

Quiet Interfaces that Help Students Think

Sharon Oviatt

Computer Science Dept.
Oregon Health & Science University
20,000 NW Walker Road
Beaverton, Oregon, 97006 USA
Tel: 503-679-8593
oviatt@csee.ogi.edu

Alex Arthur & Julia Cohen[§]

Computer Science Dept.
Oregon Health & Science University
20,000 NW Walker Road
Beaverton, Oregon, 97006 USA
alex.arthur@csee.ogi.edu
[§]Dartmouth College

ABSTRACT

As technical as we have become, modern computing has not permeated many important areas of our lives, including mathematics education which still involves pencil and paper. In the present study, twenty high school geometry students varying in ability from low to high participated in a comparative assessment of math problem solving using existing pencil and paper work practice (PP), and three different interfaces: an Anoto-based digital stylus and paper interface (DP), pen tablet interface (PT), and graphical tablet interface (GT). Cognitive Load Theory correctly predicted that as interfaces departed more from familiar work practice (GT > PT > DP), students would experience greater cognitive load such that performance would deteriorate in speed, attentional focus, meta-cognitive control, correctness of problem solutions, and memory. In addition, low-performing students experienced elevated cognitive load, with the more challenging interfaces (GT, PT) disrupting their performance disproportionately more than higher performers. The present results indicate that Cognitive Load Theory provides a coherent and powerful basis for predicting the rank ordering of users' performance by type of interface. In the future, new interfaces for areas like education and mobile computing could benefit from designs that minimize users' load so performance is more adequately supported.

ACM Classification & General Terms:

H.5.2 [Information Interfaces and Presentation]: User Interfaces—user-centered design, interaction styles, evaluation/methodology, input devices and strategies, prototyping.

Keywords: Pen-based interfaces, Performance Metrics, E-Learning and Education, Universal (or Disability Access), Input and Interaction Technologies, Handheld and Mobile Computing, Tangible UIs, Communications

INTRODUCTION

As technical as we have become, modern computing still has not permeated many important areas of our lives, including ones that we value highly like mathematics education. High school mathematics education still depends on pencil and paper, and perhaps for good reason. Other paradoxes dominate recent news, presenting challenges for the future of computing. For example, while the gap between white and black Americans in access to the internet has narrowed by 6% during the past 8 years, according to Pew national survey data [13], during the same time period Washington Assessment of Student Learning (WASL) test results revealed a substantial widening of the achievement gap in mathematics between white and minority students—from 17.9 to 31.3% between white and black 7th graders, 17.3 to 29.3% between white and Latino students, and 17.1 to 23.8% between white and American Indian students [5]. At present, only one-quarter of minority students in Washington State pass basic achievement tests in mathematics. New technologies clearly are needed that can permeate, and unite rather than divide.

Related Literature on Pen-based Interfaces & Education

Among the attractive things about pen-based interfaces are their compatibility with existing work practice in education and many other domains [1,2,7,9,10], their suitability for mobile use and collaboration [12,16], and also their support for a broad range of expressive input in different representational systems (i.e., linguistic, numeric, symbolic, diagrammatic). In contrast, graphical user interfaces (GUIs) provide good support for linguistic and numeric content, but user input involving symbols or diagrams is more poorly supported or missing altogether. Complex problem solving in domains like mathematics requires input fluency in all four representational systems, and also flexible translation among them— for example, from word problems, to diagrams, to algebraic formulas. As a result, current work practice for high school geometry and mathematics education still involves pencil and paper— that is, learning without computational support at all.

Research on pen tablet interfaces has pursued recognition-based interfaces for freehand mathematical notations and equations [10,18], rather than handwriting per se [17]. In addition, considerable work in this area has simply involved sharing of ink between instructors and students in classroom lecture contexts, for example annotations and

marks on distributed Powerpoint slides [1,2]. Recent empirical work also has explored whether pen-based, spoken, or multimodal input are faster for entering equations or preferred by users, compared with state-of-the-art GUIs that include an equation editor with a hierarchically-organized array of 50-100 symbols and templates [4]. When user input was examined for entering equations varying in difficulty on a tablet computer, and decoupled from system recognition processing, adults performed faster with all of the non-GUI interfaces and preferred writing math content [4]. However, research has yet to examine interface support for math *problem solving* rather than mechanical entry tasks, or for newer Anoto-based digital stylus and paper interfaces rather than the pen tablet interface. Research also has not yet probed how well different interfaces support math education for a range of low- and high-performing students. Given present hardcopy work practice, research also could benefit by investigating what the best route may be for transitioning math education to computational support so students' learning performance can be enhanced.

Related Literature on Cognitive Load & Learning

When learning new intellectual tasks, the cognitive effort required on the part of learners tends to be high and to fluctuate substantially. Over the past decade, Cognitive Load Theory (CLT) has maintained that in the process of learning and developing expertise it is easier to acquire new schemas and effectively automate them if instructional methods minimize demands on a students' working memory, thereby reducing their cognitive load [6,14,15,19]. Advocates of this theory typically assess the "extraneous complexity" associated with instructional methods or interfaces separately from the "intrinsic complexity" associated with a student's primary learning task, which is done by comparing performance indices of cognitive load as students use different methods. Basically, CLT has focused on the design of instructional materials with the aim of decreasing extraneous cognitive load so students' available cognitive resources can be devoted to their learning tasks.

In a series of education experiments, it was revealed that a dual-mode presentation format supported expansion of working memory and better problem solving in geometry tasks than a single visual mode [14]. Essentially, it was shown that an integrated multimodal presentation format can expand the size of available working memory in a manner that expedites classroom instruction. Other potential ways to reduce students' extraneous cognitive load and expand working memory during learning could include designing interfaces that are more similar with existing work practice, that support the representational systems required by the domain, and that focus students' attention on the math operations required for problem solution rather than on gratuitous features that may entertain but also distract (i.e., minimizing unnecessary complexity).

