Numeracy
Advancing Education in Quantitative Literacy

Volume 8 | Issue 1 Article 4

2015

Effects of Reducing the Cognitive Load of Mathematics Test Items on Student Performance

Susan C. Gillmor
University of Kansas, scgillmor@ku.edu
John Poggio
University of Kansas, jpoggio@ku.edu
Susan Embretson
Georgia Institute of Technology, susan.embretson@psych.gatech.edu

Effects of Reducing the Cognitive Load of Mathematics Test Items on Student Performance

Abstract
This study explores a new item-writing framework for improving the validity of math assessment items. The authors transfer insights from Cognitive Load Theory (CLT), traditionally used in instructional design, to educational measurement. Fifteen, multiple-choice math assessment items were modified using research-based strategies for reducing extraneous cognitive load. An experimental design with 222 middle-school students tested the effects of the reduced cognitive load items on student performance and anxiety. Significant findings confirm the main research hypothesis that reducing the cognitive load of math assessment items improves student performance. Three load-reducing item modifications are identified as particularly effective for reducing item difficulty: signalling important information, aesthetic item organization, and removing extraneous content. Load reduction was not shown to impact student anxiety. Implications for classroom assessment and future research are discussed.

Keywords
educational measurement, cognitive load theory, item writing, classroom assessment

Cover Page Footnote
Susan Gillmor is a Ph.D. candidate at the University of Kansas studying educational research, evaluation, measurement, and statistics. While she completes her dissertation research on growth modeling for teacher evaluation, Susan is a Graduate Teaching Assistant and works as a Research Associate for the Center for Assessment in Dover, NH.

John Poggio is a Professor in the Department of Educational Psychology and Research at the University of Kansas. He is the former director of the Center for Educational Testing and Evaluation where he ran the Kansas Assessment Program for 30 years. John is the co-Principal Investigator for the IES project "An Adaptive Testing System for Diagnosing Sources of Mathematics Difficulties."

Susan E. Embretson is Professor of Psychology at the Georgia Institute of Technology where she serves as Director for the Quantitative Psychology Program. Susan was the 2013 recipient of the National Council on Measurement in Education Career Contributions Award. She is the Principal Investigator for the IES project "An Adaptive Testing System for Diagnosing Sources of Mathematics Difficulties."

Introduction

With the intent of more validly assessing student understanding, educators frequently write their own assessments or select test items provided by commercially produced curricula. However, these test items can be often fraught with problems that can be distracting or confusing to students (Zorin et al., 2013). Research has shown that construct-irrelevant factors such as language complexity and item format can interfere with student performance on assessments (Haladyna et al. 2002; Shaftel et al. 2006; Martiniello 2008; Cawthon et al. 2012;). These complications can restrict the appropriateness of educational measurement, which can result in inaccurate judgments about student understanding. When a test inadvertently assesses factors that it is not intended or designed to measure, the resulting construct-irrelevant variance causes a threat to the test's validity. Test validity is an ongoing process of judging the degree to which inferences about test scores are appropriate for their proposed uses. Validity is the most central concern for test development and evaluation (AERA, APA and NCME 2014). On top of validity, the new 2014 Standards for Educational and Psychological Testing emphasize fairness in access to the construct(s) measured. According to AERA, APA and NCME. (2014: 57):

> "Standardized tests should be designed to facilitate accessibility and minimize construct-irrelevant barriers for all test takers in the target population"

The purpose of this study is to investigate the applicability of Cognitive Load Theory (CLT) to educational measurement for improving test score validity and fairness. This study evaluates a series of systematic item modifications rooted in CLT that teachers and test developers can use to reduce construct-irrelevant variance when writing mathematics assessment items. Cognitive Load Theory, defined by Sweller (1988, 1989) for instructional design purposes, originates from the cognitive sciences and rests on the assumption that the human working memory has limited capacity (Miller 1956). Research has shown that reducing the cognitive load of instructional materials facilitates learning efficiency (Clark et al. 2011). This study transfers the insights of CLT to educational measurement by curtailing extraneous cognitive load that may contribute to construct-irrelevant variance in order to more accurately measure the intended construct.

An illustrative example of a traditional item and a reduced cognitive load assessment item is shown in Table 1. The first item (A) highlights a commercially available mathematics assessment question that exhibits unnecessarily high cognitive load demands on the examinee. The item is designed to assess student understanding about a basic geometry concept: the sum of interior angles of a quadrilateral. However, due to the complexity of the original wording of the item and the inclusion of irrelevant details, it also likely measures reading comprehension among a number of other abilities inadvertently. The second item

(B) measures the same underlying construct while stripping the irrelevant variance caused by extraneous load, and thus producing a more instructionally valid measurement of student knowledge. Without being "loaded" with additional reading and complexity, student performance will likely be higher on this item than the original, which can be interpreted as a reflection of increased measurement validity.

Table 1.
Example Item in Traditional and Reduced Cognitive Load Form.

