

A Characterization of Calculus I Final Exams in U.S. Colleges and Universities

Abstract

In this study, we developed a three-dimensional framework to classify post-secondary Calculus I final exams. Our *Exam Characterization Framework* (ECF) classifies individual exam items according to the cognitive demand required to answer the item, the representational context in which the item is asked, and the format of the item. Our results from using the ECF to code 150 post-secondary Calculus I final exams from across the United States revealed that the exams generally require low levels of cognitive demand, seldom contain problems stated in a real-world context, rarely elicit explanation, and do not require students to demonstrate or apply their understanding of the course's central ideas. We compared the results from analyzing individual instructor's exams with survey data of instructor's beliefs about the conceptual orientation of their exams. Additionally, we compared the characteristics of our sample of Calculus I final exams with post-secondary Calculus I final exams administered in 1986/87.

Keywords: Calculus; Assessment; Mathematical reasoning; University level mathematics

Introduction

Course exams are among the most revealing written artifacts of the mathematical skills and understandings instructors want their students to acquire in a mathematics course. A course exam provides information about an instructor's expectations for students' level of computational fluency, their depth of understanding specific concepts, and the degree to which students are expected to make connections among the course's central ideas.

Our review of the literature related to Calculus I assessment revealed that little is known about the content of Calculus I exams administered in colleges and universities in the United States. Several studies have characterized items on mathematics exams and in textbooks according to their conceptual focus (Bergqvist, 2007; Boesen, Lithner, & Palm, 2006; Gierl, 1997; Li, 2000; Lithner, 2000, 2003, 2004; Mesa et al., 2012; Palm, Boesen, & Lithner, 2006), while others have examined the format of exam questions (Senk, Beckman, & Thompson, 1997), and the degree to which students can solve problems by imitating procedures (Bergqvist, 2007; Lithner, 2000, 2004). These approaches, however, characterize a small sample of exam and textbook items and therefore provide a limited snapshot of the mathematics valued by instructors or curriculum developers. The present study provides one response to this gap in the literature by characterizing a large number of final exams from first-semester calculus courses at a variety of post-secondary U.S. institutions.

We begin this paper by outlining our study's context. We then chronicle the development of our *Exam Characterization Framework* (ECF), including details on how the existing literature informed its development and how iterative coding led to its refinement. Next, we describe the three strands of the ECF and provide examples of coded items to clarify potentially problematic interpretations. We then present the results from coding our sample of 150 post-secondary

Calculus I final exams, collectively containing 3,735 items. This is followed by a presentation of results from comparing instructors' perceptions of their exams with our characterization of them. We conclude by comparing our characterization of the sample of 150 exams with the cognitive demand of a sample of Calculus I final exams from 1986/1987 administered at 13 different colleges and universities in the United States (Steen, 1988).

Context of Study

This study is a part of a larger initiative by the Mathematical Association of America to determine the characteristics of successful programs in college calculus. As part of a larger data corpus, faculty from 253 universities electronically submitted exams and instructor surveys. Of these 253 exams, we randomly selected 150 for use in this study. Of the instructors providing data for the larger data corpus, 48% are tenured or tenure-track faculty, 28% are other full-time faculty, 9% are part-time faculty, and 15% are graduate students. Moreover, of the 150 exams we randomly selected, 61.9% were administered at national universities, 22.7% at regional universities, 9.4% at community colleges, 4.6% at national liberal arts colleges, and 1.4% at regional colleges.¹

Literature that Informed the Development of the Exam Characterization Framework

¹ National universities are those that offer a full range of undergraduate majors as well as a host of master's and doctoral degrees. Regional universities offer a full range of undergraduate programs, some master's programs, and few doctoral programs. National Liberal Arts Colleges are schools that emphasize undergraduate education and award at least half of their degrees in the liberal arts fields of study. Regional colleges are institutions focusing on undergraduate education, awarding less than half of their degrees in liberal arts fields of study. We used the university classifications of U.S. News & World Report (www.usnews.com/education).