Specific Goals and Predictions of This Study

The general goal of this research was to prototype promising new interface directions for math education, and to

compare their ability to support students' performance during math problem solving activities. More specifically, this study compared students' performance on geometry problem solving tasks while using pencil and paper, which is existing work practice for high school math education, with their use of three alternative interface prototypes: (1) an Anoto-based digital stylus and paper interface [3], (2) a pen-based tablet interface, and (3) a graphical tablet interface that included a stylus and simplified equation editor. As interface prototypes depart more from familiar work practice ($GT > PT > DP$), it was conjectured that students would experience greater extraneous cognitive load such that performance would deteriorate. The present study also assessed whether Cognitive Load Theory (CLT) could correctly predict the rank ordering of interfaces by their ability to support performance. Performance was assessed using convergent metrics of dynamic information processing, including speed, attentional focus, meta-cognitive control, correctness of solutions, and memory.

A second goal of the present work was to present students with geometry tasks varying in intrinsic complexity from low to very high, which would provide a forcing function for evaluating how well different interfaces support student performance during a realistic range of tasks. Cognitive Load Theory predicts that as task difficulty increases, cognitive load also rises and combines with extraneous sources of load (e.g., interface-related) to degrade performance. A third and related goal was to prototype and evaluate interfaces with students ranging from low- to high-performers in math, so that new educational interfaces can be designed that are supportive of learning for all students. Cognitive Load Theory predicts that lower-performing students with less well consolidated math expertise would *experience* more cognitive load than higher-performing students. For such students, more challenging interfaces or difficult tasks would disrupt their performance disproportionately more than the higher performing students.

METHODS

Participants

Twenty public high school students who had recently completed a geometry class were included in the study as paid volunteers. All students had been using paper and pencil materials in their high school math classes, had expressed an interest in technology, and were experienced users of graphical user interfaces with keyboard and mouse input.

According to both (1) teacher records on students' classroom grades in geometry and (2) students' percentage of correct math problem solutions in the present study, approximately half of the students were classified as high-performing and the other half low-to-moderate performing. Thirteen of the volunteers were female and seven male. All students were native speakers of English, but varied in ethnic (e.g., Russian, Vietnamese, Indian, Scottish) and racial backgrounds (e.g., Caucasian, Black, Asian).

Math Problems and Difficulty Levels

Based on consultation with high school teachers, textbooks, and classroom observations, math problems were selected for inclusion in the study that students had just finished learning about in a high school geometry class. Teacher records of average student test performance on specific problems (i.e., percent correct on midterms) was used as an initial basis for classifying problems by difficulty level. Pilot testing then was conducted to confirm difficulty level classifications. Table 1 illustrates typical examples of math problems representing the different difficulty levels.

All of the math problems were *word problems* that required translation from linguistic information into symbolic and digit-based information to solve them. Since the majority were spatially-oriented geometry problems, diagrams also were helpful in solving them. In short, successful completion of the math problems required complex problem solving using all four representational systems (i.e., linguistic, symbolic, numeric, and diagrammatic), as well as translating among them. These characteristics permitted testing the ability of different interfaces' to support flexibly expressive communication patterns, which is required for extended problem solving in domains like geometry. The number, format, and type of information elements varied in problems representing different difficulty levels, such that harder problems involved more steps to solution, information presented in different formats (e.g., integers versus ratios), incidental information not required for solution, and so forth. Beyond this, all problems maintained a natural narrative order of presentation (i.e., *end* rather than *start unknown*), were contextualized real-world problems, and used familiar terminology. In case students had forgotten, any terms or equations needed to complete a problem were supplied for reference, as shown in Figure 1 (bottom left).

Difficulty	Geometry Problem Content
Low	A hamster wheel is 10 cm in diameter. If the hamster runs so the wheel turns 100 revolutions, then how far did the hamster run?
Moderate	Josh made a party hat shaped like a triangle. If it has a base of 16" and a height of 7", what is the surface area of the front of it?
High	Thad has a new silver kite shaped like a rhombus. If the diagonals are 7 cm wide and 12 cm long, what is the surface area of his kite?
Very High	The Transamerica building in San Francisco is a pyramid 800' tall, with a square base 149' on each side. What is its volume in cubic feet?

Table 1: Examples of geometry problems representing different difficulty levels

Procedure

Students were tested in pairs, and were given instructions and practice together. They were told: "We are getting ready for a summer camp that will be teaching geometry to younger high school students who don't understand math as well as you. We would like you to help us get our materials ready by trying out some different computer interfaces while you solve geometry problems. When you're done, I'll ask you a few questions about the interfaces you used, and what you think we can do to make sure kids at camp learn math successfully and have a great experience."

The student volunteers then were shown the four different sets of materials that they would be using to solve problems, including: (1) standard pencil and paper, (2) digital stylus and paper (i.e., Nokia stylus with Anoto-based paper technology), (3) tablet computer with stylus input, and (4) tablet computer with keyboard, mouse, and stylus input, which was enhanced with a simplified MathType equation editor containing 11 symbols for geometry that were not on the keyboard (e.g., square roots, powers).

For all four conditions, each problem set was presented on a Toshiba Portege laptop screen, as shown in Figure 1, which included the main word problem (top) along with any terms or equations required to complete the problem (lower left). In the two paper-based conditions students simply read the problem on the computer screen but did their work on paper. In the two tablet-based conditions they entered their work on the computer, using Windows Journal for the pen tablet condition, and either MathType or Windows Journal (i.e. using a stylus) for the mixed graphical tablet interface. Figure 1 shows the graphical tablet interface condition, with MathType (left side) and Windows Journal (right side) both open. In the pen tablet condition, Windows Journal was the only input area open, and in the two paper conditions the middle of the screen shown in Figure 1 was left blank. Like the paper conditions, the Windows Journal and MathType input environments simply provided blank space on which students could compose their input. In all conditions, students were told they could use their calculator, and they were free to use their materials any way they liked. With the graphical tablet interface, students were told they could use the keyboard and equation editor or pen input however they wished.

For each of the three computer interfaces, students were given instructions on how it worked and allowed to practice until they were familiar and had no more questions. For example, with the pen tablet interface they were oriented to the basics of how to ink, erase, undo/redo, move and resize their input, and how to scroll down to get more writing space. With the graphical tablet interface, they were shown the symbols in the equation editor and given practice using them all in a problem. They also were shown how to undo/redo and other basic capabilities, although most aspects of the graphical interface already were familiar. Beyond basic orientation to each interface, students were told to work at their own pace, and that they should just concentrate on solving each problem. If they couldn't

complete a problem, they were instructed to go on to the next.

