A Tale of Two Modes:
A Case Study in
User-centered Design’s Role in
Comparability and Construct Validity

Ellen Strain-Seymour, Ph.D.

Pearson

Introduction: UCD’s Role within User-centered Assessment Design

One merit of user-centered assessment design (UCAD) as defined by Nichols et al (2008) is its broadened view of test development. As a methodology, it takes into account a range of users beyond the test-takers and addresses a wider context for assessment including, for instance, test materials handling, report design, and the use of assessment data by students, teachers, parents, and administrators. However, it could also be said that even when narrowing the focus to computer-based testing and the validity of those assessment results, UCAD offers an additional dimension to traditional educational measurement and extends its analytical methods in critical ways during a time of increased online testing.

When UCAD is applied to even a narrow definition of test design and delivery, what Nichols calls the “gestalt” of UCAD’s “conjunctive approach” that combines user-centered design (UCD), evidence-centered design (ECD), and the best practices of conventional test development is apparent. The symbiosis that makes this approach more than the sum of its parts derives in part from the strong similarity between UCD and ECD combined with their analytical rigor honed in different areas of practice. While ECD is at home in the field of educational measurement and assessment design, UCD is a relative newcomer within test publishing. With methods sharpened in the academic field of human-computer interaction and in the applied field of technology product design, UCD’s recent appearance within test design practices is a result of the growth of computer-based testing (CBT).

Together UCD and ECD offer a comprehensive framework for tracing the integrity of test design from the identification of a construct for evaluation to the creation of items and a scoring design and finally to the test-taker’s experience of items and engagement with computer-delivered test content. UCD and ECD align in their rigorous analysis of “tasks” as the essential interpretive unit and in their concern for the conceptual models that facilitate task completion. ECD investigates construct validity by asking whether an assessment task is parallel to a classroom task and whether completion of that assessment task is indeed indicative of
students’ possession of certain knowledge and skills (Mislevy 2006). UCD has similar roots in cognitive science and also places significant emphasis on task analysis, although not exclusively for the purposes of educational measurement. UCD methods, typically intended for investigation of tasks conducted via computer applications and electronic devices, includes two general categories of task analysis. In the first, a task for which a non-computer-based equivalent does not exist is broken down into its constituent steps and cognitive processes to understand how an application’s functions and user interface support or hinder the user in achieving his or her goals. The second type of task analysis is applied when a physical or paper-based task is converted into a computer-based task. Comparative task analysis is required to understand the original task and then to design or evaluate a computer application to be used for accomplishing the same task. In either case, the emphasis is on gathering evidence from users in an attempt to understand the conceptual models that undergird a task and using that understanding to inform and improve application design.

While this paper should be read in the context of UCAD as a proposed set of practices, the emphasis is on UCD’s individual contributions to this conjunctive approach and on the type of task analysis and user research that characterizes a UCD approach applied to online test design. The paper begins with coverage of two areas that form the background to the UCD case studies that follow. First, comparability literature is re-read through the lens of usability engineering to gain a sense of application design’s potential role in ensuring comparability. Also in this section is a classification of design problems that can disadvantage online test-takers and the questions to pursue to eliminate such problems. Moving from psychometric to organizational concerns, the second background area fleshed out in this paper is a glimpse at possible organizational arrangements that can help foster a UCAD approach within a large test publishing company. The remainder of the paper is dedicated to exploring several CBT interface redesign efforts, with particular emphasis on the revision of an essay-writing interface. This case study highlights the ways in which user research and usability testing can be combined with more traditional psychometric research to interrogate construct validity, address comparability, and inform policy around the transition to computer-based testing.
UCD Applied to Comparability and Construct Validity

The test delivery environment - whether paper or computer-based - is a medium through which students must perform assessment tasks. Ideally, this environment is an invisible and faithful conduit for collecting evidence used to make inferences about student skills and knowledge. In reality, the testing interface can introduce construct-irrelevant variance that interferes with test validity (Huff & Sireci, 2001). For instance, a student’s difficulty with filling in bubbles on an answer sheet or with navigating between items in a computer-based test may result in score contamination. In the case of CBT, computer literacy, if not a part of the construct being measured, can become an “ancillary non-construct element” (Haertel & Wiley, 2003, 1). Thus, a testing interface’s intuitiveness - its ability to be effortlessly used by subjects without instruction or prior exposure - is the ultimate goal of test designers and usability engineers.

The typical mechanism for detecting CBT-introduced construct-irrelevant variance is comparability analysis of scores for paper-based and computer-based tests. A difference in performance across the two test delivery modes, assuming that the effects of other variables have been largely eliminated, is known as a mode effect: a source of performance variance that is assumed to be a product of the mode of test delivery. Currently, comparability studies play at least two important roles. Most states engaged in online testing for K-12 high-stakes tests must also provide paper-versions of those tests until schools acquire adequate technology infrastructure to support universal CBT. While dual-mode testing is practiced, fairness across paper and online forms of assessments must be maintained, as underscored by the Guidelines for Computer-Based Tests and Interpretations (APA, 1986): “...when interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cut scores obtained from conventional tests” (p. 18). In addition to addressing fairness, comparability studies provide a benchmark for the designers of computer-based testing systems. While conventional testing does not necessarily involve a total absence of construct-irrelevant variance, most agree that at least while CBT is a new experience for some students, the use of computers may introduce
more opportunities than paper for such variance. Comparability studies, when combined with UCD research, provide opportunities to identify and eliminate CBT's potential compromises in construct validity.

While published comparability studies have found mode effects in both directions - advantaging online test-takers in some cases and paper test-takers in others - more frequently, if any mode effect is present, it is to the advantage of paper test-takers (Haas & Hayes, 1986; Choi & Tinkler, 2002; Pommerich, 2004; Paek 2005). Fortunately for CBT proponents, comparability literature shows diminishing mode effects over the past two decades of computer-based testing (Clariana & Wallace, 2002). While the effect of students’ increased familiarity with computers should not be totally discounted, reading this literature through a usability engineering lens produces a convincing tale of causality: comparability differences have been and continue to be eradicated through improvements to online testing interface design and functionality.

