Journal of Educational Psychology
2008, Vol. 100, No. 1, 223–234

0022-0663/08/$12.00 DOI: 10.1037/0022-0663.100.1.223

A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load

Krista E. DeLeeuw and Richard E. Mayer
University of California, Santa Barbara

Understanding how to measure cognitive load is a fundamental challenge for cognitive load theory. In 2 experiments, 155 college students (ages = 17 to 22; 49 men and 106 women) with low domain knowledge learned from a multimedia lesson on electric motors. At 8 points during learning, their cognitive load was measured via self-report scales (mental effort ratings) and response time to a secondary visual monitoring task, and they completed a difficulty rating scale at the end of the lesson. Correlations among the three measures were generally low. Analyses of variance indicated that the response time measure was most sensitive to manipulations of extraneous processing (created by adding redundant text), effort ratings were most sensitive to manipulations of intrinsic processing (created by sentence complexity), and difficulty ratings were most sensitive to indications of germane processing (reflected by transfer test performance). Results are consistent with a triarchic theory of cognitive load in which different aspects of cognitive load may be tapped by different measures of cognitive load.

Keywords: cognitive load, measurement, education, multimedia, learning

Suppose a student viewed a narrated animation explaining how an electric motor works, such as that partially shown in the left panel of Figure 1. The lesson lasts about 6 min and explains how electricity flows from a battery and crosses a magnetic field, which in turn creates a force that moves a wire loop.

A major challenge in designing multimedia lessons such as the electric motor lesson is to be sensitive to the learner's cognitive load during learning. In particular, the lesson should be designed so that the amount of cognitive processing required for learning at any one time does not exceed the learner's processing capacity (Mayer, 2001, 2005a; Mayer & Moreno, 2003; Sweller, 1999, 2005). However, researchers have not yet reached consensus on how to measure cognitive load during learning or even whether the various cognitive load measures are tapping the same construct (Brünken, Plass, & Luetner, 2003; Paas, Tuovinen, Tabbers, & van Gerven, 2003).

According to a triarchic theory of cognitive load based on cognitive load theory (Sweller, 1999, 2005) and the cognitive theory of multimedia learning (Mayer, 2001, 2005a; Mayer & Moreno, 2003), there are three kinds of cognitive processing during learning that can contribute to cognitive load: (a) *extraneous processing*, in which the learner engages in cognitive processing that does not support the learning objective (and that is increased by poor layout such as having printed words on a page and their corresponding graphics on another page); (b) *intrinsic (or essential) processing*, in which the learner engages in cognitive processing that is essential for comprehending the material (and that depends on the complexity of material, namely the number of interacting elements that must be kept in mind at any one time); and (c) *germane (or generative) processing*, in which the learner engages in deep cognitive processing such as mentally organizing the material and relating it to prior knowledge (and that depends on the learner's motivation and prior knowledge, as well as prompts and support in the lesson). Table 1 gives examples of how each of these three facets of cognitive load can be manipulated within our multimedia learning situation involving a narrated animation about how an electric motor works.

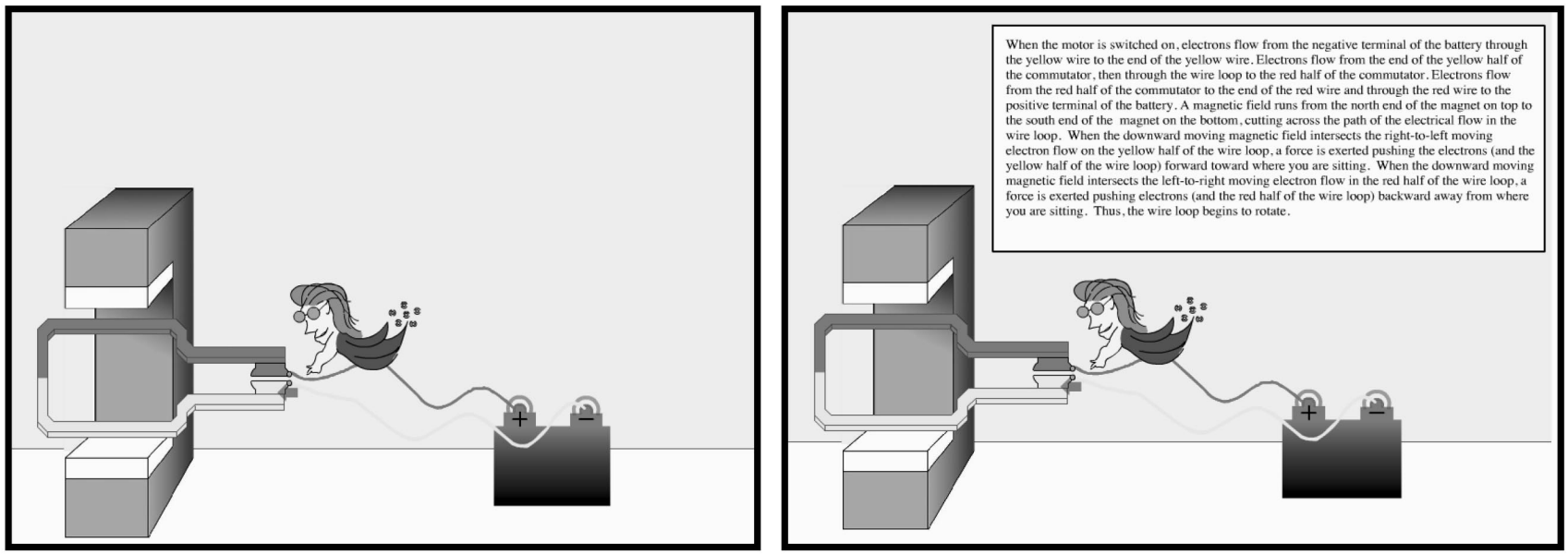
First, one way to manipulate extraneous processing in a multimedia lesson such as the electric motor lesson is through redundancy. A nonredundant lesson consists of presenting concurrent animation and narration, as summarized in the left panel of Figure 1; a redundant lesson consists of presenting concurrent animation, narration, and on-screen text, as summarized in the right panel of Figure 1. The on-screen text is redundant with the narration because both contain the same words and are presented at the same time. Redundancy can create extraneous cognitive load because the learner must expend precious cognitive capacity on reconciling the two verbal streams (i.e., checking to make sure the spoken words correspond to the printed words) and the learner must scan back and forth between the printed words and the animation (Mayer, 2005b). The processes of reconciling and scanning are forms of extraneous cognitive processing because they do not support the instructional objective (i.e., understanding how an electric motor works).