A. Traditional Item

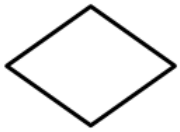
The picture below shows an old-fashioned kite. The dotted lines represent the wooden braces behind the kite's fabric.

What is the sum of the interior angles of the fabric area of the kite?

A. 45°
B. 90°
C. 180°
D. 360°

B. Reduced Load Item

What is the sum of the interior angles in the shape below?



A. 45°
B. 90°
C. 180°
D. 360°

Two guiding hypotheses structure this study:

1. Reducing the extraneous cognitive load of mathematics assessment items will improve their validity for assessing student knowledge and in turn, will improve student performance, and

2. Reducing the extraneous cognitive load of mathematics assessment items will reduce student anxiety while taking the examination.

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

Review of Literature

Some of the most important early literature about working memory is attributed to Miller (1956). Miller postulates that the human mind cannot hold more than seven (plus/minus two) "chunks" of information at any one time (Miller 1956). Cognitive Load Theory defines into three types of load that put demand on this limited memory capacity during learning or cognitive processing: intrinsic, extraneous, and germane (Paas et al. 2003).

Intrinsic cognitive load is associated with the inherent challenge or level of difficulty of the material being processed. The level of intrinsic cognitive load of a given content will vary across students; it cannot be manipulated through instructional design. The idea of intrinsic load stems from de Groot (1966) who studied novice and master chess players and found that a distinguishing trait between the two groups was the masters' ability to accurately replicate a chess board having only seen it for a short time. Experts have superior recall of presented material because the domain holds less intrinsic cognitive load.

Extraneous load, the focus of the current study, is defined as anything in the instructional materials that occupies working memory capacity but is irrelevant to the intended material. Instructional designers have dedicated much research to methods of reducing extraneous cognitive load in order to free up space in the working memory for learning and problem solving. Germane cognitive load refers to the cognitive resources dedicated to constructing new schema in long-term memory.

Germane cognitive load increases with student motivation to participate in the learning process; it is the mental effort that students dedicate to learning or solving a problem. Increased germane load contributes to new stored knowledge for the student. These three types of cognitive load—intrinsic, extraneous, and germane—are used additively with the remaining free space to comprise total working memory capacity (Mayer and Moreno 2010).

Cognitive psychologists have built a body of literature supporting the use of Cognitive Load Theory to facilitate learning and problem solving. The earliest work, in the late 1980s, introduced educational psychologists to a set of guidelines to manage the cognitive load of instructional materials. Sweller (1989) found that some instructional presentations increase cognitive demands such as examples that require students to split their attention between two sources of information, while other presentations reduce cognitive load such as providing worked examples. Recently, Clark et al. (2011) published a synthesis of the current research on cognitive load management techniques. The findings of this work are central to specifying the cognitive load item modifications used in this study.

While CLT is broadly used and accepted in the field of instructional design,

only a few previous studies have applied its principles to educational measurement. A recent series of studies tested the effects of modified items on the performance of students with severe disabilities (Kettler et al. 2009, 2011). These researchers modified items using CLT, Universal Design guidelines (Rose and Meyer 2000) and research on item development. Generally, those researchers found that reduced load items did improve student performance. Reducing the length of the item stimulus seemed to be particularly effective (Kettler et. al. 2011). They also found that reduced load items decreased item difficulty more dramatically for the students with disabilities as compared to the students they tested without disabilities.

Miller (2011) similarly confirms the applicability of CLT principles to assessment by finding that aesthetically improving computer-based test items decreases participant cognitive load while increasing participant satisfaction and performance on an e-assessment. Aesthetic factors contributing to this finding included increased contrast, organization, and flow of assessment content.

A parallel field of research in working memory and problem solving has developed simultaneously. Hitch (1978) found that the limited working memory storage contributes to calculation error during mental arithmetic. More generally, the more one needs to store in the working memory, the more likely one is to forget bits of information and make an error. Not only do these findings support the primary research hypothesis of the current study, but they suggest that test items could benefit from a reduction in numerical and arithmetic complexity when the numerical values are not necessary for the measured construct.

Other related work comes from Embretson and Wetzel (1987) and Gorin and Embretson (2006). These researchers systematically mapped the cognitive complexity of reading passage items and developed a cognitive model that identifies construct-relevant item features that contribute to item difficulty, or the item's intrinsic cognitive load. For reading passage items, as the amount of text increases so does the item difficulty due to the increased demands on the working memory. Therefore, when text is construct-irrelevant, removing the added demand on the working memory will likely improve the validity of the measurement and increase student performance.