In the earliest comparability studies involving CBT, references to the testing interface are either off-handed or absent altogether. However, with a careful reading and the benefit of industry-acquired background, a timeline of increased comparability mapped against evolution from the most rudimentary testing systems to today's full-featured systems can be constructed. For instance, some of the earliest systems lacked basic capabilities such as the ability to navigate back to a prior item to review or change an answer (Pommerich & Burden, 2000). Today's systems have a variety of tools: text highlighting, the ability to “bookmark” an item, sophisticated navigation, item marking tools, and ways to locate skipped items. Lending further support to the idea that interface design and functionality plays a key role in comparability is the fact that the areas where mode effects still linger are in subjects that require a complex interface, such as reading comprehension where solutions for balancing screen real estate between items and multi-page passages are still being refined (Higgins, Russell, & Hoffmann, 2005).
Recent research more fully highlights the role of interface design in diminishing mode effects. For instance, Muter (1996) recommends studying individual responses to interface features in order to identify any sources of variation across pencil-and-paper and computer-based tests and then redesigning the interface to accommodate differences in user behavior and preferences. Both Higgins et al (2005) and Mazzeo et al (1991) have described repeated comparability studies using iteratively redesigned interfaces. Despite these signs of the ascendant profile of usability engineers working alongside psychometricians and test designers, absent from the research is a full elucidation of how UCD principles can be applied to the pursuit of comparability and where best to employ user research and usability studies within the test development process.

**Interface-related Mode Effects**

A starting point for understanding UCD's potential contribution to comparability is a classification of the types of interface-related problems that can disadvantage online test-takers. While some interface/functionality failings can manifest themselves in ways that span multiple categories, generally an interface issue that threatens comparability can be assigned to one of the three broad categories described in Figure 1. It should be noted that the interface or interaction between test-taker and computer can be broadly understood to include not only application design but lighting, monitor quality, hardware and network reliability, and ergonomic factors that are beyond the control of user interface designers but which nonetheless may contribute to mode effects. For instance, young test-takers may be at risk for discomfort-related performance dips when taking long tests within with shared-use computer labs that are more ergonomically suitable for sixth-graders than for the small bodies of second-graders. Similarly, mobile computer labs used for testing may include laptops with small screens set up in unfavorable lighting conditions that produce screen glare. Such environmental and hardware issues are included in this first category of usability failings. More easily addressed by test and technology development researchers, however, are usability problems that stem from the interface design of testing software.
Any interface design feature that causes frustration or diverts time and energy away from the test content may contribute to a mode effect. Certainly, diminished performance could be caused by time spent learning electronic testing tools, frustration that decreases motivation, or heightened anxiety produced by what is perceived as an unfriendly testing application. Some mode effects are statistically insignificant at the item level but become measurable at the test level, while others - perhaps caused by scrolling or poor image resolution within a detailed map, for instance - are more detectable at the item level. Cognitive load theory suggests that mode effects may affect more difficult items. Functionality that falls short of being intuitive may increase demands on a test-taker’s working memory and pull cognitive resources away from the problem-solving and reasoning required to answer a test item (Sweller, 1988). Thus, usability failings could be expected to have a greater impact on students who are working in the upper ranges of their ability and who are closer to hitting the limits of their working memory. Taking into consideration all of these possible factors, mode effects would seem to be more likely to affect students who are less motivated, less computer literate, and/or less proficient in the skill being measured. The danger then is that CBT-introduced construct irrelevant variance may exaggerate pre-existing performance gaps. And indeed, some evidence of this has been found, with patterns of disproportionate effects on certain student groups, although sometimes falling short of statistical significance (Way & Fitzpatrick, 2006).

In addition to this first category - usability failings - a second category of interface-related issues that may lead to poorer performance among online test-takers is reduced functionality in comparison to what paper test-takers may be able to do with their pencils and test booklets. A combination of common sense and user research - observing test-takers, interviewing teachers, conducting focus groups with students - has led to a list of missing functions, many of which have been added to testing software in the past two decades. Test-taking strategies such as eliminating answer choices, underlining key words, and checking for skipped items are now better supported in online tests. Unfortunately, in some cases, this expanded functionality has led to usability problems as described in the first category, such as
when a virtual highlighter or online reading guide is difficult to use. Policy has also been affected by such considerations; online test-takers are frequently allowed to have scratch paper like paper-and-pencil testers so they can create graphic organizers before writing an essay or fold their paper into three-dimensional shapes in order to answer geometry items involving nets.

The final category of potential problems tends to apply more to complex assessments tasks delivered through computer-based testing. When virtual equivalents of content-specific tools and simulated environments are delivered through CBT, some aspect of the design may significantly change the nature of the task. Often rather than being an issue of poor usability, it is a case of failed equivalence. Some factor prevents test-takers from completely transferring their knowledge to this context and exercising their skills within the assessment task. For instance, the translation of physical tools - compass, ruler, microscope - into a computer-based equivalent may fail to provide students with key tactile or auditory cues or otherwise interfere with the exercise of their classroom skills within the virtual environment. This type of task inauthenticity can rear its ugly head even when the classroom or real-world skills being evaluated involve computer-based tasks. A testing environment’s re-creation of certain computer-based functions may be dissimilar to classroom applications thereby hindering student performance. This third category of interface-related design flaws results in a comparability issue when a paper version or a hands-on task is offered as an assumed equivalent. The paper version may offer more supports, such as a textual description of what must be observed in a virtual world or images of key steps that the online test-taker must perform in the simulated environment. Paper and hands-on equivalents may involve greater authenticity when drawing upon familiar, classroom tools - protractors for geometry assessments, pH strips for chemistry tests, globes for world geometry exams - rather than virtual versions of those tools.
UCD: Organizational Implications and Case Studies

A very practical concern for organizations considering a UCAD approach is where in the organization to place UCD practitioners and researchers working on issues of comparability. Small and large test publishers may have different options in this regard. Within most organizations, including many test publishing companies, UCD specialists work under the Information Technology (IT) umbrella, interacting closely with product managers, business requirements analysts, and programmers - the staff most directly involved with application design. In a small company, such placement may still offer opportunities for close interaction with content developers, psychometricians, research staff, clients, and end-users. Larger companies may have more compartmentalized departments that inhibit this kind of collaboration, with IT groups tasked with creating “content neutral” applications and only occasionally engaged in projects that involve day-to-day interaction with psychometricians and end-users. However, such companies may also have more robust testing software solutions that allow for greater customization at the item and tool design level. In this case, UCD specialists can, like subject matter experts and psychometricians, be assigned to one or more individual testing programs where they become intimate with the content, test data, and accountability requirements associated with a distinct set of tests.

When grappling with issues of how interface design can support construct validity, comparability, and task authenticity, UCD researchers’ success is dependent on several factors: familiarity with the intricacies of a testing program, integration into the daily decisions made about content design, access to the client and end-users for user research purposes, and close interaction with content experts and psychometricians. In many cases, such as in the design of innovative items (e.g., virtual labs for assessment purposes), content design and interaction design, are not easily separated. Thus, while UCD practitioners are certainly required within IT departments that direct product design, an argument can be made for also including UCD specialists within the thick of content development, working closely with staff interfacing with clients and assuring the integrity of test design down to the level of the test item. This
argument is even stronger when the testing software supports a number of options for item formatting and tool use, thereby allowing item design and delivery decisions to be made at a local and granular level to support content and psychometric concerns that are specific to an individual testing program and even to an individual item.