Afterwards, each student worked individually with their own computer display and materials during the main test session. They completed 16 math problems, including 4 problems apiece in each of the 4 conditions. In all conditions, when students were done with each problem they were instructed to click the “Submit” button (Figure 1, lower right), which then collected auto-timings from the beginning of problem presentation until submission of the answer. While working on their problems, each student wore a close-talking headset that recorded their digital speech during a think-aloud protocol. They were told: “While working on your math problems, I’m going to ask you to talk out loud about anything you’re thinking as you work. If you’re thinking about the interface you’re using or the math problem you’re working on, just go ahead & say it! No matter how miscellaneous, all your thoughts will help us develop better materials for the kids at camp.”

america building? 900 ft. 700 ft. 800 ft.”). Afterwards, students each were asked to fill out a questionnaire in which they were asked what they liked and disliked about the four interfaces, and also which interfaces helped them focus best on solving their math problems. Among the questions they were asked was, “If you had to take an AP math exam and *perform at your best*, how would you *rank order your preference* to use each interface or set of materials? (1 most preferred, 4 least preferred).

Research Design

This study involved a mixed factorial experimental design, with the main within-subject independent factors including: (1) *Type of Interface*: Paper and pencil hardcopy materials (PP), Digital stylus & paper interface (DP), Pen tablet interface (PT), and graphical tablet interface (GT), and (2) *Math Problem Difficulty Level*: Low, Moderate, High, and Very High. Each student completed a set of four problems per condition, which increased progressively in difficulty from low to very high difficulty. The specific math content

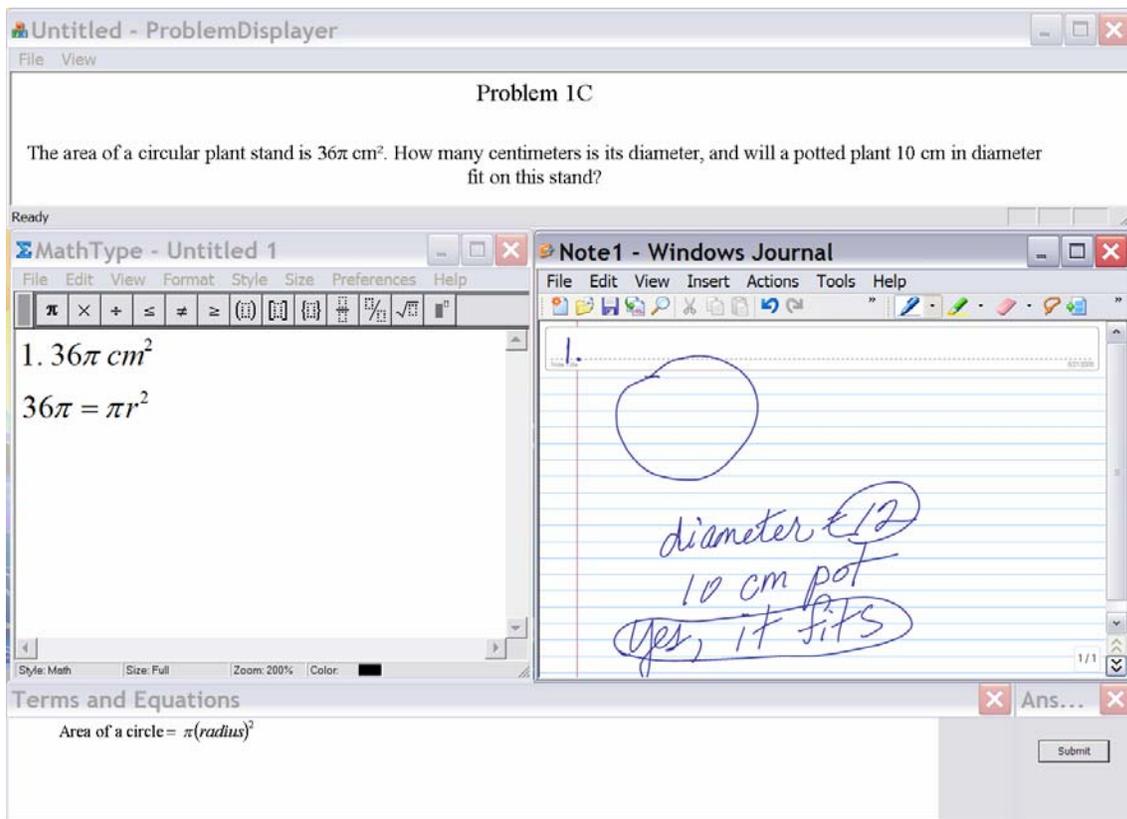


Figure 1: Student interface used to display math problems.

Following this main test session, students also each completed 16 questions on a forced-choice memory test, in which they were asked to recall both task-critical and incidental information from the high and very high difficulty math problems that they had just finished solving. Four questions apiece involved content students had seen while working in each interface condition. Each question required selecting the answer from among three alternatives that were close distractors (e.g., “Who owned the new silver kite? Thad Thomas Ted”; “How tall is the Trans-

of different problem sets that were paired with the interface conditions and also the order of presentation of the interface conditions were completely counterbalanced. The order of math content asked about during the memory test also was rotated to match that received during the session.

The main between-subject factor in this study was: (3) *Student Performance Level*: High, Low. Eleven students were classified as high performing, and nine as low performing.

Dependent Measures and Coding

The digital recordings of speech were transcribed from all sessions, and automatic timings were collected of student solution times for all problems. In addition, data coding was performed on the measures outlined below.

Time-to-Complete Problem Solutions

Timings were automatically collected from the beginning of each problem presentation until the student clicked “submit” to indicate completion of their answer. Total time to solution then was computed for each problem, and summarized for each condition. Time to solution also was computed for the subset of problems that students solved correctly, excising 90 cases where the solution was completely incorrect or not attempted at all.