In addition to improving performance, there is some evidence to suggest that reducing extraneous cognitive load may also alleviate student stress or anxiety (Miller 2011). Both anxiety and cognitive load are inversely correlated with performance because both factors consume the working memory's processing resources (Chen and Chang 2009). Ashcraft and Kirk (2001) investigated this concept and found that the aspects of mathematics performance that rely heavily on working memory are the same aspects that are most affected by mathematics anxiety. Therefore, although there is a paucity of literature examining the relationship between cognitive load and anxiety, the secondary hypothesis of the

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

current study is that reducing the cognitive load of test items will lead to reduced student anxiety during test-taking. This study builds on the previous research by systematically and purposively testing the direct effects of reducing cognitive load of assessment items on student performance and anxiety.

Methods

The study participants are 222 eighth-grade students from three regionally diverse schools in a geographically large, Midwestern state. The study was conducted in early fall and students within each participating school were randomly assigned to the experimental group or comparison condition. Table 2 shows the number of participants from each school in the comparison and experimental groups.

Table 2.
Number of Participants by School.

School	Control	Experimental	Total
1	32	30	62
2	59	69	128
3	15	17	32
Total	106	116	222

The comparison group was given a traditional test with commercially available items that were chosen to represent typical levels of cognitive load. The experimental group received a modified set of the same items which had characteristics leading to extraneous cognitive load removed. These extraneous load-reducing modifications were all adapted for educational measurement from the cognitive load studies discussed in the literature review. A complete list of the seven strategies employed and the reference from which this load-reducing strategy was taken can be found in Table 3.

The nature and number of modifications used to modify each item varied with the content and structure of each item. Due to the diversity of the items, researchers used judgment as to which strategies were necessary to remove extraneous cognitive load. When modifying items, extreme care was taken to not alter the underlying content objective or construct being evaluated; changes were made only to make the item more accessible. The intent is to reduce cognitive load of test items so students can more efficiently use their available cognitive resources for problem solving. All items were catalogued and qualitatively coded with the types of modifications employed for load reduction.

Table 3.
Strategies for Reducing Cognitive Load in Assessment Items.

Method	Description	Citation
Translation	Reduce word count and simplify language.	Kettler et. al. 2011
Visual Aid	Use diagrams to represent spatial information.	Clark et al. 2011
Signaling	Focus attention with signals and cues.	Clark et al. 2011
Weeding	Pare content down to essentials. Eliminate extraneous visuals and text.	Clark et al. 2011
Sequencing	Ask question first to give a direction to the item, and then include supporting information. This also includes ordering the answer options logically.	Clark et al. 2011
Aesthetics	Format item logically, and aesthetically. Place text near corresponding features on figures.	Miller 2011
Numerical simplicity	Use smaller, rounded, and familiar numbers when values are construct-irrelevant.	Hitch 1978

The items on both of the test forms come from an item bank that is developed, validated, and distributed by the eLearning Design Lab at the University of Kansas. The items had all previously been used as part of the Kansas accountability testing program and retired due to over-use. Permission from the test publisher has been given to release the items in this paper. The test forms were compiled for the specific purposes of this study and represent five, seventh-grade mathematics content standards; each measured by three items. Items from the bank were chosen based on the researchers' judgment as likely benefiting particularly well from editing with cognitive load-reducing strategies. Due to the relatively short length of the exams (15 test questions), and the likely multidimensional nature of the test forms, the comparison and experimental forms exhibited relatively low Cronbach's alpha reliability estimates of .754 and .656 respectively. The form with traditional items has better reliability; implications of this finding are discussed in more detail in the discussion section.

In addition to taking the cognitive assessment, all student examinees reported how they felt while taking the test by completing the state anxiety subtest of the Spielberger State-Trait Anxiety Inventory for Children (STAIC) (Spielberger and Edwards 1973). This inventory was developed for purposes of research on anxiety in children and was used with permission from the developer. The scale measures state anxiety with twenty questions that ask respondents to report how they felt at a particular moment in time (e.g. calm) on a three-point rating scale (e.g., from "very calm" to "not calm"). The STAIC test manual reports Cronbach's alpha reliability ranging from .82 for males to .87 for females, which

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

is above the accepted standard for basic research in the social sciences (Nunnally 1978). The observed Cronbach's alpha reliabilities for the STAIC in the current study were .93 for the students in the control group, and .87 for the students who received the reduced load mathematics assessment. Validity evidence presented in the test manual supports the use of the inventory for research purposes with children. The STAIC has been widely accepted and is currently associated with over 200 references in the literature (Spielberger and Edwards 1973).

The test forms were administered in the classroom during the regularly meeting mathematics class periods of the students. As approved by the Institutional Review Board at the first author's university, parents of the students were informed of this research activity and gave signed consent. The paper and pencil test forms were distributed randomly to students at their seats. The tests took approximately 30 minutes in all (the mathematics achievement test followed by the STAIC) to complete, after which teachers allowed students to work independently on school work or reading until all students had finished. At one school, two students did not complete the test form within the 45-minute class period and were allowed to stay after class until they were finished.