The interface redesign initiatives described below have been or are in the process of being undertaken in the context of a single assessment program, albeit one of the largest state-based assessment programs. Design decisions and research efforts are conducted collaboratively with the client, psychometricians, content experts, and UCD staff housed within the content group. The prototyping and ultimate delivery of customized online tools, interactive items, and other interface elements is in this case supported by the testing software’s integration of Adobe Flash. Items and the majority of student-facing interface features can be created and delivered through Flash, which sits atop a secure system with backend data handling services that connect with scoring and reporting systems. Flash’s support for rapid development with very limited technical staff has been critical to this effort to create customized solutions to content-specific comparability problems.

In the efforts described below, user-centered design practices were introduced within a client-vendor team that was already highly attuned to the specificities of the online student test-taking experience. Such awareness has been used to inform a number of on-going transitions in the state assessment program. A high-stakes testing program offered only on paper has been transitioned to dual-mode testing at the higher grade levels. A paper component is being added to an online-only, low-stakes end-of-course program that will soon become a graduation requirement and a source of accountability data. And an English language-proficiency test is being transitioned from paper only to online only. An understanding of comparability, enabled in part through usability and appropriate computer-based functionality, has been necessary to handle these transitions with sensitivity to impact on students.
This thoughtful approach to the delivery of online tests has included a number of best practices:

- When online is the primary delivery mode, clients see items in their online form throughout the item development process, so that any interaction between the conditions of online delivery and the content (e.g., image resolution of scientific illustrations delivered online, item fit with limited screen real estate, on-screen ruler use within measurement items) can be evaluated as part of the content review, rather than retrofitting for online following content review.

- External committees of teachers and curriculum specialists work through items on a laptop during meetings in order to experience the item as an online test-taker would.

- Formal observations of test-takers accompany field testing when expanding online testing to student populations without extensive computer experience (e.g., second graders with limited English language proficiency).

- Surveys are administered to students and test administrators following nearly every online test to gauge satisfaction and identify areas for improvement.

- A comparability study follows every administration involving dual-mode testing.

While surveys reveal high satisfaction with online testing among participants, psychometric data and qualitative information from the above sources has led to the identification of areas for design improvements. Some tests - namely math and reading comprehension tests - are still plagued with minor comparability issues, and less quantifiable compromises in user experience of online tests have been detected. The user research and interface redesigns described below have been a response to these findings.

- **The introduction of a pencil tool that allows students to mark the on-screen item.**

  Observation of test-takers and inspection of test booklets reveal that many students using a paper test booklet will support their problem-solving by marking a geometric figure with symbols to indicate equidistance or parallelism or with numbers to indicate dimensions. Observational information and data from surveys of online test-takers indicate that students taking the test online do one of the following:

  - Take the time to duplicate the geometric figure on scratch paper in order to make these same types of marks.
- Attempt to keep calculated dimensions and relationships in memory while working through a test item.

- Use the highlighter or dot tool to simulate such marks and experience frustration in such attempts since these tools were not designed for making such marks.

Item analysis on math tests showing mode effects suggest that items that invite multi-step, in situ calculations such as described above are more prone to mode effects. In response, a pen and a pencil tool for making freehand marks on-screen has been developed and will be submitted to usability testing prior to operationalizing.

- **The improvement of the online ruler, protractor, line-draw, and compass tools.** Feedback from teachers and users suggested that the virtual equivalents of geometry tools that can be used in combination with pencil-and-paper tests and classroom assignments could be improved to have greater ease-of-use. Computer-based tools used in eLearning applications like as Geometers Sketchpad were analyzed as well as student use of physical tools. Key issues included ruler and protractor transparency, ruler length for effective rotation on-screen, “pointy” versus safety compass use in the classroom, radius “locking” on a compass during geometry constructions (as is used for locating perpendicular bisectors), and freehand versus straightedge-assisted line drawing. Tools were redesigned with sensitivity to these issues and submitted to usability testing leading to further redesign. Another round of usability testing will be conducted with subsequent rounds of redesign and usability testing as required, prior to operationalizing.

  INSERT FIGURE 2

- **The development and testing of a design strategy for converting science lab tools to computer-based equivalents.** When converting equipment like a digital scale, triple-beam balance, microscope, and titration set-up to an on-screen equivalent, two key challenges were identified. First, a photorealistic rendering of the equipment intended to assure recognizability were found to obscure which controls were interactive and caused motivated and computer literate users to engage in “roll-over hunts” to discover functionality. Other subjects simply
failed to locate the interactive controls. In the improved design, all interactive controls were rendered in a two-dimensional form on a foreground layer on top of the photorealistic image, with each control placed in close proximity to its location on the actual tool. Secondly, using established conventions for displaying 3D movement such as movement in depth along the z-axis risked advantaging experienced gamers who are more adept in these conventions for translating between 3D and 2D environments. Thus, all z-axis movement, such as adjusting the stop-cock on a titration set-up or the focus knob on a microscope, was translated to more common button and slider controls. Designs were iteratively modified with the aid of usability testing and the input of content specialists. With the aid of content specialists, less construct relevant functionality - such as turning on the microscope light - was excluded, while key concepts (e.g., students must focus on the lowest magnification level before moving to a higher magnification level) were full realized in the design of the virtual microscope. A dimensionality study by psychometricians is being planned to compare such items to the nearest paper equivalent.

INSERT FIGURES 3 & 4

- The design of a “paging” reading passage interface. Research suggests that any item that requires navigation, scrolling, or opening a separate window may be vulnerable to mode effects (Mazzeo & Harvey, 1988; Murphy, Long, Holleran, & Esterly, 2000; Choi & Tinkler, 2002). This poses a risk for any item that does not fit on one screen or which requires navigating to a formula chart or reading passage on a separate screen or in a window that partially obscures the student’s view of the item. This risk is exacerbated in the case of multi-page reading passages presented within a scrolling window (Higgins, Russell, & Hoffmann, 2005). In addition to the required computer interaction (scrolling, navigation, managing a separate window) that may come less naturally than test booklet page turning, positional memory is not supported when a passage is displayed in a scrolling window or non-page-based way (Dillon, 1992; Haas & Hayes, 1986; Higgins, Russell, and Hoffmann, 2005). When reading a passage on paper, students may remember that content relevant to a test item was at the top of the second
page, for instance, and quickly flip back to the appropriate page in response to a test item. However, when a student scrolls through a multi-page passage, page breaks, which form a kind of spatial bookmark in memory for paper testers, are often absent. In response to comparability problems in reading comprehension tests assumed to be due to the design issues described above, a working group of content specialists and psychometricians was formed. An initial redesign has been created based on existing research and observational data and submitted to evaluation by this working group. The design includes a side-by-side passage/item display with a “paging” interface within the passage itself. Additional research and redesign will be conducted via iterative usability studies using small sample sizes followed by a larger-scale psychometric study.