Correctness of Problem Solutions

All student problems were scored for correctness of solution, with partial credit given when one or more of the following problems was evident: the final numeric answer wasn’t fully computed, only one part of a two-part question was answered correctly, labeling in units was missing from the final answer, or decimal places were incorrect or missing. The percent of correct solutions per condition then was totaled for each student.

Focus of Attention during Problem Solving

Based on the think-aloud transcriptions, each of students’ utterances were scored during their problem solutions as: (1) a *read comment*, (2) *low-level math comment* during a localized solution step (e.g., “So, it’s just 15 times 12”), (3) *high-level math insight* of a meta-cognitive nature (e.g., “Oh, it’s a volume problem”; “That’s just useless information”), (4) an *interface-related comment* (e.g., “Now, I’m gonna lasso it to make it smaller so I have more room to work”), with *negative interface comments* also scored separately (e.g., “Dang it, I mis-clicked”; “Okay, the lasso didn’t work”), and (5) *other off-task comments* (e.g., “Adrienne’s leaving, okay”). Utterances coded as high-level math comments reflected a meta-cognitive level of awareness about solving the problem, including what type of problem it was, how difficult/easy, what type of information/strategies were needed to solve the problem (e.g., translations or conversions), spatial relations among objects represented, diagnostics about the presence or nature of errors, and self-reflections by the student about their ability to solve the problem. The total number of each type of comment was summarized for each condition, as well as the ratio per total comments.

Memory for Problem Content

The total percentage of forced-choice memory questions that students recalled correctly was scored and summarized for each condition.

Self-Reported Interface Preferences

Each student’s rank-order preference for the 4 interfaces was scored from their written questionnaire, with 1 for the most preferred and 4 for least. Average student rank-order interface preferences then were summarized for each condition. Qualitative feedback also was summarized about

students’ reasons for their rank-order preferences, as well as what students liked and disliked about each interface.

Reliability

Second scoring was conducted between two independent coders on all of the scored data for correctness of problem solutions, with a 100% match obtained. Scoring of students’ focus of attention during the think-aloud also was performed by two independent raters for all data, with minor departures then resolved.

RESULTS

Data were available on 320 problem solutions and 20 hours of speech recordings during the think-aloud. For all dependent measures reported below, analyses were based on a minimum of 16 and maximum of 20 students’ data.

Time-to-Complete Problem Solutions

The average time that it took students to complete math problems while using the different interfaces is shown in Figure 2. When using paper and pencil and the digital stylus interface, students’ average speed for all data was 117.96 seconds and 115.36 seconds, respectively, not a significant difference by paired t test, $t < 1$. Likewise, when using the pen tablet and graphical tablet interfaces, they averaged 132.36 and 136.17 seconds, respectively, also not a significant difference, $t < 1$. When using the two paper-based interfaces students’ solution times averaged 116.7 seconds, compared with 134.3 seconds with the tablet-based interfaces, a significant difference by paired t test, $t = 2.55$ ($df = 15$), $p < .015$, one-tailed. In short, working on the tablet interfaces averaged 16% more time than solving the same problems on the paper-based interfaces. These findings were replicated when solution times were analyzed for problems not involving errors, and on log transformed data.

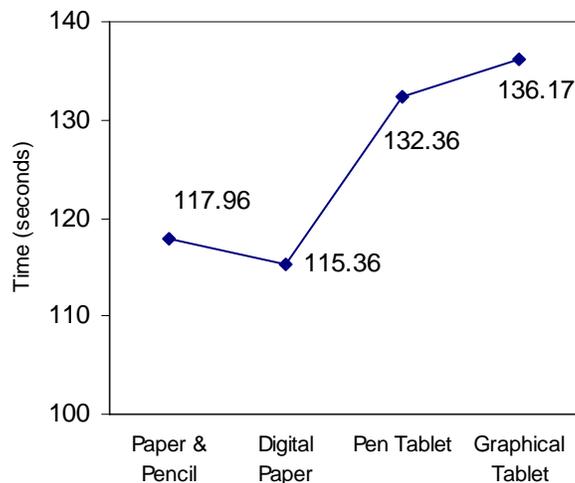


Figure 2: Average total time to complete individual math solutions using different interfaces

Correctness of Problem Solutions

Overall, students working with paper and pencil averaged 1.16 errors per condition on math problems, compared with

1.15 when using the digital stylus, 1.25 with the pen tablet, and 1.43 with the graphical tablet interface. These differences in average error rates are illustrated in Figure 3. They represent an overall drop in students' percent correct math solutions from 70.9% and 71.3% when using pencil and paper and the digital stylus interface, to 68.8% and 64.4% with the pen and graphical tablet interfaces.

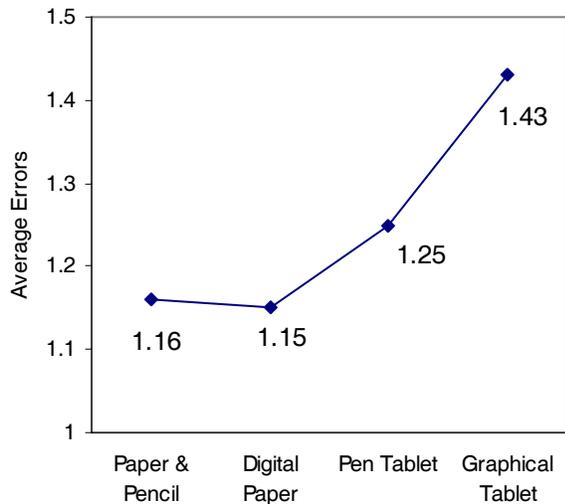


Figure 3: Average number of math errors per condition in different interfaces

As shown in Figure 4, however, the high-performing students' errors changed relatively little in the different interface conditions (.93 for pencil and paper, .71 for the pen-based interfaces, and .59 for graphical tablet interface), although the low-performing students' errors for the same conditions increased from 1.44, to 1.81, to 2.44. That is, at the same time that high performers shifted from 77% of math problems solved correctly with pencil and paper to 82% and 85% when using pen-based and graphical interfaces, the low performers dropped from 64% with pencil and paper to 55% with pen-based interfaces, and then more sharply to 39% correct with a graphical tablet interface.