Analysis

The first hypothesis of this study is that the items with reduced cognitive load will have lower item difficulties than the traditional test items. Item difficulty is defined as the proportion of students who answered the item correctly, with zero being the most difficult and one being the easiest. For the present analyses, the item difficulties serve as the dependent variables of interest, while the independent variable is the two-level group identifier (comparison, experimental). A two-group Hotelling's T^2 multivariate test statistic is used to test the student performance differences between the two, fifteen-item forms.

Additionally, as a follow-up procedure, each item is analyzed separately. For both the Hotelling's T^2 and the follow-up *t* tests an alpha of .05 was used. The researchers made an a priori decision to not make any type-1 error adjustment because there is a unique hypothesis associated with each item. Because each item had a different number and nature of modifications, it was important to test each individually with its own hypothesis, and, therefore, an error correction was deemed unnecessary.

The second hypothesis for this study is that reducing the cognitive load of test items will result in a decrease in student state anxiety while taking the test. To test differences in anxiety, an independent samples *t* test is performed with the student-level average of the anxiety measure as the dependent variable, and the treatment condition (comparison group, experimental group) as the independent variable. An a priori alpha of .05 was used for this test.

Results

The first research hypothesis predicted that the reduced-cognitive load items would result in higher student performance than the traditional test items. Hotelling's T^2 test statistic reveals large, true differences between the two forms ($F_{(15,206)} = 4.562$, $p < .001$). The omnibus Cohen's d effect size is .37, which means that the average reduction in difficulty of the reduced load items was 37% of a standard deviation. A two-way analysis of variance shows that while there are true differences in average performance across schools (an expected finding), there is no significant interaction between condition and school ($F_{(2,216)} = .990$, $p = .373$). This means that on average, student performance was higher on the treatment form by about the same amount across all three settings. Post-hoc analyses using multivariate Hotelling's T^2 show that seven of the fifteen items on the experimental form have significant differences in student performance. The results for these analyses are shown in Table 4. Six of the items on the reduced-load form resulted in significantly higher student performance, while one of the fifteen reduced-load items showed a significant reduction in student performance as compared to the comparison group. Of the six statistically significant items where reduced cognitive load resulted in improved performance, the Cohen's d effect size ranged from .31 to .71 with an average of .4.

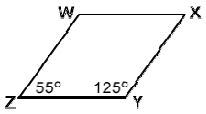
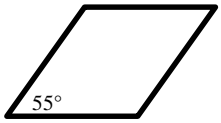
Table 4.
Results from 16 independent samples *t* tests.

	Control Form		Experimental Form				
Item	*p*	SD	*p*	SD	*t*	Sig. (2-tailed)	Cohen's *d*
All**	.56	.22	.63	.19	-2.76	.01	-.37
1	.44	.50	.47	.50	-0.33	.74	-.04
2**	.32	.47	.66	.48	-5.26	<.01	-.71
3*	.87	.34	.74	.44	2.41	.02	.32
4*	.54	.50	.67	.47	-2.60	.04	-.35
5	.59	.49	.56	.50	0.51	.61	.07
6	.53	.50	.65	.48	-1.79	.08	-.24
7*	.42	.50	.55	.50	-2.04	.04	-.27
8**	.77	.42	.91	.28	-.2.89	<.01	-.39
9	.39	.49	.51	.50	-1.83	.07	-.25
10	.62	.49	.53	.50	1.46	.15	.20
11	.50	.50	.59	.50	-1.29	.20	-.17
12	.45	.50	.41	.49	0.71	.48	.02
13*	.76	.43	.88	.33	-2.27	.03	-.31
14	.88	.33	.89	.31	-0.24	.81	-.03
15**	.29	.46	.52	.50	-3.49	<.01	-.47

Note. p is the proportion of students answering the item correctly.
*indicates statistical significance at the $\alpha = .05$ level. **indicated statistical significance at the $\alpha = .01$ level.

To help illustrate these findings, examples of a significant and non-significant item are included. Table 5 below shows the original and reduced versions of Item 13. This item shows significant differences in difficulty from 78% of students answering the original item correctly to 88% of the students answering the reduced load version correctly (t = -2.27, p = .03). No new information was added to the item to help students solve it; instead the item was made easier by translating the item language and formatting it in a way that is more accessible for all students. The state standard that is tested by this item is "Students identify angle and side properties of triangles and quadrilaterals: parallelograms have opposite sides that are parallel and congruent." To reduce the cognitive load of this item the authors used the strategies of translation, sequencing, and spatial contiguity. The unnecessary information was removed, reducing the reading load and thus the word count, the question the student was expected to answer was moved to the beginning of the item, and lastly the graphic was simplified and centered in the item.

Table 5.
Example of Significant Item 13.