A Case Study: Essay-writing Interface

While any of the above efforts could be the topic of an extended case study, the interface re-design efforts used to improve an essay-writing interface provides the benefits of (1) data from usability studies, follow-up field tests, and operational tests and (2) examples from each of the categories of design problems described in Figure 1 above. This case study’s initiating event - the discovery of a mode effect that disadvantaged online test-takers of a high-school writing test - was marked by complexity from the start. Particularly troubling was the scoring model that did not lend itself to slight score adjustments to account for mode effects. Since a score of 2 within a four-point scale on the essay portion of this state-wide test is required for graduation, the typical raw-to-scaled-score adjustments to account for slight mode effects within a 50-item test could not be easily implemented. Any whole number adjustments to essay scores meant affecting graduation eligibility for large numbers of students.

The option to take the writing test online as part of a comparability study was first offered as part of an additional re-test opportunity, with campuses volunteering for participation and students randomly assigned to condition (paper or online). The mean score
for online essay-writers was 1.56, as compared to 1.69 for paper essay-writers. The distribution of essay scores also differed across modes; 59% of paper test-takers earned a passing essay score while only 49% of online test-takers passed. Various factors were investigated, but the top two issues earmarked for investigation were (1) scorer bias and (2) interface design. Lingering in the background was also the question about whether online essay writing and paper-and-pencil writing might involve different constructs. For this reason, policy decisions and critical next steps in the transition to online hinged on the findings of our multi-pronged investigation of this mode effect.

Research supports the idea that bias among scorers can lead to overly harsh scoring of typed essays in comparison to hand-written essays (Breland, Lee, & Muraki, 2005; Way & Fitzpatrick, 2006). One possible explanation for this bias is that scorers may approach the typed essays as if they were more like final drafts than the hand-written equivalents and be less forgiving of small errors, including spelling errors that can not be hidden using ambiguous hand-written letters (the slightly open at the top ‘o’ when a student is unsure if a word is actually spelled with a ‘u’). A series of studies were conducted to compare human and automated scoring of both hand-written and typed essays and to rescore essays converted to the other mode (hand-written essays were typed and typed essays were duplicated in hand-written form). While the results of the automated essay scoring study did lend support to the theory of scorer bias, the study was marked by some data anomalies that suggested some confounding variables that diminished our confidence in our ability to accurately gauge the level of impact to ascribe to rater bias. Nonetheless, based on some evidence of scorer bias, a series of steps were taken to incorporate more type-written essays into the scorer trainer and eliminate common biases against typewritten essays.

One product of the latent semantic analysis involved in the automated essay scoring bias was comparative data on the semantic characteristics of the two sets of essays. Most notable were the differences in essay length and sentence construction across typed and hand-written exams. Since essays created in the online test environment were lengthier overall and
had longer sentences, we were interested to understand factors impacting revising/editing and student decisions around essay length. Longer essays and more complex sentence structures might be assumed to be correlated with higher scores, but since the opposite was true in this case, we investigated how the essay-writing interface might be a contributing factor.

While UCD is best employed prior to application design in order to understand and account for user behaviors, priorities, and preferences from the start, it can also be used to detect and resolve usability problems within existing applications. A usability study should balance a targeted approach with a generalized “catch-all” approach. In other words, the research method should resolve hypotheses and suspicions around specific interface features but should not be so targeted as to miss unanticipated usability issues. In this case, we wanted to conduct user research to target the issues of essay length and revision strategies, but we did not want to discount the fact that other unrelated usability issues divert energy away from best practices in essay creation and revision. Thus, we chose a combination of three UCD methods - user personas, task analysis, and usability testing - to diagnose problems and redesign the essay-writing interface.

We began our construction of user profiles and our task analysis by conducting one-on-one interviews with high school students who have taken the state writing exam on paper. We included students that fell across the continuum for computer experience but with strong representation of computer novices, with the expectation that this population may be more vulnerable to mode effects. Beyond this criterion, we aimed for representativeness in terms of gender and economics. In open-ended interviews with our subjects, we queried students about their writing processes on high-stakes tests and in their classroom activities. Based on salient characteristics that emerged during the interviews, we organized user types along two axes: writing process style and our original criterion, computer experience.

Two user groups were immediately apparent in the area of writing process: planners and extemporaneous writers. Planners reported using some combination of outlines, notes, graphic organizers, and multiple distinct drafts in both testing and homework situations. While
taking high-stakes writing tests, planners expressed concerns about having enough scratch paper and planning appropriately for the two-page maximum. Extemporaneous writers had a “dive right into it” attitude and trusted that the appropriate thinking would happen organically through the process of writing. Further analysis revealed that extemporaneous writers could be broken down into two smaller groups: less motivated students and advocates of a “flow and rewrite” approach. The key distinguisher between these two groups (beyond general investment in producing quality writing) was the presence/absence of a revision step. Unmotivated students engaged in neither essay-planning nor rewriting, while flow-and-rewrite advocates often reported spending more time re-working than creating an original draft. Unlike planners, this group of students did not always distinguish between a rough draft and a final draft; instead writing and reworking were integrated processes, without a distinct break between the two steps.

Within user-centered design, user profiles or personas are a way to both encapsulate results of user research (interviews, focus groups, surveys) and create representations of user groups that can be used to guide interface decisions. In addition to a general description of user priorities and preferences, a user profile may include a fictional name and even a stock photograph in an attempt to make the user profile more concrete to designers and programmers. (See Figure 6 for a sample user persona.) During a prolonged development phase for an application, names and faces can help the design team to keep user groups in mind long after initial user research has been conducted. At key points within the design process, these concrete representations of user groups are used to guide decision-making. For instance, all multiple-choice items within the testing software have a Reset button that returns the item to an unanswered state. The Reset button is often carried over into other item types. A discussion about this button among designers might draw upon user personas to understand implications for different user groups. Clara, the name for our diligent and well-informed student who favors planning, would review the tutorial, understand the button’s function, and would in all likelihood recognize it on the essay item. Fred, another user persona, is not computer shy but learns functionality through exploration. During usability testing, we witnessed that when the
motivation of the subject who most epitomized this user persona faltered, he took a break from the test by exploring buttons and functions, even doodling with the highlighter. For Fred, one careless, exploratory, or bored click of the Reset button could lead to the potentially devastating effect of a deleted essay. Out of the designers’ concern for Fred – who in many cases may belong to one of the student groups that schools are targeting for improvement for NCLB’s Adequate Yearly Progress (AYP) purposes – the Reset button was excluded on the essay item. Its potential disadvantage for the careless clicker far outweighed its possible utility for the student who wishes to erase and start over.