An analysis of the difference in errors between pencil and paper and the mean of the three computer interfaces indicated a significant divergence in errors between low- and high-performing students as a function of shifting from existing work practice to using any of the computer interfaces, independent $t = 7.87$ ($df = 18$), $p < .001$, one-tailed. The divergence in error rates between low and high performing students was only marginally significant when shifting from pencil and paper to the pen-based interfaces, $t = 1.69$ ($df = 18$), $p < .06$ one-tailed. However, the difference in average errors between the pen-based interfaces and the graphical tablet interface represented a significant divergence between low- and high-performing students, independent $t = 1.81$ ($df = 18$), $p < .045$, one-tailed. In short, when the low-performing students shifted from pencil and paper to using a pen-based interface or graphical tablet interface they made 25.7% and 69.4% more math errors, respectively, whereas in contrast the high-

performing students made 23.7% and 36.6% fewer errors in the same context.

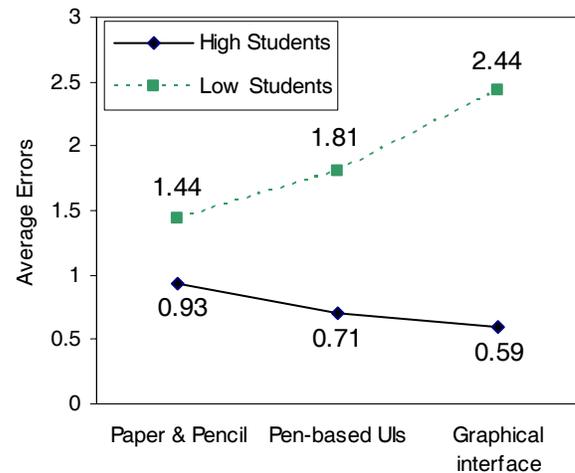


Figure 4: Difference between low- and high-performing students in math errors per condition as a function of pen-based (PP, DP) versus graphical tablet (GT) interfaces.

As corroboration that the task difficulty levels used in this study resulted in effective increments in the degree of difficulty experienced by students, analyses confirmed that the average number of errors increased from .88 per student per condition on the low difficulty problems, to .85 on the moderate ones, 1.29 on high difficulty, and 1.98 on very high difficulty problems. These differences in performance errors basically validated and anchored the difficulty levels. Low-performing students also made uniformly more errors than high-performing students at all difficulty levels, independent $t = 6.54$, $df = 18$, $p < .001$, one-tailed.

Focus of Attention during Problem Solving

Based on the think-aloud content collected as students worked, the total number of all such comments did not differ between interface conditions, paired $t_s < 1$. Likewise, there was no significant difference in total think-aloud utterances between low- and high-performing students, independent $t = 1.24$ ($df = 16$), N.S. The total number of all math comments and low-level math comments also did not differ between interface conditions, paired $t_s < 1$. However, as illustrated in Figure 5, the percent of students' high-level math comments of the total declined to 8.9% in the graphical tablet interface condition, compared with 17.9% while using paper and pencil, 17.3% with the digital stylus interface, and 14.8% with the pen tablet. These percentages did not differ significantly between paper and pencil, digital stylus, and pen tablet conditions, paired $t_s < 1$, although high-level math comments when using the graphical tablet interface were significantly fewer than with either pencil and paper, paired $t = 3.26$ ($df = 15$), $p < .0025$ one-tailed, or the digital stylus interface, paired $t = 1.75$ ($df = 15$), $p < .05$ one-tailed, or the pen tablet interface, paired $t = 3.20$ ($df = 15$), $p < .003$ one-tailed. This represented a substantial 50.3% drop in high-level math comments by students

while solving problems in the graphical tablet interface condition, compared with such comments when using pencil and paper. Low-performing students experienced a sharper drop (58.6%) in high-level math comments than did high-performing students (42%). Since the total number of think-aloud comments did not differ between interfaces, analyses based on ratios and total numbers replicated.

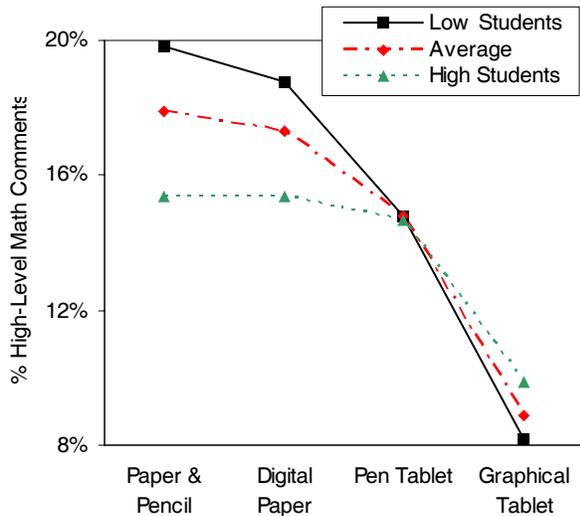


Figure 5: Average percent of high-level math comments of the total for low- and high-performing students when using different interfaces

Think-aloud comments about interfaces used, including negative interface comments, were evident in 94% and 81% of the students, respectively. The frequency of both types of comment, which reflected distractions from students' ability to focus on solving their math problems, did not differ significantly between low- and high-performing students. However, interface-related comments increased from an average of .69 per session when using pencil and paper, to 1.38 with digital stylus and paper, to 2.94 with the pen tablet, and a high of 5.25 when using the graphical tablet interface. As shown in Figure 6, this represented 1.5%, 3.0%, 5.8% and 10.7% of all think-aloud comments, respectively.