Traditional Item (Control Form)	Reduced Load Item (Experimental Form)
Properties of quadrilateral WXYZ are described below.  side XY is parallel and congruent to side WZ side WX is parallel and congruent to side ZY The measure of ∠WZY is 55° The measure of ∠XYZ is 125° Which term describes quadrilateral WXYZ? A. Parallelogram B. Rectangle C. Square D. Trapezoid	What do we call this shape? The opposite sides are parallel and the same length.  A. Parallelogram B. Rectangle C. Square D. Trapezoid

Item 9, shown in Table 6, is an example of an item where the load-reducing strategies did not result in improved student performance. Although the difficulty is less for the reduced load version than the original item, with the percentages of students answering the item correctly being 51% and 39% respectively, this is not a statistically significant difference (t = -1.83, p = .07). The Cohen's d effect size for this item is .25, which is small, meaning that either the item modifications did

not go far enough to reduce the cognitive load, or that the original item started with relatively low load. This item tests the state standard that reads: "Students identify, state, and continue a pattern presented in various formats including numeric (list or table), algebraic (symbolic notation), visual (picture, table, or graph), verbal (oral description), kinesthetic (action), and written using positive rational numbers including arithmetic and geometric sequences (arithmetic: sequence of numbers in which the difference of two consecutive numbers is the same, geometric: a sequence of numbers in which each succeeding term is obtained by multiplying the preceding term by the same number)." The load-reducing strategies used were translation, sequencing, spatial contiguity, and signaling. The first three strategies were used to reduce the word count, restructure to begin the stem with the question, and improve the visual design (e.g. centering the number sequence, reformatting the fractions). The signaling strategy is used by adding and bolding the "?" in number sequence in order to focus the students' attention to what it is the item is asking them to do.

Table 6.

Example of Non-significant Item 9.

Traditional Item (Control Form)	Reduced Load Item (Experimental Form)
A number pattern is shown below. 1/3, $^1/_2$, 2/3, 5/6 The pattern continues. Which rule could be used to find the next number in the pattern? A. Add 1/3 to the previous term. B. Add 1/6 to the previous term. C. Multiply the previous term by 1/3. D. Multiply the previous term by 1/6.	Which rule finds the next number in the sequence below? $\frac{1}{3}$ $\frac{1}{2}$ $\frac{2}{3}$ $\frac{5}{6}$? A. add $\frac{1}{3}$ B. add $\frac{1}{6}$ C. multiply by $\frac{1}{3}$ D. multiply by $\frac{1}{6}$

The second hypothesis of this study predicts that students given the reduced load form will experience less state anxiety while taking the assessment than students assigned to the traditional (control) form. Only 204 participants are used in this study because 18 students did not choose to fill out the STAIC at all. Of the 204 students that did respond to at least one item on the anxiety inventory, the completion rate was high with an average of 19.82 items completed out of the 20. A negative correlation between anxiety and performance (r = -.350, p < .001) confirms prior literature on the relationship and serves as validity evidence for the STAIC measure in the current study (Seipp 1991). However, the independent

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

samples *t* test does not result in statistically significant differences in the average reported state anxiety for those who received the reduced load version as compared with the traditional form. The Cohen's *d* effect size for this mean difference is .2, which means that the average anxiety level of the experimental group is at the 58th percentile of the comparison group. This difference is small and not substantively meaningful.

Table 7.
Results from STAIC (α = .904)

Control Form			Experimental Form				
n	Mean	SD	n	Mean	SD	Sig.	Cohen's *d*
94	1.68	0.4	110	1.61	0.29	.197 †	0.207

†Equal variances not assumed

Discussion

As hypothesized, students who were randomly assigned the reduced cognitive load form performed better than the students assigned the form with typical assessment items. Removing the extraneous cognitive load of assessment items proved effective in increasing the accessibility of the measured construct, resulting in enhanced performance and potentially test fairness. Stripping items of extraneous cognitive load was successful at increasing the accessibility of the tested content. A higher proportion of students answering the items correctly resulted from the students in the treatment group having a greater opportunity to demonstrate their knowledge. Unfortunately, evidence of improved test score validity is not supported by the reliability analysis. The traditional form exhibited a higher Cronbach's alpha reliability estimate than the reduced-load form. This means that removing extraneous cognitive load was not effective at increasing measurement precision. This reduced reliability could be an artifact of the reduced variability given the improvement in performance. Likely due to the short nature of the test forms, neither form achieved acceptable reliability, and replication of the study with longer forms will be necessary to further understand the nature of the impact of reducing the cognitive load on the item covariances.

Although, overall, the effect of reducing the cognitive load of test items had a positive impact on student performance, differences between the control and experimental forms were not found on all items. Items were analyzed individually in order to better understand the unique contributions of the different load-reducing techniques. An explanation for the variability of results across the items comes from a deeper look at the types of the modifications used on each of the items. Informative patterns emerge when we use this information in conjunction with our results from the fifteen *t* tests. Testing each item separately gives insight

into which strategies are most useful for improving student performance. The most effective strategies were deemed those producing significantly improved items. Signaling, aesthetics, and weeding are the three extraneous cognitive load-reducing strategies that produced significantly improved student test performance in mathematics. The signaling method refers to directing the students' attention to key words in the test item. For example, on Item 2, shown in Table 8, students likely performed statistically better than their peers in the control group because the test item signaled that "least" is an important word for their attention.