INSERT FIGURE 6

For our re-design of the essay interface, we relied on the matrix in Figure 7 for our user personas and continued our interview process to understand how these user groups might be impacted by using a computer to write essays on the exam as opposed to writing by hand. All of our subjects had experienced the paper-and-pencil version of the test, but none had experienced the online version of the assessment since the prior comparability research study had been limited in scope. The table below shows our six user groups and represents predicted reactions to the possibility of a computer-based essay item.

INSERT FIGURE 7

Our comparative task analysis was further fleshed out through student survey data, information from curriculum specialists on writing instruction, interviews with scoring center personnel, and the viewing of type-written and hand-written essays. We also performed a heuristic analysis of the current interface to arrive at theories about possible usability problems of the type identified as Category One issues in Figure 1. A number of questions were also posed in regard to Category Two type problems: “supports” and “functionality” allowed on
paper but not on computer. What kinds of things are students doing on the paper version of the test that may not be supported online? Are they including sketches, starring/underlining text, rearranging paragraphs through arrows, making their writing smaller as they approach the end of their two pages in order to fit more? If so, are these mechanisms having a positive effect on their scores? It turned out that these were rarities, and raters were instructed to ignore most of these types of markings (except paragraph rearrangement). However, without the benefit of the user research, we could not be sure that paper was not providing some kind of support for the writing process that did not exist online. In terms of Category Three type problems, we spent a large portion of our interview time investigating computer-based writing methods as compared to hand-writing, both in the classroom and in the context of an exam. Were students asked to write interchangeability on paper and using computers for their classroom assignments? Did their methods differ when writing on computer? What were their methodologies when using word processing applications such as Microsoft Word and are those methods transferable to our essay-writing interface within the online test? Is there anything prohibiting the transfer of their preferred computer-based writing methods to the assessment?

Using these questions to analyze our transcripts of the user interviews, we boiled down our findings to a number of key observations:

- Over half of the student interviewees report changing their writing methods when moving from paper-based to computer writing. Nearly all of these students reported that the methods they use when writing on a computer are their preferred and more typical style of writing.

- Due to differences in preferred writing methods and varying levels of comfort with computers, the transition from paper to online is likely to affect different user types differently.

- Editing tools within the online interface are critical since many students shift time spent in pre-writing over to time spent revising when moving from handwritten essays to computer-based writing.
• Space management is important to students when a page limit is imposed, such as on the state writing assessment.

The interface was then re-designed to provide more visible editing tools, while still supporting keyboard shortcuts and complete interchangeability across button-triggered editing functions and keyboard-controlled editing functions. The amount of new functionality was minimized to avoid advantaging students with greater exposure to word processing. The writing prompt was included on the same page as the essay (unlike in the prior design where the writing prompt was on a preceding screen). An open design with ample “white space” was balanced with the need to show as much of the student’s essay as possible. And a concrete sense of the amount of space remaining was provided through a “page preview” that resembled both the two-page test booklet and Microsoft Word’s print preview function.

A one-on-one usability test was then conducted with ten students. Each student was asked to explore the interface and then use it to write an essay. Using screen-recording software and voice-triggered audio recording, the students’ activities were recorded using the think-aloud protocol. Following instructions to “think aloud”, when students encountered interface functionality, they gave voice to their assumptions about a button’s function, expressed their surprise or frustration over any of their misconceptions about functionality, and talked through their discoveries regarding interface features. Following this exploration, essay writing was occasionally interrupted by additional discoveries about the interface, which students again described out loud for capture by the audio recording software. Completion of the essay was followed by an interview and then a survey for rating satisfaction with the interface and prioritizing functionality.

Overall, the interface rated well with users. However, some remaining usability problems were identified and changed in the final design. For instance, the highlighter tool had not been redesigned; it had been moved intact from the original interface to the new design. Not only did several students have difficulty using the tool as intended, additional frustration
was caused by the fact that the highlighter does not work on student-typed text (only on item text). Two students gave up on the tool, while another spent three minutes attempting to make it work. To avoid potential student confusion over whether scorers would see highlights on essay text, the highlighter was removed rather than re-designed. Another area of confusion was the page preview tool, which students generally liked but which also raised some questions. The preview of space remaining functioned like a progress bar showing generic text in paragraphs that gradually filled the two pages, representing the percentage of the character maximum that had been used. If half of the character maximum had been used, one of the two pages was filled with text regardless of how many lines a student had used in their essay. The use of a character maximum mapped uncomfortably to the page metaphor. A student could use up the character maximum by typing one character followed by a line return, filling up dozens of “pages” this way, while the page preview would display two pages filled with paragraphs of text. While our savviest subjects took a moment or two to understand the approach, other students were confused or ignored the page preview because it did not seem to be a match to their essays. In response, policy was changed to use a line maximum rather than a character maximum. The page preview was redesigned to show the exact line breaks of the student’s essay. Filling the two pages shown in the page preview coincided with using the maximum number of lines, regardless of how many characters were in each line of text. This technique was seen to be more consistent with paper-based space management techniques; double line returns between paragraphs could be deleted if extra space were needed in the same way that the hand-writer can write smaller as s/he reaches the bottom of the second page.

By the end of our UCD-based study, we had found design problems in all three categories described in Figure 1. Usability failings such as in the highlighter tool could have been causing students wasted time and frustration. Paper supports, such as the ability to easily see how much of the two pages a student had filled, were missing from the original online design, which featured only a scrolling window that increased in size as text was added until a student tried to type that one character that exceeded the character maximum. And lastly, while the transference of paper-based writing methods from the classroom to the assessment
were facilitated, the transferal of computer-based writing methods - which are far more
focused on revision than on pre-writing - were not fully supported due to weak editing tools in
the original online design.

The small sample size of 10 students used in the usability study was initially a cause for
concern among psychometricians who had not had prior exposure to usability testing. The goal
of usability testing is not statistical validity; it is the discovery of usability problems through
behavioral observation. Although results may vary with the heterogeneity of your user groups,
generally accepted usability testing guidelines suggest that additional information gained with
a larger sample size is limited after the first 10-12 users. The majority of usability problems
will be encountered with six users, with virtually no additional issues discovered with users
beyond the initial 15 tested (Nielsen & Landauer, 2003). Available time and funding should be
dedicated to additional iterations of designs and usability testing rather than on high numbers
of subjects for any single round of usability testing.

