducting a pilot study could provide the evidence needed to assume comparability of performance across devices, which would allow device transition plans to move forward beginning in 2023 with no need for more elaborate and expensive bridge studies.

One challenge for the pilot study would be to determine which content areas to include. The NAEP assessment schedule shows reading and mathematics to be administered in 2023, 2025, 2027, and 2029; science in 2023 and 2027; and civics and U.S. history in 2025 and 2029. Writing will not be administered until 2029, presumably after the device transition is complete. However, based on the recent digital writing results, there is perhaps more reason to be concerned about writing than the other subjects. One possible strategy, similar to what was recommended for the cognitive labs, would be to assemble samples of items from multiple content areas, making sure to include items that involved significant engagement in terms of device/peripheral use.

The study could focus on Grades 4 and 8 as Grade 12 is assessed less frequently. Assumptions for the study might be as follows:

1. For each assessment and grade level, participating students would take two blocks of content, plus a background questionnaire. The number and composition of the blocks across all of the possible content areas to be studied would need to be worked out.
2. Consistent with the typical NAEP two-stage sampling design, a representative sample of schools with respect to demographics and geographic location will first be selected.
3. Within each selected school, a fixed number of students would then be sampled to participate in the study. Half of the students would be randomly assigned to test using Surface Pros and half using Chromebooks.
4. The background questionnaire would include items to assess demographics, more general computer and device familiarity, and specific items about ease of use and device familiarity, such as those outlined in earlier in this section.
5. Performance comparisons would focus on item-level performance and relationships between performance and background characteristics.

Results of the pilot study would ideally be available in time to inform whether more elaborate bridge studies would be needed.

Option 4: Bridge Studies. Our understanding of the NAEP device transition plans suggest that the first introduction of Chromebooks will be in 2023, with perhaps one-third of the overall device inventory being available around that time. If results of the pilot studies indicate a need for bridge studies, we assume the first bridge study can occur with the 2023 assessment, which will include reading, mathematics, and science.[3] One way to approach this device-transition bridge study would be to model the approach used in the 2017 bridge study to support the implementation of digital reading and mathematics assessments (Jewsbury et al., 2020). To do this, however, we would suggest that a full complement of Surface Pros would need to be utilized with a standard NAEP sample to support results with appropriate levels of precision. Chromebooks also would be utilized to augment the sample and to support the evaluation of comparability and, if necessary, the linking of assessment results based on Chromebook administration to results based on Surface Pro administration. In the 2017 bridge study, the new administration conditions (DBAs) were administered to the full sample, and the old administration conditions (PBAs) augmented the sample. Unlike in 2017 where there was reason to report results based on the DBAs, it would not seem to matter whether 2023 reported results are based on Surface Pro or Chromebook administration, so this difference in linking direction would not seem to be of concern.

If the 2023 bridge study for reading, mathematics, and science indicates comparability across devices, it might be justifiable to generalize these findings to civics and U. S. history, and assume that results of NAEP administrations using Chromebooks and Surface Pros are interchangeable, such that no additional bridge studies would be needed. However, if the data suggest the need for bridge studies as part of the 2025 NAEP administration, some decisions will have to be made about how Chromebooks will be deployed. A phased transition approach would suggest an additional infusion of Chromebooks in 2025, perhaps an additional one-third of the overall NAEP administration inventory. But if a bridge study is needed in 2025, it might be worth considering whether that infusion should be delayed so that the 2023 approach to comparability and linking could essentially be repeated using similar proportions of device types. Another option would be to control the sampling so that reading and mathematics could be administered using only Chromebooks, and civics and U.S. history would be administered with an appropriate mix of Surface Pros and Chromebooks to permit an assessment of comparability and, if necessary, an estimate of the linking relationships.

By 2027, plans suggest the transition to Chromebooks would be complete, and the bridge study results from 2023 could be applied to reading, mathematics, and science results. In 2029, results for reading, mathematics, civics, and U.S. history will presumably have been bridged across devices. However, it may be worth considering an additional bridge study for writing in 2029, where the majority of students would test using Chromebooks, but an augmented sample would test using Surface Pros to permit comparability comparisons and possibly linking of results to take into account any device effects that were detected.

[3] Technology and Engineering Literacy (TEL) also will be administered in 2023 at Grade 8, but our understanding is that laptops are used for TEL assessments and therefore are not relevant to considerations of Surface Pro versus Chromebook administration.