There was no significant difference in students' frequency of commenting about the interface between the digital stylus and pencil and paper conditions, Wilcoxon Signed Ranks test, $z = 1.47$, N.S., or between the pen tablet and graphical tablet conditions, Wilcoxon Signed Ranks test, $z < 1$. However, the frequency of interface comments increased significantly between the paper-based and tablet-based interfaces, which averaged 1.03 and 4.09 comments per session (297% increase), Wilcoxon Signed Ranks test, $z = 3.42$, $p < .0005$, one-tailed. More specific comparisons also confirmed that UI comments were substantially higher in both the pen tablet interface and graphical tablet interfaces than when using paper and pencil, Wilcoxon Signed Ranks tests, $z = 2.68$, $p < .0035$ (one-tailed), and $z = 2.84$, $p < .0025$, (one-tailed), respectively. They also were higher

in both the pen tablet and graphical tablet interfaces than when using the digital stylus and paper, Wilcoxon Signed Ranks tests, $z = 2.51$, $p < .006$ (one-tailed), and $z = 2.60$, $p < .0045$ (one-tailed). Compared with pencil and paper, interface comments increased a substantial 326% when using the pen tablet, and 661% with the graphical tablet interface.

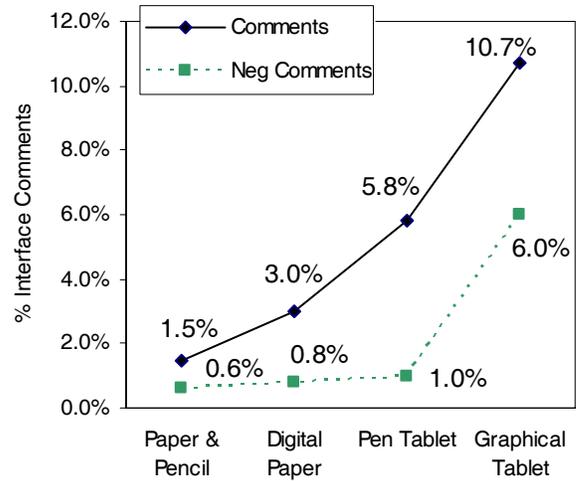


Figure 6: Average percent of interface comments and negative interface comments of the total when using different interfaces

Negative interface comments averaged just .25, .38, and .50 during the pencil and paper, digital stylus, and pen tablet interfaces, respectively (0.6%, 0.8% and 1.0% of all think-aloud comments, as illustrated in Figure 6), but increased to 2.94 (6.0%) when using the graphical tablet interface. This represented a 1,076% increase over the negative comments that occurred when students used pencil and paper. Wilcoxon Signed Ranks tests confirmed that the pencil and paper, digital stylus, and pen tablet conditions did not yield any significant differences in negative interface comments, $z_s < 1$. However, there were significantly more negative interface comments when students used the graphical tablet condition than with pencil and paper, Wilcoxon Signed Ranks test, $z = 2.38$, $p < .009$, one-tailed, or when using the digital stylus interface, $z = 2.53$, $p < .006$, one-tailed, or with the pen tablet, $z = 2.28$, $p < .015$, one-tailed.

Memory for Problem Content

The percentage of content that students recalled correctly from the math problems they just finished solving also varied when using the different interfaces, as illustrated in Figure 7. There were no significant differences for students overall or for high-performing students in memory for math content as a function of the interface, but low-performing students did differ in their ability to remember information depending on the interface they had used. Low-performing students did not differ significantly in their ability to remember math information after using paper and pencil versus the digital stylus interface, Wilcoxon paired sign test, $z < 1$, N.S. However, they were significantly less able to retain information when using the pen

tablet interface than paper and pencil, Wilcoxon paired sign test, $z = 1.73$, $p < .045$, one-tailed. They also were less able to retain math information when using the graphical tablet interface, compared with paper and pencil, Wilcoxon Signed Ranks test, $z = 1.89$, $p < .03$, one-tailed. When memory was examined separately for the two student subgroups, the high-performing students did not demonstrate any significant difference due to the interface, $z < 1$ N.S., but the low-performing students were significantly less able to retain information immediately after using the two tablet-based interfaces, compared with the paper-based interfaces, Wilcoxon Signed Ranks test, $z = 2.07$, $p < .02$, one-tailed. This represented a 12% drop in their memory for recently seen math content specifically after using the tablet interfaces.

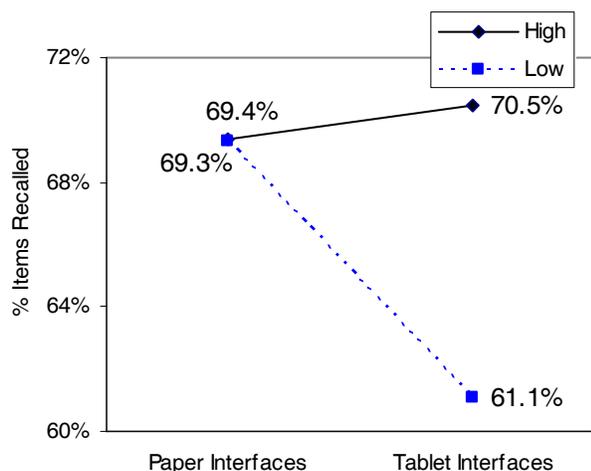


Figure 7: Percent of items recalled correctly by low- and high-performing students in the paper- (PP, DP) versus tablet-based (PT, GT) interfaces

Self-Reported Interface Preferences

Students' overall preference rankings were 1.93 and 1.85 for paper and pencil and the pen tablet interface, respectively, which was not a significant difference by Wilcoxon Signed Ranks test, $z < 1$, N.S. However, paper and pencil was more attractive to students than the digital stylus interface (average rank 2.63), Wilcoxon Signed Ranks test, $z = 2.09$, $p < .02$, one-tailed. It also was more attractive to students than the graphical tablet interface (average rank 3.60), Wilcoxon Signed Ranks test, $z = 3.57$, $p < .001$, one-tailed. Likewise, the digital stylus interface was preferred over the graphical tablet interface, $z = 2.86$, $p < .002$, one-tailed.

Further analysis revealed that 100% of the high performers ranked the paper-based interfaces as preferred over the tablet interfaces if they had to take an AP exam and perform their best (Wilcoxon Signed Ranks test, $z = 2.25$, $p < .015$, one-tailed). In contrast, 62.5% of the low performers actually ranked the tablet-based interfaces as preferred over the paper interfaces in the same circumstance, a reversal illustrated in Table 2. A Wilcoxon Signed Ranks test confirmed that low-performing students did not have a similar

preference for the paper interfaces over the tablet ones, $z < 1$.