In the case of our analysis of online essay writing, usability testing provided that which
could not be gained from looking at the comparability data alone. The qualitative data
received from usability testing provided context for understanding the mode effects we were
witnessing and a basis for believing that our new design had the potential to diminish the mode
effects. Usability testing also helps with an additional check: it can be verified that
functionality added to provide improved comparability does not introduce new usability issues.
However, follow-up comparability studies are required to assure that the mode effects have
indeed been eliminated with a revised design. In this case, a follow-up study was conducted in
an unmotivated research study in which students were randomly assigned to one of three
conditions: paper, the original interface, and the revised interface. A total of 825 campuses
participated with at least 1000 students assigned to each condition. The study design was
similar to the original comparability study, with random assignment to condition followed up
with matched samples comparability analyses (MSCA). Student profiles were matched across all
three conditions using gender, ethnicity, and ability levels (using prior test scores). Five
hundred bootstrap replications were conducted to form sampling distributions and to estimate the standard error of the scale score and raw score equivalents. Item level statistics were obtained for each selected (bootstrapped) sample, and averaged over the 500 iterations to obtain the sampling distribution of the item means. The resulting mean for the paper version was 1.83 out of 4 in comparison to 1.89 for students using the original online interface and 2.08 for those using the revised interface. Additional analyses on the essay performance indicated that females, African-Americans, and Hispanics were most affected by the improved interface and were more likely to obtain an essay score of 2 or higher with the revised interface than with either the original interface or on paper. Not only did these results give us reason to believe that our improved design had been successful; the data suggested that our scorer training to eliminate rater bias against type-written essays had been successful.

The live administration that followed (using only the revised interface in addition to paper) showed similar results, but with gains for all ethnic groups taking the online test using the revised interface in comparison to paper. (The reduced motivation involved with a research study that does not provide students with a score seemed to be a larger factor for white males whose performance was least affected by the new interface within the research study.) Following MSCA, the mean was 1.72 for online test-takers in comparison to 1.57 on paper. Sixty-one percent of online test-takers received a 2 or higher on the essay, while only 55% of paper test-takers received a passing score of 2 or higher.

If comparability was our goal, then we overshot our mark. If we had only our before-and-after comparability data, then we may have been left in a quandary over policy decision offering both paper and online versions of the state writing test. However, our user research revealed that greater authenticity may be involved in the online essay test for many students, which may explain the improved results. For schools with fewer technical resources, writing on computers may be less emphasized in the curriculum. These may be the same schools that will not be able to offer all of their students the option of taking the essay test online. Luckily, the
paper test may be the more authentic assessment task for many of these students. The recommendation to schools is that the mode of assessment matches the mode of instruction. In the end, issues of comparability, however important for assuring fairness, can not outweigh considerations of construct validity; students should have the option of taking the test online or on paper, in order to be assessed in the mode where they will be most able to demonstrate their writing abilities. And the information needed to guide such decisions and to optimize the online testing experience must come from a combination of user-centered design methods and psychometric analysis.

This study points to two directions for the future. First, in an era of online testing, the research methods of test developers must be multi-disciplinary, drawing on the synchronicities between user-centered design, evidence-centered design, and psychometric methods to assure construct validity within computer-based testing and comparability within dual-mode testing. Secondly, in the near future UCAD researchers and practitioners might no longer be trying to address the weaknesses of computer-based testing to bring it up to the benchmark of paper-based testing. Instead, that benchmark may be surpassed. Online functionality may be used to improve the quality and authenticity of assessment, with these kinds of improvements being the motivator that drives the transition from dual-mode testing to online only testing.
References


### Figure 1: Classification of User Interface-related Mode Effects

<table>
<thead>
<tr>
<th>ISSUE</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POOR USABILITY</strong></td>
<td>Are any aspects of the interface frustrating? Do they consume time to understand or operate, or otherwise divert energy and attention away from test content? Poor text or image quality, confusing functionality, eye strain-inducing background/text color combinations, inferior monitor contrast or luminescence, poor ergonomics in shared use computer labs.</td>
</tr>
<tr>
<td><strong>REDUCED FUNCTIONALITY / REMOVAL OF PAPER SUPPORTS</strong></td>
<td>Are there things that test-takers can do on paper that can not be done online? Might the absence of these supports impact performance? Lack of ability to: mark in margins, sketch on the test item, change answers easily, gauge progress through test booklet, eliminate answers, mark for review, refer back to booklet pages using scratch paper bookmarks.</td>
</tr>
<tr>
<td><strong>INAUTHENTICITY / LACK OF EQUIVALENCE WITH CLASSROOM TOOLS</strong></td>
<td>Are tasks performed in a very different way using the computer-based testing interface than they would be in the classroom or real world? Does this interfere with the transferal of skills and techniques from the learning environment to the assessment environment? Online testing tools – calculator, ruler, compass, text editing – or computer simulations within performance-based assessment that are unlike their real-world or classroom counterparts.</td>
</tr>
</tbody>
</table>

### Figure 2: Tool Usability Improvement

![Diagram of a floor plan showing a building's interior and exterior dimensions.](image-url)
Figure 3: Representing Z-axis Interactivity in Virtual Equipment - Titration
Figure 4: Representing Z-axis Interactivity in Virtual Equipment - Microscope
All My Babies Are Gone Now

From Loud and Clear
By Anna Quindlen

1 All my babies are gone now. I say this not in sorrow but in disbelief. I take great satisfaction in what I have today: three almost-adults, two taller than me, one closing in fast. Three people who read the same books I do and have learned not to be afraid of disagreeing with me in their opinion of them, who sometimes tell vulgar jokes that make me laugh until I choke and cry, who need razor blades and shower gel and privacy, who want to keep their doors closed more than I like. Who, miraculously, go to the bathroom, zip up their jackets, and move food from plate to mouth by themselves. Like the brick wall I bought for the bathroom with a rubber squeegee at its center, the baby is buried deep within each, barely discernible except through the unreliable haze of the past.

2 Everything in all the books I once poured over is finished for me now. Pseudology: T. Berry Brazelton, Dr. Spock. The ones on sibling rivalry and sleeping through the night and early childhood education, all grown obsolete. Along with Goodnight Moon and Where the Wild Things Are; they are battered, spotted, well used. But I suspect that if you flipped the pages, dust would rise like memories.

What those books taught me, finally, and what the woman on the playground taught me, and the well-meaning relations and the older parents at cocktail parties—what they taught me was that they couldn’t really teach me very much at all. Raising children is presented at first as a true-blue test, then becomes multiple choice, until finally, for crying out loud, you realize that it is an endless essay. No one knows anything. One child responds well to positive reinforcement, another can only be managed with a stern voice and a timeout. One boy is facile at killing, his brother at two. When my first child was born, parents were told to put baby to bed on his belly so that he would not choke on his own spit-up. By the time my last arrived, babies were put down on their backs because of research on sudden infant death syndrome.