Second, one way to manipulate intrinsic processing in the electric motor lesson is through the complexity of the sentences. A low-complexity sentence (such as Sentence 1 in Figure 2) requires the learner to hold only a few concepts in working memory to understand the essential point, whereas a high-complexity sentence (such as Sentence 7 in Figure 2) requires the learner to hold many concepts in working memory to understand the essential point. Therefore, measures of cognitive load should reveal lower load after a low-complexity sentence than after a high-complexity sen-

Correspondence concerning this article should be addressed to Krista E. DeLeeuw or Richard E. Mayer, Department of Psychology, University of California, Santa Barbara, CA 93106. E-mail: deleeuw@psych.ucsb.edu or mayer@psych.ucsb.edu

Non-Redundant

Redundant



Narration: "When the motor is switched on, electrons flow from the negative terminal of the battery through the yellow wire to the end of the yellow wire. Electrons flow from the end of the yellow half of the commutator, then through the wire loop to the red half of the commutator…"

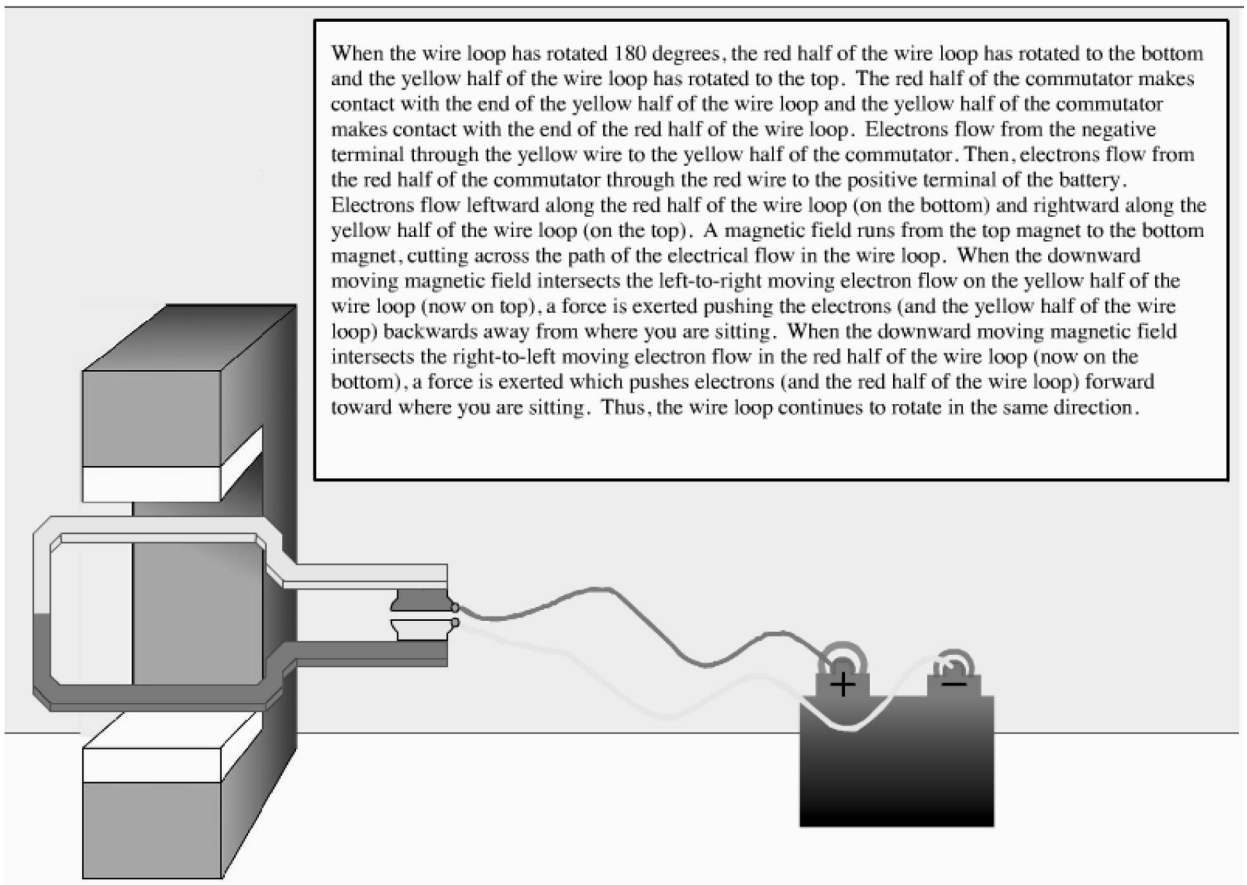
Figure 1. Frames from nonredundant and redundant lessons on how an electric motor works.

Table 1
Three Ways of Creating Cognitive Load in Multimedia Learning

Type of cognitive load	Example of cognitive load manipulation
Extraneous load	Redundancy: Redundant presentation has concurrent animation, narration, and on-screen text; nonredundant presentation has concurrent animation and narration.
Intrinsic load	Complexity: A high-complexity sentence involves many interacting concepts; a low-complexity sentence involves few interacting concepts.
Germane load	Transfer: Low-transfer students are less likely to have engaged in germane processing during learning; high-transfer students are more likely to have engaged in germane processing during learning.

tence. Sentence complexity can create intrinsic cognitive processing because the learner needs to coordinate more pieces of information essential to comprehension (Mayer, 2005a; Sweller, 1999).

Third, one way to examine differences in germane processing in the electric motor lesson is to compare students who score high on a subsequent test of problem-solving transfer with those who score low. High-transfer learners are more likely to have engaged in higher amounts of germane processing during learning than are low-transfer learners (Mayer, 2001, 2005a). Figure 3 shows a transfer test question along with answers from a



Low-complexity: Sentence 1. "When the loop has rotated 180 degrees, the red half of the wire loop has rotated to the bottom and the yellow half of the wire loop has rotated to the top."
High-complexity: Sentence 7. "When the downward moving magnetic field intersects with the left-to-right moving electron flow on the yellow half of the wire loop (now on top), a force is exerted pushing electrons (and the yellow half of the wire loop) backwards away from where you are sitting."

Figure 2. Examples of low-complexity and high-complexity sentences in a lesson on how an electric motor works.