In short, the data revealed a reversal in interface preferences between the student subgroups, which may have reflected different beliefs about how well they expected the paper versus tablet interfaces would support their math performance. This difference also revealed a *performance-preference paradox* (illustrated in Table 2), which was that students with low math competence whose performance degraded most when using the tablet interfaces nonetheless preferred them, whereas high performers uniformly preferred the paper interfaces that best supported performance.

	Prefer Paper	Prefer Tablet	% Correct Paper	% Correct Tablet
Low Students	37.0%	63.0%	57.5%	50.0%
High Students	100.0%	0.0%	82.5%	80.0%

Table 2: Preference for the tablet (PT, GT) versus paper (PP, DP) interfaces (left), and math performance levels in these interfaces (right) for low- versus high-performing students

When using the graphical tablet interface in which students had free choice to type or use pen input, 50% of students mixed text and pen input (e.g., using text for digits and formulas, pen for diagrams and labels). Another 37.5% only used pen input, and 12.5% only used text. Overall, 37% of all input was text-based and 63% was pen input.

When using *pencil and paper*, typical student comments included: "If I had to focus and do my best on an AP math test, I'd use paper and pencil." "I'm most used to it, and it was the easiest, most efficient, and accurate." When using the *digital stylus interface* they said: "It was closer to my normal process, so I was not distracted." "I stayed focused, just like paper and pencil." But also: "I'd like it better if it was a digital pencil I could erase." Reactions to the *pen tablet interface* included: "It was a little distracting because it was so fascinating to play with." "My writing looked funnier." "It was cool. I could draw, erase and change things easily." In contrast, students criticized the *graphical tablet interface*: "Sometimes clicking and choosing the right symbol was confusing and distracting. I was focusing on the computer, not the problem." "I knew I could do it but it was just faster to write." "It was not that helpful at solving problems. Okay for input though." On the importance of pen input and diagramming for math, students said: "I'm a visual learner. I like to draw pictures to help me think clearly." "I need visualizations to figure out the problems."

DISCUSSION

For the same students completing the same math problems, performance was just as fast when using the digital stylus interface as paper and pencil (115.36 and 117.96 seconds), although working on the pen and graphical tablet interfaces averaged a significant 16% slower than the two paper-based interfaces. With respect to solving their geometry

problems correctly, students also performed as well using the digital stylus interface as with paper and pencil, averaging 71.3% and 70.9% correct solutions, respectively. In contrast, their performance dropped to 68.8% and 64.1% on the pen and graphical tablet interfaces. In short, of the three interface options explored, only the digital stylus and paper interface preserved students' ability to complete work as quickly and accurately as existing pencil and paper work practice.

However, the lower performing students' ability to solve math problems correctly and to remember the problem content they had just worked on both were selectively disrupted when using the tablet interfaces, but especially with the graphical tablet interface, as summarized in Figures 4 and 7. Whereas high-performing students actually improved slightly from 77% of problems solved correctly with pencil and paper, to 82% when using the pen-based interfaces and 85% with the graphical tablet interface, in contrast the low-performing students' correct solutions dropped from 64% with pencil and paper, to 55% with the pen-based interfaces and more sharply to 39% with the graphical tablet interface. Parallel trends were revealed for students' recall of problem content. The high-performing students correctly recalled 69.4% and 70.5% of math content they had just worked on after using the paper and tablet interfaces, and likewise low-performing students recalled 69.3% of math content after using the paper-based interfaces. However, after working on the tablet interfaces the low-performing students only were able to recall 61.1% of the same information. In short, the low-performing students incurred a handicap when using the pen and graphical tablet interfaces that high-performing students simply did not experience. With respect to the impact of introducing new technology, this substantial performance divergence that was generated between low- and high-performing student subgroups when using the tablet interfaces constituted a form of *digital divide*.

Student think-aloud data yielded valuable insights into how their focus of attention and ability to work changed as a function of the different interfaces. The frequency of spontaneous comments about the interface, which revealed the extent to which students were distracted from focusing their attention on solving math problems, increased significantly between the paper and tablet interfaces. Compared with existing pencil and paper work practice, interface comments increased a substantial 326% when using the pen tablet and 661% with the graphical tablet interface. However, as illustrated in Figure 6, negative interface comments only were significantly higher in the graphical tablet condition. In fact, compared with pencil and paper, negative comments about the interface increased a substantial 1,076% while trying to solve math problems with the graphical tablet interface.

As students became more distracted with the tablet-based interfaces, especially the graphical tablet interface, their high-level meta-cognitive math comments correspondingly declined, as illustrated in Figure 5. Although students' low-

level procedural math comments were unaffected by the interface, their ability to think at a more abstract, strategic, and self-reflective level about the process of solving math problems declined significantly when using the graphical tablet interface. Compared with paper and pencil, students' high-level math comments decreased 50.3% when using the graphical tablet interface. These results are consistent with previous research indicating that writers' high-level planning was reduced during computer-supported word processing compared with hardcopy composition, whereas their lower-level planning was not [8].

The convergent pattern of results in this study indicates that Cognitive Load Theory provides a coherent and powerful basis for making quantitative rank-order predictions about user performance with different interfaces. As predicted by CLT, interfaces that depart more from familiar work practice generated greater extraneous cognitive load, such that performance became slower and less accurate. In addition, the lower-performing students with less well consolidated geometry expertise experienced more load than high-performing students, which interacted with extraneous load generated by the interfaces to disrupt their problem solving and memory performance specifically while using the tablet interfaces. As predicted by Cognitive Load Theory, student performance also declined as the intrinsic difficulty level of problems increased. In the future, it will be important that new interfaces for education, mobile computing, and other areas be designed to minimize cognitive load so users can focus on the intrinsic difficulty of real-world tasks.