As a new parent this ever-shifting certainty is terrifying, and then soothing. Eventually you must learn to trust yourself. Eventually the research will follow.

I remember fifteen years ago poring over one of Dr. Brazelton’s wonderful books on child development, in which he describes three different sorts of infants: average, quiet, and active. I was looking for a sub-quiet codfish (see: slug) for an eighteen-month-old who did not walk. Was

What is the primary difference between the two reading selections?

A. The essay is told from the perspective of a mother, while the poem is from that of a child.
B. The events in the story took place many years ago, while those in the poem are more recent.
C. The essay mentions fewer people than the poem does.
D. The issues explored in the essay are more important than those in the poem.
Clara

Clara is an 11th grader who attends a public school in a suburban area of Austin. She describes herself as a “pretty good” student but reports hoping to raise her GPA to a 3.75 by the end of the year in order to get into a good college.

Clara says she is usually one of the first students to finish state-standardized tests, which she finds relatively easy. However, when it is an essay-based test, she says she is consistently one the last students to finish. She begins with an outline using the scratch paper and then writes out a few key sentences that pop to her mind, which she wants to be sure to include somewhere in her essay. Then she writes an entire rough draft on the scratch paper before she transfers her essay to the test booklet.

Quotes from Clara:

“I just can’t get enough scratch paper! Too bad we can’t use index cards for making and organizing notes during the essay test.”

“I’m always scared I’m going to write too much or not enough, so I write the whole rough draft and then figure out how big to make my writing in order to make it two pages exactly. I don’t want the person scoring my essay to think that I didn’t make the most of the space available to me.”
### Figure 7: User Type Matrix

<table>
<thead>
<tr>
<th></th>
<th>Computer Novices</th>
<th>Computer Experts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Planners</strong></td>
<td>Transference to CBT expected to have minimal impact on writing method. Most planning work conducted on scratch paper and then transferred to the final delivery environment, whether using test booklet or online interface. Essay length indications, potentially using paper metaphors such as page length, important to this group. Fatigue from handwriting an essay twice is a factor for this group.</td>
<td>Transference to CBT expected to involve some planning steps transferred to the online writing space versus scratch paper. May move some Planners closer to the profile of Rewriters, since some Planners report different writing behaviors when using a computer. Essay length indications, such as used in other applications (character max or Word’s “print preview”) important to this group.</td>
</tr>
<tr>
<td><strong>Unmotivated</strong></td>
<td>Transference to CBT could encourage rewriting due to ease of editing, but risk of usability issues could further compromise performance by increasing frustration. Some expected correlation between this group and poor typing skills, which could result in a greater disadvantage for this group.</td>
<td>Unmotivated writers who are computer experts may have subject-specific motivation issues that affect ELA more than math and science assessments. Comfort with the computer could increase motivation in comparison to the paper-and-pencil test.</td>
</tr>
<tr>
<td><strong>Rewriters</strong></td>
<td>An essay item requiring hand-writing pushes some members of this group closer to the profile of a Planner, since some Rewriters report different writing behaviors when using a computer. Highly visible and easy-to-use editing tools very important to this group.</td>
<td>This group expresses the most excitement over computer-based essay items replacing hand-written essays on standardized tests. Some flow-and-rewrite behaviors may have been encouraged by learning writing skills with the use of the computer. Editing tools that are similar to those used in other computer-based environments very important to this group.</td>
</tr>
</tbody>
</table>
Figure 8: Design Submitted for Usability Testing

Write an essay about a critical decision you were forced to make.

Outline:
- Introduction: many difficult decisions, one stands out
- Background: my parents' divorce, my mom's move to CA
- Criteria
  - friends
  - school
  - my siblings' choices
- What I decided
- Why I changed my mind
- Conclusion: sometimes it's not about the decision you make, but about making the best of any decision you make and having the courage to admit that you may have made the wrong decision.

I have made many difficult decisions in my life and expect that many hard decisions are yet to come. However, as I stand back and examine those decisions that
## Figure 9: Comparability Study Data

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Original Interface</th>
<th>Revised Interface</th>
<th>Paper and Pencil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>49%</td>
<td>50%</td>
<td>47%</td>
</tr>
<tr>
<td>Female</td>
<td>51%</td>
<td>50%</td>
<td>53%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>43%</td>
<td>51%</td>
<td>55%</td>
</tr>
<tr>
<td>Afr-Amer.</td>
<td>16%</td>
<td>9%</td>
<td>7%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>39%</td>
<td>35%</td>
<td>36%</td>
</tr>
<tr>
<td>Other</td>
<td>2%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>Mean ELA Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC items</td>
<td>38.56</td>
<td>39.26</td>
<td>39.72</td>
</tr>
<tr>
<td>Essay</td>
<td>1.85</td>
<td>2.08</td>
<td>1.82</td>
</tr>
<tr>
<td>Percentage at Each Essay Score Point</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>9%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>1</td>
<td>24%</td>
<td>20%</td>
<td>28%</td>
</tr>
<tr>
<td>2</td>
<td>44%</td>
<td>44%</td>
<td>51%</td>
</tr>
<tr>
<td>3</td>
<td>21%</td>
<td>26%</td>
<td>15%</td>
</tr>
<tr>
<td>4</td>
<td>3%</td>
<td>6%</td>
<td>1%</td>
</tr>
</tbody>
</table>
Appendix A: Information Survey

1. What grade will you be entering in the fall?
   
   
   10 – 1
   11 – 3
   12 – 5
   Graduated – 1

2. What is the name of your high school?
   American Youth Works
   Anderson High
   Gonzalo Garza Independence High School (2 students)
   LBJ Liberal Arts and Sciences Academy
   McNeil (3 students)
   McCallum
   Out-of-state

3. What kind of student do you consider yourself?
   
   Fair – 0
   Average – 3
   Good – 6
   Exceptional – 1

4. Have you taken tests on the computer before?
   
   Yes – 8
   No – 2

   If you answered yes, what type of tests?
   Career, benchmark, class tests, IQ, Compass, college placement, MC, aptitude

   Did they include writing an essay? Yes – 2       No – 6

5. Which best describes your level of comfort with computers?
   
   I’m an expert computer user. – 3
   I’m fairly comfortable with computers – 7
   I only rarely use computers – 0

6. Do you have a computer at home that you can use for school work?
   
   Yes – 10
   No – 0

7. How often do you use Microsoft Word or some other word processing application?
   
   Often – 9
   Sometimes – 1
   Rarely – 0
   Never – 0
8. How would you rate your typing skills?
   - Fast and error-free – 1
   - Pretty good – 6
   - Average – 2
   - Lousy – 1