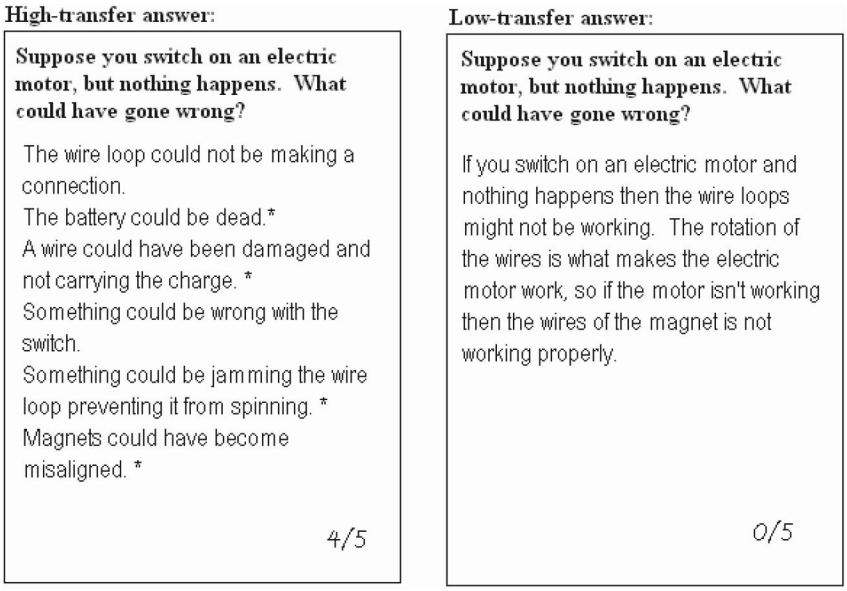


Figure 3. Examples of posttest answers from low-transfer and high-transfer students. Correct idea units are indicated by asterisks.

high-transfer student (in the left panel) and answers from a low-transfer student (in the right panel). Differences in transfer performance can be used to infer differences in germane processing during learning.

In this study, we examine the sensitivity of three commonly used techniques for measuring cognitive load—response time to a secondary task during learning, effort ratings during learning, and difficulty ratings after learning—to each of these three aspects of cognitive load. First, as shown in the first line of Table 2, we implemented a secondary task in our electric motor lesson. The secondary task was a visual monitoring task in which learners were asked to detect a periodic color change from pink to black in the background of the animation and to press the space bar as fast as possible each time this color change occurred. Response time has occasionally been used as a measure of cognitive load in multimedia research, with longer response times indicating greater cognitive load. When more cognitive resources are used by the primary task (learning from the narrated animation), fewer resources will be available to devote to the secondary task (noticing and responding to the background color change), resulting in a longer response time on the secondary task (Brünken, Steinbacher, Schnotz, Plass, & Leutner, 2002; Chandler & Sweller, 1996; Marcus, Cooper, & Sweller, 1996; Sweller, 1988).

Second, as shown in the second line of Table 2, mental effort ratings were implemented in our electric motor lesson. We asked the learner to rate his or her current level of mental effort at points throughout the lesson on a 9-point scale ranging from 1 (*extremely low*) to 9 (*extremely high*). Self-reported mental effort ratings have been used somewhat more frequently in multimedia research than dual-task measurement of load (Paas et al., 2003; Paas & van Merriënboer, 1994; Paas, van Merriënboer, & Adam, 1994).

Third, as shown in the bottom line of Table 2, difficulty ratings were implemented in our electric motor lesson by asking the learner to make a retrospective judgment after the lesson concerning the lesson's difficulty, using a 9-point scale ranging from 1 (*extremely easy*) to 9 (*extremely difficult*). Self-reported difficulty ratings have been used somewhat less frequently in multimedia research (Kalyuga, Chandler, & Sweller, 1999; Mayer & Chandler, 2001).

A major goal of this study is to determine whether these three measures of cognitive load tap the same underlying construct (suggesting a unitary theory of cognitive load) or whether they are

Table 2
Three Ways of Measuring Cognitive Load in Multimedia Learning

Type of measure	Implementation of measure
Response time to secondary task	At each of eight points in an animated narration, the background color slowly changes from pink to black, and the learner's task is to press the spacebar as soon as the background color changes.
Effort rating	At each of eight points in an animated narration the learner is asked to rate "your level of mental effort on this part of the lesson" on a 9-point scale ranging from 1 (*extremely low mental effort*) to 9 (*extremely high mental effort*).
Difficulty rating	At the end of the lesson, the learner is asked to rate "how difficult this lesson was" on a 9-point scale ranging from 1 (*extremely easy*) to 9 (*extremely difficult*).

sensitive to different manipulations of cognitive load (suggesting a model of cognitive load that is more consistent with a triarchic theory). According to a unitary theory of cognitive load, each of the three measures of cognitive load should be sensitive to each of the three types of cognitive load (i.e., indicating higher load for redundant vs. nonredundant groups, high-complexity vs. low-complexity sentences, and low-transfer vs. high-transfer learners), and the three should be highly correlated with one another. In other words, they should indicate one unitary measurement of cognitive load. According to a triarchic theory of cognitive load, it is possible that different measures of cognitive load may be sensitive to different types of cognitive load. If cognitive load is not a unitary construct and the measures are not all sensitive to the same type of load, then we should find that the three measures of cognitive load are not highly correlated with one another. If this is the case, the measures should be evaluated in terms of which aspect of the construct each measure best taps into. Predictions of possible ways in which the measures may relate to the three aspects of cognitive load in a triarchic theory are listed in Table 3. Alternatively, the possibility that the measures may gauge altogether different underlying constructs should be considered. We examined these questions in two experiments.

Experiment 1

In Experiment 1, we tested the predictions of the unitary theory of cognitive load and the triarchic theory of cognitive load by examining the sensitivity of each of three common metrics of cognitive load—response time to a secondary task, mental effort rating, and difficulty rating—to manipulations intended to create differences in extraneous, intrinsic, and germane processing.

Method

Participants and Design

The participants were 56 college students. There were 16 men and 40 women, and their ages ranged from 18 to 22. We used a 2×2 mixed design with the between-subjects factor being redundancy (redundant vs. nonredundant lesson) and the within-subject factor being complexity (high- vs. low-complexity sentences within the lesson). Concerning redundancy, 28 participants served in the nonredundant group (in which they received a computer-based lesson containing animation and narration), and 28 participants served in the redundant group (in which they received the same computer-based lesson containing the animation and narration plus on-screen text). Concerning complexity, four of the sentences in the lesson were identified as particularly high-complexity sentences (because they contained many interacting concepts), and four of the sentences in the lesson were identified as particularly low-complexity sentences (because they contained few interacting concepts). For each participant, cognitive load measurements were taken after each of the eight target sentences (for response time and effort rating) and after the entire lesson (for difficulty rating).

Materials and Apparatus

The computer-based materials consisted of a pretest questionnaire, a mental effort rating questionnaire, and two computer-based multimedia lessons on how an electric motor works. A Flash program designed for this experiment presented the multimedia lessons and collected data on the pretest questionnaire, response time to secondary tasks, and mental effort ratings. The paper-based materials consisted of a difficulty rating sheet and seven posttest problem sheets.

Pretest questionnaire. The computer-based pretest questionnaire assessed participants' previous experience with and knowledge about electronics and asked them to report their age, gender, SAT scores, and year in college.