SUMMARY AND CONCLUSIONS

In this white paper, we proposed a framework for considering device and interface features that may affect student performance on digital NAEP assessments. To develop the framework, we first examined research literature on the comparability of assessments delivered by computer versus paper as well as literature addressing the comparability of delivery across different digital devices (e.g., computer, laptop, Chromebook, tablet). Although relevant to future NAEP digital assessments, this latter literature is sparse and ambiguous. Thus, we delved further into interdisciplinary research on cognitive load theory (CLT) and human-computer interactions (HCIs), and theoretical bases of assessment related to construct validity and replicability. From these foundations, we developed an approach to device and interface effects rooted in an understanding of the cognitive and sensorimotor demands of a task. We applied that perspective to research related to specific NAEP assessment content areas, first to writing and the close connections between writing and reading, and then to considerations for nonwriting content areas, such as mathematics, social studies, science, and interactive computer tasks (ICTs).

From this foray into the research and theoretical connections between device/interface features and assessment content, we proposed the framework presented in Table 1 that condenses a wide number of variables across the categories of screens and input devices, test delivery system tools, and interface elements. Primarily focused on usability, the framework can serve as a checklist of sorts but also can be applied to current NAEP assessments and the current NAEP delivery interface to consider where gaps in existing research might lie and/or what research might be done to best illuminate potential causes of device or interface effects.

Ultimately, our attempts to uncover such effects were not rewarded: there was no apparent smoking gun to explain NAEP device effects in writing and no magic bullet to help ensure that such effects will not occur in the future. We explored two additional considerations relevant to device and interface effects. One is the tension between usability and familiarity: the idea that familiarity with an interface can overcome weaknesses in usability features and that even a highly usable interface may negatively impact performance if important features are unfamiliar to the student. A second is the potential for device effects to be introduced or exacerbated if students are not motivated to give their best effort. Both of these considerations might be further explored in future research.

In the final section of this paper, we put forth suggestions for research and data collections to further understanding about the impact of device and interface familiarity on performance as NAEP transitions to a next-generation delivery system and pursues plans to replace the current Microsoft Surface Pro devices with Chromebooks. We ordered this discussion by first proposing simpler, low-cost efforts, such as additional questionnaire items and cognitive labs, and then describing more elaborate (and expensive) research and bridge studies that might be undertaken. This array of future work involves trade-offs between confidence in preserving the NAEP scale and the cost and complexity of necessary studies.

Recognizing the impact of technology evolution on NAEP and further uncertainties introduced by unprecedented recent events, we hope that the literature review, framework, and research suggestions will prove useful as the NAEP program moves into the future.

POSTSCRIPT

While wrapping up this project, the appearance of a novel coronavirus shifted from a distant phenomenon to what looked to be a short-term dilemma to an ongoing crisis with grave economic and national health consequences. Attempts to manage the pandemic and keep students safe have radically reshaped schooling, including assessment. Most states cancelled spring 2020 high-stakes testing, the College Board's Advanced Placement tests were rapidly transitioned to online tests, and ACT and SAT testing was postponed and reduced. It is becoming apparent that these events are prompting some lasting changes with possible implications for NAEP and NAEP's digital testing strategy. For this reason, we have added this postscript to address these implications and pose some questions that may only be answerable in hindsight.

In the immediate term, it has become necessary to adjust NAEP assessment plans for 2021. Due to COVID-19 safety requirements and a lack of additional contingent funding, NAEP is, at the time of writing, planning on implementing a thin sample alternative, with smaller state samples and no Trial Urban District Assessments (TUDAs). There is no guarantee of national estimates for the assessment because it is not yet clear what level of participation will be obtained, but there will be state estimates where participation allows. These plans may shift again, depending on the state of schools by January 2021 and who is willing to participate.

Despite this changing landscape, or possibly in part because of it, the purpose of "the nation's report card" remains clear. With the disruptions, distractions, and rapid changes in course for teachers and students who swapped their notebooks and whiteboards for online learning tools, we will want to know how we fared. Where did we succeed, and where does the year-to-year trend reveal some insurmountable challenges in keeping education on track and students focused amidst a pandemic?