Although the graphical tablet interface represented the state-of-the-art with respect to students' current computer support and all were experienced GUI interface users, nonetheless this interface was consistently ranked the least preferred for supporting mathematics. In contrast, students rated the pen tablet interface as highly as pencil and paper, and the digital stylus interface was intermediate. Judging from students' self-reports and their high frequency of negative interface comments, the graphical tablet interface was consistently disliked. When asked which interface they would use if they had to perform their best on an AP exam, 100% of high-performing students said they would prefer the paper-based interfaces. However, a reversal occurred for low-performing students, 63% of whom said they would prefer using the tablet interfaces even though their performance actually was more poorly supported by them. This *performance-preference paradox* reflects weaker meta-cognitive skills in the lower-performing students, who clearly were less aware than high-performing students of the tools they needed to perform well [20]. The low-performing students also may have been more vulnerable to the illusion that technology would compensate for their performance deficits, or perhaps make their work easier.

Our society seems to have a very strong need to believe that our lives will be uniformly better with the introduction of technology, but that simple assumption clearly is faulty. As technology is increasingly introduced into classrooms,

one important implication of the present results is that lower-performing students may well not benefit equally or have intuitions that are as accurate as higher-performing students about how to make best use of their new computational tools. This underscores that introducing new interfaces into the educational system risks exacerbating pre-existing performance differences between low- and high-performing student groups, rather than closing the gap.

Of the interface alternatives compared in the present work, the digital stylus and paper interface that most closely approximated existing work practice yielded better support for performance than either of the tablet interfaces, and it also contained the least extraneous complexity. Of the tablet interfaces, the pen tablet supported performance better than the graphical tablet interface, because pen input is so familiar to students and also supports the widest range of representational systems required for expressing math (e.g., including diagrammatic and symbolic information). Future work should pursue further prototyping and performance comparisons of pen-based interfaces, including testing with both recognition-based and transmission-based systems. Future research also could benefit from longitudinal follow-ups of the present work, assessment of other skills such as learning and generalization, and investigation of other educational domains such as science. In the future, interfaces will be needed that can provide better support for extended problem solving in domains like math and science education, rather than mechanical tasks like word processing.

ACKNOWLEDGMENTS

Thanks to Rachel Coulston, Marisa Flecha-Garcia and Rebecca Lunsford for recruiting students and pilot testing math problems. This research was supported by DARPA Contract No. NBCHD030010. Any opinions, findings or conclusions are those of the authors, and do not reflect the views of DARPA or the Department of the Interior.

REFERENCES

1. Abowd, G. Classroom 2000: An experiment with the instrumentation of a living educational environment, *IBM Systems Journal*, 1999, vol. 38, no. 4, pp. 508-530.
2. Anderson, R., Hoyer, C., Wolfman, S. and Anderson, R. A study of digital ink in lecture presentation. In *Proceedings of CHI'04 Human Factors in Computing Systems* (April 24-29, Vienna Austria), ACM/SIGCHI, NY, 2004, pp. 567-574.
3. Anoto technology, <http://www.anoto.com/>
4. Anthony, L., Yang, J. and Koedinger, K. Evaluation of multimodal input for entering mathematical equations on the computer, In *Proceedings of CHI'05 Human Factors in Computing Systems* (April 2-7, Portland Oregon), ACM/SIGCHI, NY, 2005, pp. 1184-1187.
5. Bach, D. Math gaps grows for minority students: Difference in WASL scores shows significant jump at seventh grade, *Seattle Post-Intelligencer*, Nov. 14, 2005.
6. Baddeley, A. *Working Memory*, Oxford University Press, NY, 1986.
7. Cohen, P. and McGee, D. Tangible multimodal interfaces for safety-critical applications, *Communications of the ACM*, 2004, vol. 47, no. 1, pp. 41-46.
8. Haas, C. Does the medium make a difference? Two studies of writing with pen and paper and with computers, *Human Computer Interaction*, 1989, vol. 4, pp. 149-169.
9. Landay, J. and Myers, B. Sketching interfaces: Toward more human interface design, *IEEE Computer*, March 2001, vol. 34, no. 3, pp. 56-64.
10. LaViola, J. and Zeleznik, R. MathPad2: A System for the creation and exploration of mathematical sketches, *ACM Transactions on Graphics, Proceedings of SIGGRAPH*, 2004, 23(3):432-440.
11. Leapfrog, (<http://www.leapfrog.com>), 2006.
12. Liao, C., Guimbretiere, F. and Hinckley, K. Papier-Craft: A system for interactive paper, In *Proceedings of UIST'05*, ACM/SIGCHI, NY, 2005, pp. 241-244.
13. Marriott, M. Blacks turn to internet highway, and digital divide starts to close, *New York Times*, March 31, 2006.
14. Mousavi, S., Low, R. and Sweller, J. Reducing cognitive load by mixing auditory and visual presentation modes, *Journal of Educational Psychology*, 1995, vol. 87, no. 2, 319-334.
15. Paas, F., Tuovinen, J., Tabbers, H. and van Gerven, P. Cognitive load measurement as a means to advance cognitive load theory, *Educational Psychologist*, 2003, vol. 38, no. 1, pp. 63-71.
16. Pederson, E., McCall, K., Moran, T. and Halasz, F. Tivoli: An electronic whiteboard for informal group meetings, In *Proceedings of INTERCHI'93 Human Factors in Computing Systems* (April 24-29, Amsterdam), ACM/SIGCHI, NY, 1993, pp. 391-398.
17. Plamondon, R. and Srihari, S. On-line and off-line handwriting recognition: A comprehensive survey, *IEEE PAMI*, Jan. 2000, vol. 22, no. 1, pp. 63-84.
18. Smithies, S., Novins, K. and Arvo, J. Equation entry and editing via handwriting and gesture recognition, *Behavior and Information Technology*, 2001, vol. 20, pp. 53-67.
19. van Merriënboer, J. and Sweller, J. Cognitive load theory and complex learning: Recent developments and future directions, *Educational Psychology Review*, 2005, vol. 17, no. 2, pp. 147-177.
20. Winne, P. H. and Perry, N. E. Measuring self-regulated learning. In *Handbook of Self-Regulation* (ed. by M. Boekaerts, P. Pintrich & M. Zeidner), Academic Press, Orlando FL., 2000, pp. 531-566.