Multimedia lesson. The nonredundant program (exemplified in the left portion of Figure 1) presented a narrated animation about how electric motors work that was used by Mayer, Dow, and Mayer (2003). The redundant version was identical except that it also presented on-screen text in paragraph form that was identical to the words spoken in the narration (as exemplified in the right panel of Figure 1). The on-screen text was redundant with the spoken narration, which is thought to increase extraneous cognitive load, reflected in lower scores on tests of transfer (Mayer, 2005b). Both versions of the lesson also included two levels of sentence complexity. We analyzed each sentence or clause in the script in terms of the number of interacting elements, or the number of idea units that must be kept in mind at one time for comprehension to take place. We then chose the four points of the lesson with the highest number of interacting elements, which we call *high sentence complexity*, and the four points with the lowest number of interacting elements, which we call *low sentence complexity*. We used these eight points during the lesson to implement the secondary task and the mental effort question.

The lesson was approximately 6 min long and described the cause-and-effect steps in the operation of an electric motor, which consisted of a battery, electrical wires, a commutator, a wire loop, and a pair of magnets.

Secondary task and mental effort rating. At eight points during the animation determined by the analysis of sentence complex-

Table 3
Possible Relations Among Three Measures of Cognitive Load and Three Types of Cognitive Load

Type of measure	Type of cognitive load
Response time	Extraneous—When material includes redundant words, the learner wastes cognitive capacity, resulting in slower response times to a secondary task.
Effort rating	Intrinsic—When material is more complex, the learner must work harder to comprehend it, resulting in higher effort ratings.
Difficulty rating	Germane—Learners who perform well on the transfer test had capacity available for deeper processing during learning, reflected in lower difficulty ratings.

ity, the background color of the animation gradually changed from pink to black. Participants were instructed to press the space bar as quickly as they could when they saw the color begin to change. Four of the color changes occurred at the end of high-complexity sentences and four occurred at the end of low-complexity sentences. The program recorded the students' response times (RTs; in milliseconds) to press the space bar for each of the eight events and calculated an average RT for high-complexity sentences and low-complexity sentences for each student. When the student pressed the space bar, the program paused the animation and presented a question regarding the amount of mental effort the participant was currently experiencing. The question asked participants to "please rate your level of mental effort on this part of the lesson" and gave them a Likert-type scale ranging from 1 (*extremely low mental effort*) to 9 (*extremely high mental effort*) from which to choose their response. The program computed and recorded the mean effort rating for each student for the four high-complexity sentences and the four low-complexity sentences. After indicating their answer by clicking the corresponding rating, participants had a choice of two buttons to press: continue or replay. Although participants believed that these buttons allowed them to either replay the last section of the animation or to continue from where they left off, the buttons actually performed identical functions; they both replayed the last sentence and continued onward. All students opted for the continue button.

Difficulty rating and posttest. The difficulty rating scale consisted of a sheet of paper containing the instruction "Please indicate how difficult this lesson was by checking the appropriate answer"; response choices ranged from 1 (*extremely easy*) to 9 (*extremely difficult*). The posttest consisted of seven sheets of paper, each with a question printed at the top and space for participants to write their answers. The questions tested participants' ability to transfer information they had learned about electric motors to problem-solving situations and were identical to those used by Mayer et al. (2003). The questions were (a) "What could you do to increase the speed of the electric motor, that is, to make the wire loop rotate more rapidly?" (b) "What could you do to increase the reliability of the electric motor, that is, to make sure it would not break down?" (c) "Suppose you switch on an electric motor, but nothing happens. What could have gone wrong?" (d) "What could you do to reverse the movement of the electric motor, that is, to make the wire loop rotate in the opposite direction?" (e) "Why does the wire loop move?" (e) "If there was no momentum, how far would the wire loop rotate when the motor is switched on?" and (f) "What happens if you move the magnets further apart? What happens if you connect a larger battery to the wires? What happens if you connect the negative terminal to the red wire and the positive terminal to the yellow wire?"

Apparatus. The apparatus consisted of five Sony Vaio laptop computers with 15-in. screens and Panasonic headphones.

Procedure

Participants were randomly assigned to the redundant or nonredundant group and were seated in a cubicle at a work desk with an individual laptop computer. There were between 1 and 5 participants in each session. Figure 4 summarizes the procedure. First, the experimenter briefly described the study, and each participant read and signed an informed consent form, which was collected by the experimenter. Second, the experimenter logged the participants in on their respective computers and asked them to complete the pretest questionnaire presented on the computer screen. Next, participants read on-screen instructions about how to perform the secondary task, and then they completed a practice task. The practice task was identical to the experimental dual task, except that it presented an animation on an unrelated topic (i.e., how brakes work). After the practice task, the experimenter inquired as to whether all participants were comfortable with the task, and participants were given a chance to ask questions. Once all participants indicated that they fully understood the task, they continued on to the experimental phase of the experiment in which the electric motor lesson was presented. Participants in the redundant group received the redundant version and participants in the nonredundant group received the nonredundant version. After the lesson, participants were given the difficulty rating sheet to complete at their own pace, followed by each of the seven posttest sheets. The posttest questions were presented one at a time in the order indicated in the *Materials and Apparatus* section, and participants were given a limit of 3 min to write their answers to each question.

Results and Discussion

Scoring

RT measurements more than 3 standard deviations from the mean of any of the RT measurements were replaced with that participant's series mean, with respect to whether the outlying RT occurred at a high- or low-complexity sentence. We opted to replace the outlying RT rather than simply exclude each outlying RT or the entire participant, with the assumptions that a RT of more than 3 standard deviations above the mean indicated the participant's inattention and that averaging across that participant's low- or high-complexity trials (depending on the type of trial in which the outlier occurred) would provide a sufficient estimate of how that participant would have performed on that trial had he or she given the task the appropriate attention. In cases in which more than one RT was an outlier within a participant's series, that participant was excluded from further analysis because there was not enough information to make a sufficient estimate of RT on those trials. This resulted in 2 participants being excluded, leaving 26 in the redundant group and 28 in the nonredundant group ($N = 54$).

Answers on the transfer test were coded for correct answers, out of a total possible 25. The list of acceptable answers was generated by Mayer et al. (2003). For example, the acceptable answers for the third question about what went wrong were "the wire loop is stuck," "the wire is severed or disconnected from the battery," "the battery fails to produce voltage," "the magnetic field does not intersect the wire loop," and "the wire loop does not make contact with the commutator." Students did not receive credit for vague answers, such as "Something is wrong with the battery," but did get credit for correct answers that were worded differently than in the lesson. For all but one question, there was more than one possible acceptable answer. Each acceptable answer generated by the participant earned 1 point; we summed the number of points across all of the questions to obtain a total transfer score, resulting in a range from 0 to 16 correct idea units, with a median of 6.5.