Although the full range of equity issues within testing was outside the scope of this paper, discussion of the consequences of the 2020 pandemic provides an opportunity to touch upon this topic. One difficult-to-prove/disprove equity concern arises from the possibility that inexperience with digital devices hinders student performance on digital assessments. The presence of computers and digital devices in student households and, by extension, the increased opportunity to develop computer proficiency, are often correlated with higher socioeconomic statuses. This question of who has access to a digital device at home or through their school became critical with the shift to remote learning in response to state lockdowns and attempts to socially distance starting in March 2020. "Digital divide" issues in education have never disappeared, but a new spotlight was shone on them as questions were asked about whose parents were essential workers and unable to monitor at-home schooling, who had room for each child to have a quiet learning space, and, of course, who had the appropriate technology at hand. The resulting mobilization effort to support student learning from home can be considered a mixed bag from an equity-in-testing perspective.

On one hand, some districts report that large device purchases this spring and summer have put them years ahead within their projected schedules for achieving one-to-one device-to-student ratios. However, devices are only part of the solution; some students' lack of home internet access was once characterized as a "homework gap." With remote learning, this gap is not confined to homework and thus is less easily ignored. Relying on the $13.2 billion in emergency K–12 funding from the CARES Act, past or future bonds, and/or community-based philanthropy, districts have purchased devices and mobile hot-spots, in addition to equipping school buses with Wi-Fi to be parked near apartment buildings and housing complexes with large numbers of students. If expanding device availability and internet access for the purposes of academic and nonacademic tasks helps to bridge the digital divide, then this purchasing frenzy may be a net gain.

On the other hand, numerous aspects of the current education environment may exacerbate inequity. Despite the hundreds of millions of dollars being spent on Chromebooks and iPads ($100 million in Los Angeles Unified District alone), some needs inevitably remain unmet, and further stimulus funding for education was stalled as of the start of the 2020–21 academic year. With Zoom-based classrooms and homework turned in online, the consequences of the digital divide—even if narrowed—have become more impactful. Attempts to level the playing field with a learning-optimized physical environment, social services offered by school nurses and counselors, and afterschool care now take a backseat to the logistical issues of distributing devices, hot spots, and free lunches. However, only a few dimensions of the at-home learning experience can be improved through funding and policy.

From a digital assessment perspective, a shift from pencil-and-paper quizzes and tests administered in the classroom to more online testing appears likely for states and districts leveraging remote learning and online teaching tools while the nation awaits a COVID-19 vaccine. Whether the increased familiarity with online testing's advantages during this time encourages its use in a postpandemic nation remains to be seen. If so, then similarity between classroom testing and large-scale assessment delivery, in addition to greater access to and comfort with digital devices, might be advantageous. Similarly, we might anticipate increased use of computer-based interactive science labs due to either remote learning or health concerns related to the equipment sharing and close physical proximity required by small-group lab work. To break up "talking head" instruction with other types of activities, leverage students' internet access, and assess student skills in an open-book way, teachers may assign problem-solving tasks involving web-based research. If this is the case and if such changes in instructional strategies have a lasting impact, then greater alignment between classroom tasks and two of NAEP's assessment strategies—interactive computer tasks and Technology and Engineering Literacy tests—could be one outcome.

Whether enduring shifts in instructional strategies have a positive impact on NAEP validity is a longer term question. Meanwhile, two more immediate and less amorphous questions related to NAEP's digital testing strategy arise from these recent events. One question is whether the second phase of research and development to investigate the potential use of school-provided equipment for NAEP delivery (alluded to earlier in this paper) should be pursued more aggressively. A second question is whether the phase-in of NAEP-purchased Chromebooks over several years of assessments still makes sense. Considering these questions suggests the need for focused research that either piggybacks on a NAEP administration or is designed and carried out separately from a NAEP administration. Specifically, this research would evaluate and establish the comparability of NAEP assessments administered on school-provided equipment. A major concern with moving in this direction is obviously budgetary, but if NAEP is able to defer and ultimately abandon the costs of purchasing and maintaining devices to support annual assessments, a convincing return on investment case for funding the needed research might be made. In particular, as there are currently no NAEP assessments scheduled for 2022, research targeted for this year might make particular sense.

REFERENCES

Adler, M. J., & van Doren, C. (1972). *How to read a book*. Simon and Schuster.

Alvès, R. A., Castro, S. L., de Sousa, L., & Strömqvist, S. (2007). Influence of typing skill on pause-execution cycles in written composition. In M. Torrance, L. Van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications. Studies in Writing* (Vol. 20, pp. 55–65). Kluwer.

Anthony, L., Yang, J., & Koedinger, K. (2005). Evaluation of multimodal input for entering mathematical equations on the computer. In *Conference on Human Factors in Computing Systems Proceedings* (pp. 1184–1187). https://doi.org/10.1145/1056808.1056872

Arms, V. M. (1983). The computer and the process of composition. *Pipeline*, *8*, 16–18.