Table 8.
Example of Item 2 using Signaling Strategy.

Traditional Item (Control Form)	Reduced Load Item (Experimental Form)
Alice, Brad, Cory, and Derek each had a pizza for lunch. Alice ate 3/10 of her pizza; Brad ate 42%, Cory ate 2/5 of his pizza, and Derek ate 45%. Who had the most pizza left after lunch? A. Alice B. Brad C. Cory D. Derek	Which value below is the **least** amount of pizza? A. $\frac{3}{10}$ of a pizza B. 42% of a pizza C. $\frac{2}{5}$ of a pizza D. 45% of a pizza

Note. All load reductions for this item are signaling, translation, aesthetics, and weeding.

The aesthetics modification involves managing the item content and white space in an organized manner. The layout of the item should not be confusing or distracting. Such unneeded distractions can lead to more difficult items arbitrarily. The item shown in Table 1 is an example where aesthetic modifications likely contribute to a significant reduction in item difficulty. The confusing and distracting details in the image were removed and it was enlarged and centered in the item. Due to these aesthetic changes, the test item task became notably more accessible for students and, thus, arguably fairer.

The weeding modification results in the most drastic changes to the language and content of the item. This technique refers to stripping the item of any construct-irrelevant or unnecessary information. It is not surprising that this modification leads to significant improvement in performance. Mayer and Moreno (2010) found that the median effect size of this modification to instructional materials is .7. Special precaution must be taken when using this item-modification technique not to remove all context from an item that tests a content standard requiring an application of skills to a "real-life" scenario. One can still remove extraneous information from an application item, but care must be taken as not to change the tested construct. An example of one of the items that significantly improved performance with the help of the weeding modification is Item 8, shown in Table 9. Item 8 assesses the students' ability to identify and

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

continue a pattern or sequence of numbers. Due to the difference in difficulties between the versions of the items, we can infer the additional information presented in the original item hindered student performance in a construct-irrelevant way.

Table 9.
Example of Item 8 using Weeding Strategy.

Traditional Item (Control Form)	Reduced Load Item (Experimental Form)
The amount of overtime dollars earned by Sam at his job for 4 of his 5 workdays last week is shown in the chart below. Sam's Overtime Earnings by Day Day: Mon | Tues | Wed | Thu | Fr Overtime Earnings: $7.80 | $11.60 | $15.40 | ? | $23.00 If the pattern in overtime earnings for all 5 days creates an arithmetic sequence, what was the amount of overtime earnings on Thursday? A. $18.00 B. $18.40 C. $19.20 D. $20.00	What number takes the place of the question mark in the pattern below? 7.8 11.6 15.4 **?** 23.0 A. 18.0 B. 18.4 C. 19.2 D. 20.0

Note. All load reductions for this item are weeding, aesthetics, translation, and signaling.

One of the fifteen modified items resulted in significantly lower student performance. Students who took the traditional version of Item 3 performed *better* than students on who took the reduced load version, as displayed in Table 10. This difference in the unanticipated direction may be explained by the contextual clues provided by the *names of the cities* in the traditional item. Students who do not know scientific notation may still be able to answer the traditional item correctly because the participants' Midwestern geographical location would likely make them familiar with the relative distances between the given cities. In the case of this item, the lower average student performance on the experimental form may actually serve as evidence of improved item validity.

An additional possible factor that could have contributed to the observed differences among the items is that some mathematics content standards may be more suited for cognitive load modifications than others. Looking to the results, all but one tested standard had at least one item that was significantly improved. The standard that did not have any of its associated items exhibit significantly different performance between forms assesses student ability to write linear expressions that represent real-world problems using variables and symbols. Item 10, shown in Table 11, is an example of one of the items measuring this standard.

Table 10.
Significant Item in Unexpected Direction.

Traditional Item (Control Form)	Reduced Load Item (Experimental Form)
Students calculated the distances from Topeka, KS to various cities/ Which of the cities is the furthest from Topeka? Chicago, IL 589 miles Denver, CO 5.4×10^2 miles Nashville, TN 6.19×10^2 miles Pierre, SD 5.7×10^2 miles A. Chicago, IL B. Denver, CO C. Nashville, TN D. Pierre, SD	Which distance is the furthest? A. 590 miles B. 5.4×10^2 miles C. 6.20×10^2 miles D. 5.7×10^2 miles

Table 11.
Example Item 10 from Unaffected Content Area.