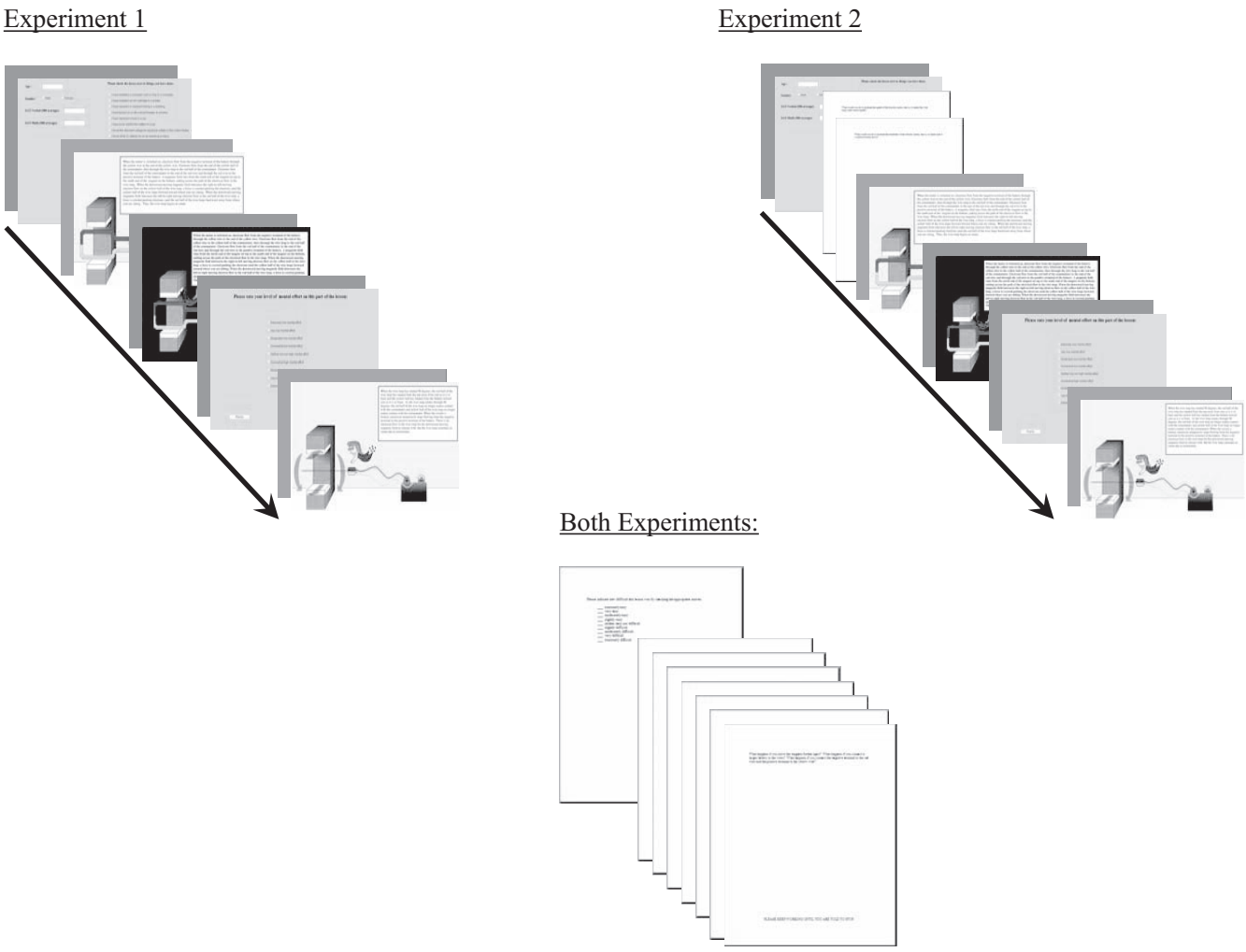


Figure 4. Summary of procedure.

Assignment to low- versus high-transfer groups was based on a median split in which students who scored 7 points or more were assigned to the high-transfer group ($n = 27$) and students who scored 6 points or less were assigned to the low-transfer group ($n = 27$).

Response Time to the Secondary Task

If response time to a secondary task during learning is a valid measure of cognitive load, it should differ on the basis of the manipulations thought to affect cognitive load. That is, response time should be longer for the redundant group than for the nonredundant group, on more complex sentences in the lesson than on less complex sentences, and for low-transfer learners than for high-transfer learners.

The first row of data in Table 4 shows the mean response times (and standard deviations), first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and finally for the low- and high-transfer learners. A box around two means (and their standard deviations) indicates that we found some difference between them, with bold lines indicating a significant difference at $p < .05$ and light lines indicating a nonsignificant difference (at $p < .10$) but an effect size greater than .20. We conducted a 2 (redundant vs. nonredundant) $\times$ 2 (low- vs. high-complexity sentences) mixed analysis of variance (ANOVA), with redundancy as a between-subjects factor and complexity as a repeated-measures factor. First, there was a marginal main effect for redundancy, with the redundant group producing marginally higher response times to the secondary task than the nonredundant group, $F(1, 52) = 3.77$, $p = .06$, $d = .53$. Although this trend was not statistically significant, the effect size is in the medium range, which suggests that the lack of significance may have been the result of a small sample size (Cohen, 1988). Second, there was a marginal main effect for complexity, with high-complexity sentences requiring marginally longer RTs to the secondary task than did low-complexity sentences, $F(1, 52) = 3.85$, $p = .06$, $d = .25$. Again, we note that although this trend was not statistically significant, the effect size was in the small range. No significant interaction was found between complexity and redundancy, $F(1, 52) = 0.004$, *ns*. Third, a t test revealed that students who scored low on transfer performance did not differ significantly on re-

Table 4
Means (and Standard Deviations) for Three Types of Cognitive Load Manipulations Based on Three Measures of Cognitive Load: Experiment 1

	Type of cognitive load					
	Extraneous load: Redundancy (Which cognitive load measure(s) is sensitive to redundancy?)		Intrinsic load: Complexity (Which cognitive load measure(s) is sensitive to sentence complexity?)		Germane load: Transfer (Which cognitive load measure(s) is sensitive to transfer performance?)	
Measure of cognitive load	Redundant	Nonredundant	High	Low	High (n = 27)	Low (n = 27)
Response time (ms)	2,657 (825)	2,249 (719)	2,555 (1,035)	2,337 (714)	2,477 (933)	2,414 (636)
Effort rating	4.97 (1.54)	5.49 (1.41)	5.43 (1.55)	5.05 (1.50)	5.30 (1.57)	5.18 (1.42)
Difficulty rating	5.15 (1.54)	5.36 (1.47)			4.63 (1.39)	5.89 (1.34)

Note. $N = 54$. Boxes with bold lines indicate significant difference ($p < .05$). Boxes with light lines indicate $.05 < p < .10$ and effect size greater than $d = .20$.

sponse time compared with students who scored high on transfer performance, $t(52) = -0.29$, *ns*. We furthermore computed a Cronbach's alpha for the four RTs at low-complexity points in the lesson and for the four RTs at high-complexity points to investigate the internal reliability of this measurement. Although the low-complexity points showed a low reliability ($\alpha = .33$), high-complexity points were shown to be reliable ($\alpha = .70$).