Backes, B., & Cowen, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Educatio*n, *68*, 89–103. https://doi.org/10.1016/j.econedurev.2018.12.007

Barkaoui, K. (2013). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241–259. https://doi.org/10.1177/0265532213509810

Berninger, V. W., Cartwright, A. C., Yates, C. M., Swanson, H. L., & Abbott, R. D. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades: Shared and unique functional systems. *Reading and Writing*, *6*, 161–196.

Blume, H. (2013). L.A. school board OKs $30 million for Apple iPads. *Los Angeles Times.* http://articles.latimes.com/2013/jun/18/local/la-me-ln-lausd-chooses-ipads-for-pilot-20130618

Bourdin, B., & Fayol, M. (1994). Is written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology, 29*(5), 591–620. https://doi.org/10.1080/00207599408248175

Bracey, G. W. (1996). Altering the motivation in testing. *Phi Delta Kappa*, *78*(3), 251–252. http://www.jstor.org/stable/20405760.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*(3)*,* 191–205. https://doi.org/10.1207/S15324818AME1603_2

Chaparro, B. S., Phan, M. H., Siu, C. Y., & Jardina, J. R. (2014). User performance and satisfaction of tablet physical keyboards. *Journal of Usability Studies*, *9*(2), 70–80.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Crockett, Z. (2019, September 22) Is the era of the $100+ graphing calculator coming to an end? *The Hustle*. https://thehustle.co/graphing-calculators-expensive/

Davis, L. L., Kong, X., McBride, Y., & Morrison, K. (2016). Device comparability of tables and computers for assessment purposes. *Applied Measurement in Education*, *30*(1), 16–26. https://doi.org/10.1080/08957347.2016.1243538

Davis, L. L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment, 20*(3), 180–198. https://doi.org/10.1080/10627197.2015.1061426

Davis, L. L., & Strain-Seymour, E. (2013a, June). *Digital devices research*. Paper presented at the CCSSO National Conference on Student Assessment, National Harbor, MD.

Davis, L. L., & Strain-Seymour, E. (2013b). *Keyboard interactions for tablet assessments*. Pearson.

Davis, L. L., Strain-Seymour, E., & Gay, H. (2013). *Testing on tablets: Part II of a series of usability studies on the use of tablets for K-12 assessment programs*. Pearson. https://docplayer.net/19176443-Testing-on-tablets-part-ii-of-a-series-of-usability-studies-on-the-use-of-tablets-for-k-12-assessment-programs.html.

Deane, P. (2011). *Writing assessment and cognition* (Research Report No. RR-11-14). Educational Testing Service. http://doi.org/10.1002/j.2333-8504.2011.tb02250.x

Deane, P., Sabatini, J. P., & O'Reilly, T. (2011, December). *English language arts literacy framework*. Educational Testing Service.

Debue, N., & van de Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology, 5,* Article 1099. https://doi.org/10.3389/fpsyg.2014.01099

DePascale, C., Dadey, N., & Lyons, S. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, *31*(1), 30–50. https://doi.org/10.1080/08957347.2017.1391262

De Smet, M., Leijten, M., & Van Waes, L. (2018). Exploring the process of reading during writing using eye tracking and keystroke logging. *Written Communication: An International Quarterly of Research, Theory, and Application*, *35*(4), 411–447. https://doi.org/10.1177/0741088318788070

Duran, R., Zhang, T., Sañosa, D., & Stancavage, F. (2019). *Effects of visual representations and associated interactive features on student performance on National Assessment of Educational Progress science scenario-based tasks*. American Institutes for Research, NAEP Validity Studies Panel.

EdWeek. (2017). *Market brief: Amazon, Apple, Google and Microsoft: How 4 tech titans are reshaping the ed-tech landscape*. https://marketbrief.edweek.org/wp-content/uploads/2017/05/Edweek-Market-Brief-Tech-Titans-Research-Report.pdf

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychology Review, 102*(2), 211–245. https://doi.org/10.1037/0033-295x.102.2.211

Eskenazi, M. A., & Folk, J. R. (2017). Regressions during reading: The cost depends on the cause. *Psychonomic Bulletin and Review*, *24*(4), 1211–1216. https://doi.org/10.3758/s13423-016-1200-9

Fishbein, B. (2018). *Preserving 20 years of TIMSS trend measurements: Early stages in the transition to the eTIMSS assessment*. Doctoral dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Boston College. https://dlib.bc.edu/islandora/object/bc-ir%3A107927

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*(6) 381–391. https://doi.org/10.1037/h0055392

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Goldberg, A., Russell, M., Cook, A., & Russell, E. M. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment*, *2*, 2–51.