Traditional Item (Control Form)	Reduced Load Item (Experimental Form)
Dinah has some packs of gum. Gary has one less than twice as many packs of gum as Dinah. Which equation represents the relationship between the number of packs of gum Dinah has (d) and the number of packs of gum Gary has (g)? A. $g = 2d - 1$ B. $g = 1 - 2d$ C. $g = 2d + 1$ D. $g = 2 + 2d$	Which equation shows that the number of apples Gary has (g) is 1 less than 2 times the number of apples Dinah has (d)? A. $g = 2d - 1$ B. $g = 1 - 2d$ C. $g = 2d + 1$ D. $g = 2 + 2d$

There are at least three possibilities for why the items measuring this content standard show similar student performance levels on both forms: 1) the items already have low cognitive load, 2) the experimental form version of the items did not reduce the load enough, or 3) the students have not been exposed to the content, in which case, reducing the cognitive load of the item will not affect the measurement. Since both groups scored correctly at levels significantly higher than chance, the third possibility can be safely ruled out. In order to disentangle the first two possible causes, more research is needed to determine the actual levels of cognitive load that are associated with both versions of these items.

There were no significant differences between the two forms in student anxiety as measured by the Spielberger State Anxiety Inventory for Children. Although there is good theoretical evidence to support this hypothesis, the data did not confirm what was expected. A study by Vytal et al., (2012) may help to provide insight into why there were no significant results. Those authors found that when participants were engaged in a high-cognitive load activity they

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

experienced reduced anxiety. In other words, when the executive function of participants was completely occupied by the task at hand, the anxiety of the participants was reduced; there was no available cognitive space for anxiety to occupy. However, under a low-cognitive load condition, participants were more susceptible to anxiety. This theory is not supported by the current study as there is no detectible relationship between anxiety and cognitive load in the data. However, the study by Vytal et al. (2012) does help shed light on a possible reason why the data do not support our initial hypothesis. Additionally, the test forms were administered in a low-stakes context for the examinees for the purposes of the research. The average anxiety level across conditions is only slightly higher than national norms for middle school students in a relaxed state (Speilberger and Edwards, 1973). Therefore, it is likely that there were mostly low levels of anxiety associated with this test in general.

In conclusion, this study supports the use of cognitive load-reducing strategies for increasing test accessibility and validity. This preliminary research provides a solid foundation on which further study must continue to build the validity argument associated with reduced load test items.

Concluding Remarks

This study provides teachers and test developers with research-based strategies and evidence for improving assessment of student knowledge. Cognitive load theory is appropriate for providing guidance in the item writing process. Items that signal the test taker of important information, are aesthetically well-organized, and are stripped of extraneous information can improve student performance. When these cognitive load-reducing techniques are employed thoughtfully, teachers can have greater confidence that student responses are a reflection of student understanding rather than factors unrelated to the measured construct. Not only are these findings relevant for the classroom, they have particular importance for item writing at the published standardized achievement test level. Cognitive load theory is one way to begin thinking about minimizing construct-irrelevant variance and thus increasing test validity.

Integrating knowledge from the cognitive sciences can improve our ability to create accessible and valid tests for all students. However, this transfer is still in its infancy and more research must be completed. First, research on the relationship between cognitive load and anxiety deserves more attention from cognitive scientists and educational researchers. This aspect of the findings from this study is inconclusive and the other empirical literature in the area is mixed. Cognitive load theory provides many other avenues of investigation and attention for educational measurement. Evidence from this investigation was certain: extraneous features can lead to and result in increased and, thus, unnecessary

cognitive load. These features are irrelevant to, and therefore interfere with the measurement of the intended cognitive target(s) of the items. Future research should continue to investigate the effects of reduced cognitive load items on performance of special populations such as English Language Learners and the cognitively impaired. These groups of students particularly may benefit from reduced cognitive load items.

Ensuring that items are accessible to all students not only increases test validity, but contributes to overall test fairness. Ensuring accessibility, and in turn fairness, through the use of universal design principles is an added emphasis in the new Standards for Educational and Psychological testing. This study suggests that the principles dictated by Cognitive Load Theory may be a new avenue by which increased test fairness can be achieved.

Acknowledgments

We would like to thank the Center for Research on Learning for providing the technical support and guidance for making this project possible, specifically Dr. Edward Meyen, Dr. Diana Greer, Daniel Spurgin, and Thomas Shorock. Thank you to the Institute for Educational Statistics for funding this project as part of a larger grant award.

References

AERA, APA and NCME. See American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (US). 2014. *Standards for educational and psychological testing*. American Educational Research Association.

Ashcraft, Mark H., and Elizabeth P. Kirk. 2001. The relationships among working memory, math anxiety, and performance. *Journal of experimental psychology: General* 130(2): 224–237. http://dx.doi.org/10.1037/0096-3445.130.2.224

Bannert, Maria. 2002. Managing cognitive load—recent trends in cognitive load theory. *Learning and instruction* 12(1): 139–146. http://dx.doi.org/10.1016/S0959-4752(01)00021-4

Cawthon, Stephanie W., Alyssa D. Kaye, L. Leland Lockhart, and S. Natasha Beretvas. 2012. Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. *Journal of school psychology* 50(3): 293–316. http://dx.doi.org/10.1016/j.jsp.2012.01.002

Chen, I-Jung, and Chi-Cheng Chang. 2009. Cognitive Load Theory: An empirical study of anxiety and task performance in language learning. *Electronic Journal of Research in Educational Psychology* 7(2): 729–746.