Mental Effort Ratings During Learning

For self-reported mental effort during a lesson to be considered a valid measure of cognitive load, several conditions should be met. First, lessons containing animation with redundant text and narration should cause learners to rate their mental effort higher overall than nonredundant lessons containing only animation and narration. Second, learners should rate their mental effort significantly higher at difficult (high-complexity) points in the lesson than at simpler (low-complexity) points in the lesson. Third, students who score low on transfer should rate their mental effort significantly higher overall than those who score high on transfer (alternatively, higher effort may indicate more germane processing, which should lead to higher scores on the transfer test).

The second row of data in Table 4 shows the mean mental effort ratings (and standard deviations) for each of three comparisons, first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and last for the low- and high-transfer learners. We conducted a 2 (redundant vs. nonredundant) $\times$ 2 (high- vs. low-complexity sentences) mixed ANOVA, with redundancy as a between-subjects factor and complexity as a repeated-measures factor. First, there was no significant effect for redundancy, with the redundant group and the nonredundant group rating their mental effort equivalently, $F(1, 52) = 1.67$, *ns*. Second, in contrast, there was a significant main effect of complexity in which participants rated their mental effort as higher on high-complexity sentences than on low-complexity sentences, $F(1, 52) = 17.78$, $p < .001$, $d = .25$. No significant interaction was found between redundancy and complexity, $F(1, 52) = 0.004$, *ns*. Third, a t test revealed that students who scored low on transfer performance did not differ significantly on mental effort ratings than students who scored high on transfer performance, $t(52) = -0.30$, *ns*. Cronbach's alpha indicated that measures of mental effort were reliable both at low-complexity points ($\alpha = .84$) and at high-complexity points ($\alpha = .90$).

Overall Difficulty Ratings After Learning

The last row of data in Table 4 shows the mean lesson difficulty ratings (and standard deviations) for each of two comparisons: between the redundant and nonredundant groups and between the low-transfer and high-transfer learners. There were no separate difficulty ratings for low and high complexity because this rating was only administered at the end of the lesson. If the difficulty rating is a valid measure of cognitive load, then the mean difficulty ratings should differ between redundant and nonredundant versions of that lesson. In addition, the mean difficulty rating of the low-transfer group should be greater than the mean difficulty rating of the high-transfer group. Finally, although we would expect learners to rate high-complexity sentences as more difficult than low-complexity sentences, there was no way to examine this because difficulty ratings were solicited only for the entire lesson after learning.

First, a t test showed that ratings of difficulty did not differ between the redundant and nonredundant groups, $t(52) = 0.50$, $p = .62$, which could indicate either that this rating scale is not sensitive to changes in extraneous load or that the redundant version of the multimedia lesson created no more cognitive load than the nonredundant version. The results of a second t test revealed that those who scored lower on the transfer test tended to rate the lesson as more difficult than those who scored higher, $t(52) = 3.39$, $p < .001$, $d = .92$.

Are Commonly Used Metrics of Cognitive Load Related to One Another?

If all of the methods we used to measure cognitive load did indeed measure the same construct, then we would expect to see significant correlations among all of the measures. Specifically, we would expect response time on a secondary task, self-reported mental effort, and difficulty rating to all be positively correlated. As shown in Table 5, however, we did not find this pattern of

Table 5
Correlation Matrix of Dependent Measures for Both Groups: Experiment 1

Measure	1	2	3	4
1. Dual-task reaction time	—	.27*	.20	.07
2. Self-report mental effort		—	.26	.19
3. Lesson difficulty rating			—	−.22
4. Score on transfer test				—

Note. $N = 54$.
* $p < .05$.

correlations. Of the six pairwise correlations, only one reached statistical significance; response time on the secondary task was significantly correlated with mental effort ($r = .27$, $p = .05$), but at a modest level. The correlation results do not support the contention that various cognitive load measures tend to measure the same thing, as proposed by the unitary theory of cognitive load.

Are Different Cognitive Load Measures Sensitive to Different Types of Cognitive Load?

Table 6 shows whether each cognitive load measure detected a significant difference for each type of cognitive load and the effect size indicated by each measure for each type of load. The pattern that emerges from these results suggests that each of the three measures—response time, effort rating, and difficulty rating—was sensitive mainly to one aspect of cognitive load. Response times were most sensitive to redundancy (which was intended to create extraneous cognitive processing), effort ratings during learning were most sensitive to complexity (which was intended to create intrinsic cognitive processing), and difficulty ratings after learning were most sensitive to transfer performance (which was intended to tap germane processing). Given the distinct pattern that was generated in Experiment 1, it is worthwhile to determine whether the pattern could be replicated in a similar study.

Overall, these commonly used measures of cognitive load do not appear to be related to one another in a clear-cut fashion. In contrast, the pattern of the results suggests that these measures may be tapping different aspects of cognitive load or different constructs altogether. The results are more consistent with the triarchic theory of cognitive load than with a unitary theory of cognitive load.

Experiment 2

In Experiment 1, an interesting pattern of results emerged in which each of the three measures of cognitive load tended to be sensitive to a different aspect of cognitive load. However, the many measures of cognitive load may have been somewhat intrusive to the learning task, resulting in measures that may reflect distraction rather than cognitive load per se. To encourage learners to focus on deep cognitive processing as their goal, in Experiment 2 we provided learners with *pretest questions* that we asked them to be able to answer at the end of the lesson. These pretest questions were shown by Mayer et al. (2003) to improve scores on the transfer test for this lesson. We reasoned that if students were more motivated to learn the information, they would pay closer attention to the lesson and therefore provide more valid measurements of cognitive load.

Method

Participants and Design

Participants in Experiment 2 were 99 college students (33 male, 66 female) ranging in age from 17 to 22. Half of the participants were randomly assigned to the nonredundant group ($n = 49$) and half to the redundant group ($n = 50$). The design of Experiment 2 was identical to that of Experiment 1.

Materials, Apparatus, and Procedure

The materials, apparatus, and procedure of this experiment were identical to those of Experiment 1, with the addition of two pretest questions. The pretest questions provided the participants with knowledge about the type of information they were expected to learn during the lesson. They were presented on an 8.5- × 11-in. sheet before the lesson and consisted of Questions a and b from the transfer test (i.e., "What could you do to increase the speed of the electric motor, that is, to make the wire loop rotate more rapidly?" and "What could you do to increase the reliability of the electric motor, that is, to make sure it would not break down?"). Participants were told that the pretest questions were "representative of the types of questions that would be asked later." Participants were allowed to look over the pretest question sheet until they felt comfortable with the material, at which time they handed the sheet back to the experimenter. Then the experimenter reminded participants of the instructions, answered any questions, and instructed

Table 6
Effect Sizes for Three Types of Cognitive Load Manipulation Created Based on Three Measures of Cognitive Load

	Type of cognitive load					
	Extraneous load (redundant vs. nonredundant)		Intrinsic load (high vs. low complexity)		Germane load (high vs. low transfer)	
Measure of cognitive load	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Response time	.53†	.48*	.25†	*ns*	*ns*	*ns*
Effort rating	*ns*	.35*	.25**	.16**	*ns*	*ns*
Difficulty rating	*ns*	*ns*			.92**	.48*

Note. Cohen's *d* is the measure of effect size. Difficulty rating does not apply to intrinsic load. Exp. = Experiment.
† $p < .10$. * $p < .05$. ** $p < .01$.

them to begin the computer program. From this point on, the procedure was identical to Experiment 1.