Graham, S. (1999). The role of text production skills in writing development: A special issue—I. *Learning Disability Quarterly, 22*(2), 75–77. https://doi.org/10.2307/1511267

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, *104*(4), 879–896. https://doi.org/10.1037/a0029185

Gunawardena, W. (2013). *Relationship of hand size and keyboard size to typing performance metrics* (Unpublished doctoral dissertation). Ohio University, Athens, OH.

Hayes, J. (1996). A new framework for understanding cognition and affect in writing. In R. Indrisanio & J. Squire (Eds.), *Perspectives on writing: Research, theory, and practice* (pp. 6–44). International Reading Association.

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.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human-computer interaction. *Computers in Human Behavior, 26*(6), 1278–1288.

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

Hoyle, W., Bartha, M., Harper, C., & Peres, S. (2013). Low profile keyboard design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1348–1352.

Inhoff, A. W., Kim, A., & Radach, R. (2019). Regressions during reading. *Vision*, *3*(3), Article 35. https://doi.org/10.3390/vision3030035

Jakwerth, P. R., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions*. Report commissioned by the NAEP Validity Studies Panel. American Institutes for Research.

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., & Rust, K. (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study*. (White paper). Educational Testing Service.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report* (Report No. 19-TR-20, pp. 3–9). Educational Testing Service.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. http://doi.org/10.1037/0033-295X.87.4.329

Kalyuga, S. (2010). Schema acquisition and sources of cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 48–64). Cambridge University Press. http://doi.org/10.1017/CBO9780511844744.005

Kane, M. T., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge. https://doi.org/10.4324/9781315708591-2

Keng, L., Kong, X. J., & Bleil, B. (2011, April). *Does size matter? A study on the use of netbooks in K-12 assessment*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kim, J., Aulck, L., Bartha, M., Harper, C., & Johnson, P. (2014). Differences in typing forces, muscle activity, comfort, and typing performance among virtual, notebook, and desktop keyboards. *Applied Ergonomics*, *45*(6), 1406–1413. https://doi.org/10.1016/j.apergo.2014.04.001

Kim, J., Aulck, L., Thamsuwan, O., Bartha, M., & Johnson, P. (2013). The effects of virtual keyboard key sizes on typing productivity and physical exposures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 887–891. https://doi.org/10.1177/1541931213571193

Kiplinger, V. L., & Linn, R. L. (1995). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment*, *3*(2), 111–133. https://doi.org/10.1207/s15326977ea0302_1

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS ONE*, *13*(9), Article e0203629. https://doi.org/10.1371/journal.pone.0203629

Kruger, J., Hefer, E., & Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: Eye-tracking and dynamic audiovisual texts. *Proceedings of the 2013 Conference on Eye Tracking South Africa*, *62,* 29–31. https://doi.org/10.1145/2509315.2509331

Larres, P. M., Ballantine, J., & Whittington, M. (2003). Evaluating the validity of self-assessment: Measuring computer literacy among entry-level undergraduates within accounting degree programmes at two UK universities. *Accounting Education, 12*(2), 97–112. https://doi.org/10.1080/0963928032000091729

Li, D. L., Yi, Q., & Harris, D. (2016). *Evidence for paper and online ACT comparability: Spring 2014 and 2015 comparability studies* (ACT Working Paper 2016-02). https://www.act.org/content/dam/act/unsecured/documents/Working-Paper-2016-02-Evidence-for-Paper-and-Online-ACT-Comparability.pdf

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human–Computer Interaction*, *7*(1)*,* 91–139. https://doi.org/10.1207/s15327051hci0701_3

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. http://doi.org/10.1017/CBO9780511811678

McFarland, M. (2017, May 12). Can we finally retire the overpriced TI-84 calculator? *CNN Business*. https://money.cnn.com/2017/05/12/technology/ti-84-graphing-calculator/index.html

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Messick, S., Beaton, A., & Lord, F. (1983). *A new design for a new era* (NAEP Report 83-l). Educational Testing Service.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* Routledge. https://doi.org/10.1111/jedm.12255

NAEP Alliance. (2018). *NAEP 2018 VA POC lessons learned*. Unpublished presentation.