9. Do your teachers provide instruction on writing techniques using a computer to write essays?
   - Yes – 1
   - Only a little bit – 7
   - Never – 2

10. Do you prefer to write essays for school using the computer or using pen/pencil and paper?
    - Computer – 6
    - Pen/pencil and paper – 4

11. For each of the following writing tasks, indicate whether your preference is to perform this task using the computer or pen/pencil and paper:
    - Jotting down notes:  
      - Computer – 3  
      - Pen/pencil and paper – 7
    - Outlining:  
      - Computer – 2  
      - Pen/pencil and paper – 6
    - Rough draft:  
      - Computer – 6  
      - Pen/pencil and paper – 3
    - Final draft:  
      - Computer – 10  
      - Pen/pencil and paper – 0

12. Do you use graphic organizers when organizing your thoughts to write an essay?
    - Yes – 3
    - No – 7

13. Do you regularly engage in other kinds of typed writing? (check all that apply)
    - Text messaging using a cell phone or PDA – 8
    - Instant messaging using a computer – 8
    - Blogging – 3
    - Participating in a chat room or discussion board – 3
    - Other: work – 1

15. For which types of tests would you prefer to take online? (Check all that apply – or none.)
    - Multiple-choice tests – 8
    - Tests involving reading one or more 1-2 page reading passages – 5
    - Tests involving writing an essay – 6
Appendix B: Post-test Ease-of-use Questionnaire

Students rated the screen as fairly easy or a little confusing.

How easy to use was this essay-writing screen?

<table>
<thead>
<tr>
<th>Rating</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Easy</td>
<td>1</td>
</tr>
<tr>
<td>Fairly Easy</td>
<td>5</td>
</tr>
<tr>
<td>A Little Confusing</td>
<td>4</td>
</tr>
<tr>
<td>Not at all Easy</td>
<td>0</td>
</tr>
</tbody>
</table>

All students rated cut/copy/paste as either very easy or fairly easy.

How easy to use was the cut, copy, and paste functionality?

<table>
<thead>
<tr>
<th>Rating</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Easy</td>
<td>5</td>
</tr>
<tr>
<td>Fairly Easy</td>
<td>5</td>
</tr>
<tr>
<td>A Little Confusing</td>
<td>0</td>
</tr>
<tr>
<td>Not at all Easy</td>
<td>0</td>
</tr>
</tbody>
</table>

Only one student reported not using cut, copy, and paste in their computer usage.

Do you use cut, copy, and paste in other computer applications?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>I rarely use the computer for text editing</td>
<td>1</td>
</tr>
</tbody>
</table>

Most students use CTRL short-cuts, with two students right clicking in order to cut and paste.

If you answered “yes” to the question above, what is the most typical way for you access such functions?

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Through a pull-down menu</td>
<td>0</td>
</tr>
<tr>
<td>By right-clicking the mouse</td>
<td>2</td>
</tr>
<tr>
<td>By using Control C, V, and X</td>
<td>0</td>
</tr>
<tr>
<td>Other: N/A</td>
<td>2</td>
</tr>
</tbody>
</table>

While writing an essay for a TAKS test, would you use cut, copy, and paste functionality?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most definitely, can’t imagine a test without it</td>
<td>3</td>
</tr>
<tr>
<td>I would use it frequently</td>
<td>3</td>
</tr>
<tr>
<td>I might use it occasionally</td>
<td>2</td>
</tr>
<tr>
<td>I would not use it. Delete and backspace on the keyboard are enough for me</td>
<td>2</td>
</tr>
</tbody>
</table>

Which of the following best describes how you might be feeling during the writing of an essay on a test?

<table>
<thead>
<tr>
<th>Feeling</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I’m most interested to know how much I have written and whether it is enough</td>
<td>4</td>
</tr>
<tr>
<td>I’m most concerned that I’ll run out of space to write.</td>
<td>4</td>
</tr>
<tr>
<td>I’m not concerned about the length of my essay while writing it.</td>
<td>1</td>
</tr>
<tr>
<td>I have a good sense of how much I’ve written without having to look.</td>
<td>1</td>
</tr>
</tbody>
</table>

Space remaining preview seems to be a little confusing most likely due to counting characters, not lines.

How easy to use was the preview of “space remaining”?

<table>
<thead>
<tr>
<th>Rating</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Easy</td>
<td>1</td>
</tr>
<tr>
<td>Fairly Easy</td>
<td>6</td>
</tr>
<tr>
<td>A Little Confusing</td>
<td>3</td>
</tr>
<tr>
<td>Not at all Easy</td>
<td>0</td>
</tr>
</tbody>
</table>

Regardless of the slight confusion, all students said they thought it was helpful.

Which of the following best describes your reaction to the “space remaining” preview?

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I didn’t know what it was.</td>
<td>0</td>
</tr>
<tr>
<td>It made me nervous and overly aware of how much I had written.</td>
<td>0</td>
</tr>
<tr>
<td>I thought it was helpful.</td>
<td>10</td>
</tr>
<tr>
<td>I didn’t use it, but it didn’t bother me.</td>
<td>0</td>
</tr>
</tbody>
</table>

If you used the undo button two times in a row after doing two cut operations, what would you expect to happen? Would it undo the cut and then redo it? Or would it undo both cuts?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Redo</td>
<td>5</td>
</tr>
<tr>
<td>5 Multiple Undos</td>
<td>5</td>
</tr>
</tbody>
</table>

Rank the following in the order of importance to you (1 = most important):

<table>
<thead>
<tr>
<th>Rank</th>
<th>Description</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The ability to cut and paste</td>
<td>33</td>
</tr>
<tr>
<td>2-tie</td>
<td>The ability to copy</td>
<td>45</td>
</tr>
<tr>
<td>4.</td>
<td>The ability to undo</td>
<td>42</td>
</tr>
<tr>
<td>3.</td>
<td>The ability to see what portion of two pages I’ve filled</td>
<td>35</td>
</tr>
<tr>
<td>6.</td>
<td>The ability to highlight the text in the writing prompt</td>
<td>57</td>
</tr>
<tr>
<td>1.</td>
<td>The ability to see the writing prompt without scrolling</td>
<td>30</td>
</tr>
<tr>
<td>1.</td>
<td>The ability to see as much of my essay at once without scrolling</td>
<td>33</td>
</tr>
</tbody>
</table>

What other functionality would you like to see added to an essay-writing interface?

Roll-over descriptions for buttons, ability to insert annotated reminders, highlighting in your essay, pre-writing screen, ability to bullet, italic/bold formatting, separate redo, revision history, centering