Results

Scoring

Outlying RTs to the secondary task were dealt with in the same manner as in Experiment 1. Again, participants who had more than two outlying RTs in a series were excluded from further analyses. This resulted in the exclusion of only 3 participants, leaving 47 in the nonredundant group and 49 in the redundant group ($N = 96$). Answers on the transfer test were coded in the same fashion as in Experiment 1.

Scores on the transfer test ranged from 0 to 13 correct idea units, with a median of 6.0, out of a possible total of 25. A subset of the data was coded by a second rater, resulting in a correlation of .772 ($p = .009$). A median split resulted in 45 high scorers (scoring 7 points or more) and 51 low scorers (scoring 6 points or less).

RT to the Secondary Task

As in Experiment 1, we expected RT on the secondary task to be longer for the redundant group than for the nonredundant group, longer for high-complexity sentences in the lesson than for low-complexity sentences, and longer for low-transfer learners than for high-transfer learners.

The first row of data in Table 7 shows the mean RTs (and standard deviations) for three comparisons, first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and last for the low- and high-transfer learners. As in Table 4, a box around two means (and their standard deviations) indicates that we found some difference between them, with bold lines indicating a significant difference. We conducted a 2 (redundant vs. nonredundant) $\times$ 2 (low vs. high complexity) mixed ANOVA with redundancy as a between-subjects factor and complexity as a repeated measures factor. There was a significant main effect for redundancy, with the redundant group producing longer RTs than the nonredundant group, $F(1, 94) = 5.44$, $p = .02$, $d = .48$. However, no main effect of complexity was found, $F(1, 94) = 1.67$, *ns*, nor was there a significant interaction between trial type and redundancy, $F(1, 94) = 0.93$, *ns*. These results show that redundant on-screen text caused longer RTs and indicate that participants who saw the nonredundant version had more free cognitive resources. However, unlike in Experiment 1, RT did not appear to be sensitive to the number of interacting elements (i.e., sentence complexity) at a given point in the lesson. Finally, a *t* test showed that there was no significant difference between low- and high-transfer learners on their RTs to the secondary task, $t(94) = 1.68$, *ns*. Cronbach's alpha showed internal reliability for RT measurements both at low-complexity points ($\alpha = .79$) and at high-complexity points ($\alpha = .76$).

Mental Effort Rating During Learning

As in Experiment 1, on the basis of unitary theory, we expected mental effort ratings to be higher (indicating more mental effort expended) for the students learning from the redundant lesson than for those learning from the nonredundant lesson. We also expected learners to rate their mental effort higher at difficult (high-complexity) sentences in the lesson than at easier (low-complexity) sentences in the lesson. Finally, we expected that learners who scored low on the transfer test would rate their mental effort higher overall than those who scored high on the transfer test.

The second row of data in Table 7 shows the mean RTs (and standard deviations) for three comparisons, first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and last for the low- and high-transfer learners. We conducted a 2 (redundant vs. nonredundant) $\times$ 2 (high vs. low complexity) mixed ANOVA with redundancy as a between-subjects factor and trial type as a repeated measures within-subjects factor. First, unlike in Experiment 1, ratings of mental effort differed significantly between the redundant and nonredundant lessons, $F(1, 94) = 4.17$, $p = .04$, $d = .35$. Second, we again found a significant main effect of complexity; learners rated their mental effort as higher on high-complexity sentences than on low-complexity sentences, $F(1, 94) = 20.36$, $p < .001$, $d = .16$. No significant interaction of complexity and redundancy was found, $F(1, 94) = 0.21$, *ns*. Third, a *t* test showed that learners who scored low on the transfer test had overall mental effort ratings similar to those who scored high on the transfer test,

Table 7
Means (and Standard Deviations) for Three Types of Cognitive Load Manipulations Based on Three Measures of Cognitive Load: Experiment 2

	Type of cognitive load					
	Extraneous load: Redundancy (Which cognitive load measure(s) is sensitive to redundancy?)		Intrinsic load: Complexity (Which cognitive load measure(s) is sensitive to sentence complexity?)		Germane load: Transfer (Which cognitive load measure(s) is sensitive to transfer performance?)	
Measure of cognitive load	Redundant	Nonredundant	High	Low	High ($n = 45$)	Low ($n = 51$)
Response time (ms)	2,918 (872)	2,520 (797)	2,677 (869)	2,769 (974)	2,569 (848)	2,859 (847)
Effort rating	5.67 (1.59)	4.99 (1.67)	5.47 (1.66)	5.21 (1.68)	5.58 (1.58)	5.13 (1.71)
Difficulty rating	5.33 (1.83)	5.21 (1.74)			4.82 (1.92)	5.67 (1.56)

Note. Boxes with bold lines indicate significant difference ($p < .05$).

$t(94) = -1.31$, *ns*. Cronbach's alpha showed internal reliability for mental effort measurements both at low-complexity points (α = .90) and at high-complexity points (α = .90).

Overall Difficulty Rating After Learning

As in Experiment 1, according to the unitary theory, we expected learners in the redundant group to rate the difficulty of the lesson as higher than those in the nonredundant group. In addition, we expected that learners who scored low on the transfer test would rate the lesson as more difficult than those who scored high.

The last row of data in Table 7 shows the mean lesson difficulty ratings (and standard deviations) for two comparisons—for the redundant and nonredundant groups and for the low- and high-transfer learners. There were no separate difficulty rating scores for low and high complexity because this rating was only administered at the end of the lesson. First, there was no significant difference between the redundant and nonredundant groups; learners in both groups rated the difficulty of the lesson similarly, $t(94) = -0.31$, *ns*. In contrast, there was a significant difference between high- and low-transfer learners; learners who scored low on the transfer test tended to rate the lesson as more difficult than those who scored high, $t(94) = 2.38$, $p = .02$, $d = .48$.

Are Commonly Used Metrics of Cognitive Load Related to One Another?