National Center for Education Statistics. (2003). *NAEP validity studies: An agenda for NAEP validity research* (Working Paper No. 2003-07). U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/pubs2003/200307.pdf

National Center for Education Statistics. (2019). *Technical summary of preliminary analyses of NAEP 2017 writing assessments.* U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/subject/writing/pdf/2017_writing_technical_summary.pdf

Nielsen, J. (1993). *Usability engineering.* Academic Press.

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*(2), 135–157. https://doi.org/10.1207/s15326977ea0302_2

Oviatt, S. (2006). Human-centered design meets cognitive load theory: Designing interfaces that help people think. *Proceedings of the 14th ACM international conference on multimedia,* 871–880. https://doi.org/10.1145/1180639.1180831

Pereira, A., Hsieh, C., Laroche, C., & Rempel, D. (2014). The effect of keyboard key spacing on typing speed, error, usability, and biomechanics: Part 2. *Human Factors*, *56*(4), 752–759. https://doi.org10.1177/0018720813502524

Pereira, A., Lee, D., Sadeeshkumar, H., Laroche, C., Odell, D., & Rempel, D. (2013). The effect of keyboard key spacing on typing speed, error, usability, and biomechanics: Part 1. *Human Factors*, *55*(3), 557–566. https://doi.org/10.1177/0018720812465005

Perie, M., Moran, R., & Lutkus, A. D. (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (NCES 2005–464). U.S. Government Printing Office.

Piolat, A., Roussey, J. Y., Olive, T., & Amada, M. (2004). Processing time and cognitive effort in revision: Effects of error type and of working memory capacity. In L. Alla, L. Chanquoy., & P. Largy (Eds.), *Revision cognitive and instructional processes. Studies in writing* (Vol. 13, pp. 21–38). Springer. https://doi.org/10.1007/978-94-007-1048-1_3

Pisacreta, D. (2013, June). *Comparison of a test delivered using an iPad versus a laptop computer: Usability study results.* Paper presented at the CCSSO National Conference on Student Assessment, National Harbor, MD.

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

Proctor, T. P., Chuah, S. C., Montgomery, M., & Way, W. D. (2020). *Comparability of performance on the SAT® suite of assessments across pencil-and-paper and computer-based modes of administration*. The College Board. https://collegereadiness.collegeboard.org/pdf/comparing-performance-paper-digital-tests-sat-suite-assessments.pdf

Russell, M. (2018). Recent advances in the accessibility of digitally delivered educational assessments. In S. Elliott , R. Kettler, P. Beddow, & A. Kurz (Eds.) *Handbook of accessible instruction and testing practices* (pp. 247–262). Springer, Cham.

Sisley, J., Kia, K., Johnson, P., & Kim, J. (2017). Effects of key travel distances on biomechanical exposures and typing performance during ultralow key travel keyboards. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 981–985. https://doi.org/10.1177/1541931213601727

Steedle, J., McBride, M., Johnson, M., & Keng, L. (2016). *Spring 2015 digital devices comparability research study*. The Partnership for Assessment of Readiness for College and Career.

Strain-Seymour, E., Craft, J., Davis, L. L., & Elbom, J. (2013). *Testing on tablets: Part I of a series of usability studies on the use of tablets for K-12 assessment programs* (White paper). Pearson.

Supovitz, J. (Host). (2020, December 17). The nation's troubling report card: Equity and diversity K–12 policy [Audio podcast episode]. In *Research minutes: Education research and policy podcast*. Consortium for Policy Research in Education. https://www.researchminutes.org/episode/the-nations-troubling-report-card/

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22,* 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J., van Merrienboer, J. G., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. https://doi.org/10.1023/A:1022193728205

Van Waes, L., & Schellens, P. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics*, *35*(6), 829–853. https://doi.org/10.1016/S0378-2166(02)00121-2

Watzman, S., & Re, M. (2012). Visual design principles for usable interfaces: Everything is designed: Why we should think before doing. In J. A. Jacko (Ed.), *The human–computer interaction handbook—Fundamentals, evolving technologies and emerging application* (3rd ed., pp. 315–340). CRC Press. https://doi.org/10.1201/9781410615862

Way, W. D., Davis, L. L., & Strain-Seymour, E. (2008). *The validity case for assessing direct writing by computer* (White paper). Pearson. http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TheValidityCaseforOnlineWritingAssessments.pdf?WT.mc_id=TMRS_The_Validity_Case_for_Assessing_Direct

Yu, L., Lorié, W., & Sewall, L. (2014, April). *Testing on tablets*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.