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4

Clark, Ruth C., Frank Nguyen, and John Sweller. 2011. *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. John Wiley & Sons, 2011.

de Groot, Adriaan D. 1966. Perception and memory versus thought: Some old ideas and recent findings. In *Problem solving*, ed. B. Kleinmuntz, 19–50. New York: Wiley.

Embretson, Susan E., and C. Douglas Wetzel. 1987. Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement* 11(2): 175–193. http://dx.doi.org/10.1177/014662168701100207

Frey, Bruce B., Stephanie Petersen, Lisa M. Edwards, Jennifer Teramoto Pedrotti, and Vicki Peyton. 2005. Item-writing rules: Collective wisdom. *Teaching and Teacher Education* 21(4): 357–364. http://dx.doi.org/10.1016/j.tate.2005.01.008

Gorin, Joanna S., and Susan E. Embretson. 2006. Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement* 30(5): 394–411. http://dx.doi.org/10.1177/0146621606288554

Haladyna, Thomas M., Steven M. Downing, and Michael C. Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education* 15(3): 309–333. http://dx.doi.org/10.1207/S15324818AME1503_5

Heo, Misook, and Anthony Chow. 2005. The impact of computer augmented online learning and assessment tool. *Journal of Educational Technology & Society* 8(1).

Hitch, Graham J. 1978. The role of short-term working memory in mental arithmetic. *Cognitive Psychology* 10(3): 302–323. http://dx.doi.org/10.1016/0010-0285(78)90002-6

Kettler, Ryan J., Stephen N. Elliott, and Peter A. Beddow. 2009. Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education* 84(4): 529–551. http://dx.doi.org/10.1080/01619560903240996

Kettler, Ryan J., Michael C. Rodriguez, Daniel M. Bolt, Stephen N. Elliott, Peter A. Beddow, and Alexander Kurz. 2011. Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education* 24(3): 210–234. http://dx.doi.org/10.1080/08957347.2011.580620

Martiniello, Maria. 2008. Language and the performance of English-language learners in math word problems. *Harvard Educational Review* 78(2): 333–368.

Mayer, R., and R. Moreno. 2010. Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning. In *Cognitive load; Theory and application*, ed. J. L. Plass, R. Moreno, and R. Brunken, 131–152. New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511844744.009

Miller, Charles. 2011. Aesthetics and e-assessment: The interplay of emotional design and learner performance. *Distance Education* 32(3): 307–337. http://dx.doi.org/10.1080/01587919.2011.610291

Miller, George A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2): 81–97. http://dx.doi.org/10.1037/h0043158

Miller, Heather, and Jacqueline Bichsel. 2004. Anxiety, working memory, gender, and math performance. *Personality and Individual Differences* 37(3): 591–606. http://dx.doi.org/10.1016/j.paid.2003.09.029

Nunnally, Jum C. 1978. *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Paas, Fred, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist* 38(1): 1–4. http://dx.doi.org/10.1207/S15326985EP3801_1

Rose, David, and Anne Meyer. 2000. Universal Design for Learning. *Journal of Special Education Technology* 15(1): 67–70.

Shaftel, Julia, Evelyn Belton-Kocher, Douglas Glasnapp, and John Poggio. 2006. The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities" *Educational Assessment* 11(2): 105–126. http://dx.doi.org/10.1207/s15326977ea1102_2

Seipp, Bettina. 1991. Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research* 4(1): 27–41. http://dx.doi.org/10.1080/08917779108248762

Spielberger, Charles Donald, and C. Drew Edwards. *State-trait Anxiety Inventory for Children: STAIC: How I Feel Questionnaire: Professional Manual*. Mind Garden, 1973.

Sweller, John. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2): 257–285. http://dx.doi.org/10.1207/s15516709cog1202_4

——. 1989. Cognitive technology: Some procedures for facilitating learning and problem solving in mathematics and science. *Journal of Educational Psychology* 81(4): 457–466. http://dx.doi.org/10.1037/0022-0663.81.4.457

Vytal, Katherine, Brian Cornwell, Nicole Arkin, and Christian Grillon. 2012. Describing the interplay between anxiety and cognition: from impaired performance under low cognitive load to reduced anxiety under high load. *Psychophysiology* 49(6): 842–852. http://dx.doi.org/10.1111/j.1469-8986.2012.01358.x

Zorin, Barbara, Patricia D. Hunsader, and Denisse R. Thompson. Assessments: Numbers, context, graphics, and assumptions. *Teaching Children Mathematics* 19, no. 8 (2013): 480–488. http://dx.doi.org/10.5951/teacchilmath.19.8.0480

DOI: http://dx.doi.org/10.5038/1936-4660.8.1.4