Table 8 shows the correlations among the dependent variables. If these methods are all measuring a unitary or overall level of cognitive load, we would expect them all to be significantly positively correlated with one another. However, RT to the secondary task was not positively correlated with any of the other measures of cognitive load. It was, however, significantly negatively correlated with scores on the transfer test ($r = -.30$, $p = .003$), indicating that participants who scored lower on the transfer test tended to take longer to respond to the secondary task, which is the result we would expect if cognitive load were causing a slower RT. Mental effort ratings were positively correlated with difficulty ratings ($r = .33$, $p = .001$) but with no other measure. Difficulty ratings were negatively correlated with transfer scores ($r = -.22$, $p = .03$) but with no other measure. As can be seen by comparing Tables 5 and 8, the measures were more correlated with one another in Experiment 2 than in Experiment 1, but most of the relations were weak or nonsignificant.

Table 8
Correlations Between Dependent Measures for Both Groups: Experiment 2

Measure	1	2	3	4
1. Dual-task reaction time	—	.12	.13	−.30**
2. Self-reported mental effort		—	.33**	.11
3. Lesson difficulty rating			—	−.22*
4. Score on transfer test				—

Note. $N = 96$.
* $p < .05$. ** $p < .01$.

Are Different Cognitive Load Measures Sensitive to Different Types of Cognitive Load?

Table 6 shows whether each cognitive load measure detected a significant difference for each type of cognitive load and the effect size indicated by each cognitive load measure for each type of cognitive load. The pattern of results for Experiment 2 is quite similar to that in Experiment 1.

In Experiment 2, it was again apparent that RTs on the secondary task were most sensitive to redundancy than to other manipulations of cognitive load. Because redundancy was intended to create extraneous cognitive load, we can conclude that RTs were most sensitive to differences in extraneous load. Ratings of mental effort were most sensitive to the complexity of the sentences, which was intended to create intrinsic cognitive load. Finally, ratings of the overall difficulty of the lesson were most sensitive to differences in the learning outcomes of the students, in terms of their scores on the transfer test, which is an indication of germane cognitive load. Taken together with the findings from Experiment 1, these results show that these measures are tapping separable aspects of cognitive load.

Conclusion

Summary of Results

Across two experiments, we found that different measures of cognitive load were sensitive to different types of cognitive load: (a) RT to the secondary task was most sensitive to manipulations of extraneous processing (reflected in longer RTs for the redundant group than for the nonredundant group), (b) effort ratings during learning were most sensitive to manipulations of intrinsic processing (reflected in higher effort attributed to high-complexity sentences than to low-complexity sentences), and (c) difficulty ratings after learning were most sensitive to differences related to germane processing (reflected in higher difficulty reported by low-transfer learners than by high-transfer learners). In both experiments, we also found that different measures of cognitive load were not highly correlated with one another. This work is consistent with recent findings reported by Ayres (2006), in which a subjective rating measure administered after each problem was sensitive to differences in intrinsic processing, and by Brünken, Plass, and Leutner (2004), in which a dual-task measure was sensitive to differences in extraneous processing.

Theoretical Implications

If cognitive load is a unitary construct, reflecting an overall amount of cognitive resources allocated to the task, all types of manipulations of the learning situation (be it a manipulation of the study materials or of the motivation of the learner, etc.) should cause a corresponding change in the amount of load, and all measures that are directly related to cognitive load should correlate with one another. However, if cognitive load is composed of, or influenced by, different elements, as proposed by Sweller (1999) and Mayer (2001), then different manipulations of the learning situation can cause different types of cognitive load to vary in distinguishable ways. In this case, it may be possible that some measures are more sensitive to one type of change in cognitive load than to others.

On the theoretical side, this pattern of results is most consistent with a triarchic theory of cognitive load, in which cognitive processing during learning is analyzed into three capacity-demanding components—extraneous processing, intrinsic (or essential) processing, and germane (or generative) processing. More important, this study provides empirical support, replicated across two experiments, for a dissociation among these three types of cognitive load. The results are not consistent with a unitary theory of cognitive load, which focuses on the overall amount of cognitive processing during learning.

Practical Implications

On the practical side, the results provide some validation for each of the three measures of cognitive load—RT to a secondary task, mental effort rating during learning, and difficulty rating made retrospectively immediately following learning—because each measure was sensitive to a particular cognitive load manipulation. An important practical implication is that different measures of cognitive load should not be assumed to measure overall cognitive load, but may be effectively used to measure different types of cognitive load. In particular, when the goal is to assess the level of extraneous cognitive load, RT to a secondary task appears to be most appropriate; when the goal is to assess the level of intrinsic cognitive load, mental effort ratings during learning may be most appropriate; and when the goal is to detect the learner's level of germane cognitive load, a simple difficulty rating immediately after learning may prove most useful.

Limitations and Future Directions

Although a similar pattern of results was obtained across two experiments, there is still a need for replication studies involving different instructional materials and different learners. It is possible that given different learning materials or different learner characteristics, we may find a different pattern of results. Regarding the intercorrelation of the measures in the two experiments, it is possible that we found mostly small, nonsignificant effects in Experiment 1 because of a small sample size, but this argument does not hold for Experiment 2 because the sample size was nearly double.

Consistent with research on the expertise reversal effect (Kalyuga, 2005), we suspect that the pattern of results might depend on the prior knowledge of the learner. In our experiments, the learners generally had low prior knowledge, so we suggest that future research should also take the learner's level of prior knowledge into account.

In this study, we focused on three specific measures of cognitive load, but there are alternative ways to implement each measure and there are certainly other measures of cognitive load. Similarly, we examined three cognitive load manipulations, but alternative methods of manipulating cognitive load should still be investigated using similar measures of load.

One problem with the current study is that the various measures of cognitive load were somewhat intrusive and may have created an unnatural learning situation. Finding unobtrusive and valid measures of each type of cognitive load continues to be a challenge for multimedia researchers. However, this challenge is worth meeting because the concept of cognitive load plays a central role in most theories of instructional design.

References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16,* 389–400.

Brünken, R., Plass, J. L., & Luetner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38,* 53–62.

Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 32,* 115–132.

Brünken, R., Steinbacher, S., Schnotz, W., Plass, J., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49,* 109–119.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10,* 151–170.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kalyuga, S. (2005). Prior knowledge principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 325–338). New York: Cambridge University Press.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia learning. *Applied Cognitive Psychology, 13,* 351–371.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology, 88,* 49–63.

Mayer, R. E. (2001). *Multimedia learning.* New York: Cambridge University Press.

Mayer, R. E. (2005a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 31–48). New York: Cambridge University Press.

Mayer, R. E. (2005b). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 183–200). New York: Cambridge University Press.

Mayer, R. E., & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology, 93,* 390–397.

Mayer, R. E., Dow, G., Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology, 95,* 806–812.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38,* 43–52.

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63–72.

Paas, F., & van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6,* 51–71.

Paas, F., van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual & Motor Skills, 79,* 419–430.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285.

Sweller, J. (1999). *Instructional design in technical areas.* Camberwell, Victoria, Australia: ACER Press.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 19–30). New York: Cambridge University Press.

Received December 12, 2006
Revision received August 22, 2007
Accepted September 4, 2007 ■