A major product of this study is the *Exam Characterization Framework* (ECF) that provides a means by which one can code mathematics exam items and achieve a reasonably objective characterization. It was our goal to locate or create a framework that would characterize items relative to their conceptual versus procedural orientation, the representational context in which the questions were asked (e.g., word problems, formulas, graphs) and the format type of the questions. In the first stage of developing our framework we examined other frameworks (e.g., Li, 2000; Lithner, 2004; Mesa et al., 2012; Smith et al., 1996) that have been used to characterize items on exams and in textbooks. While our review of the literature allowed us to identify issues that were important for us to consider, it did not uncover a framework that was suited for our purpose. The ECF framework emerged from 12 cycles of our scoring exam items, refining our characterization of ECF constructs, and rescoring items until we were satisfied that our coding constructs were accurately characterizing the exam items in our sample.

Item Coding Frameworks

Li (2000) characterized items focused on the addition and subtraction of integers in American and Chinese mathematics textbooks. Her examination of these textbook items identified three dimensions of problem requirements: (a) *mathematical feature*, (b) *contextual feature*, and (c) *performance requirements*. The mathematical feature strand refers to the number of computational procedures needed to solve a problem. Items were coded as either *single computation procedure required* (S) or *multiple computation procedures required* (M). The contextual feature dimension of her framework refers to the nature of contextual information contained in a problem statement, such as whether the problem statement was situated in a *purely mathematical context in numerical or word form* (PM) or *illustrative context with pictorial representation or story* (IC) (Li, 2000, p. 237). The performance requirement dimension of Li's

framework is further partitioned into two categories: (a) *response type*, and (b) *cognitive requirement*. Response type refers to the contextual feature of the solution that a task elicits whereas cognitive requirement refers to the mental act that one engages in while solving a task.

Lithner (2004) developed a cognitive framework to classify calculus textbook items according to various strategies that could be employed to complete the item. His interest in coding textbook items was based on the belief that most calculus textbook items could be completed without considering the intrinsic mathematical properties of the item. Lithner's framework includes six reasoning practices one may employ when solving calculus textbook exercises, including reasoning based on identification of similarities, reasoning based on what seems to be true based on past experiences, and repeated algorithmic reasoning.

Smith et al. (1996) propose a taxonomy for classifying the cognitive demand of assessment tasks that evolved from the six intellectual behaviors in the cognitive domain of Bloom's taxonomy (Bloom et al., 1956).² Smith et al. adapted Bloom's initial taxonomy for use in a mathematical context and proposed the taxonomy categories: (1) *factual knowledge*; (2) *comprehension*; (3) *routine use of procedures*; (4) *information transfer*; (5) *application in new situations*; (6) *justifying and interpreting*; (7) *implications, conjectures, and comparisons*; and (8) *evaluation*. The primary purpose of their taxonomy was to assist instructors in writing exam items.

Anderson and Krathwohl (2001) developed a revision of Bloom's taxonomy (Bloom et al., 1956) to reflect developments in cognitive psychology occurring after the publication of Bloom's original taxonomy. They also modified the categories in the cognitive domain of

² These six intellectual behaviors include: *knowledge, comprehension, application, analysis, synthesis, and evaluation*.

Bloom's original taxonomy to make the revised taxonomy more suited for classifying teaching goals (e.g., defining learning objectives).

Mesa et al. (2012) analyzed the characteristics of examples in college algebra textbooks along the four dimensions proposed by Charalambous, Delaney, Hsu, and Mesa (2010): (1) *cognitive demand*, (2) *the expected response*, (3) *the use of representations*, and (4) *the strategies available for verifying the correctness and appropriateness of the solution*. The cognitive demand dimension consists of the categories: *memorization*, *procedures without connections*, *procedures with connections*, and *doing mathematics*. Tasks that require the cognitive behavior of memorizing can be solved by simply recalling information from memory without enacting a procedure. Tasks that require students to employ procedures without connections can be solved by applying rehearsed procedures without attending to connections to other mathematical ideas or real world contexts. Tasks that require students to employ procedures with connections ask the students to apply a procedure in which they draw connections to other mathematical ideas or real world contexts. Finally, tasks that expect students to do mathematics require the student to investigate fundamental mathematics concepts that promote the generation of new knowledge.

The expected response dimension of Mesa et al.'s analytical framework refers to the solution that an example provides by identifying whether the solution only offers an answer, gives an answer with an explanation or justification of the process undertaken to arrive at that answer, or contains both an answer and a mathematical sentence. Moreover, Mesa et al. identified the representation of the statement of the example and the solution provided according to five representation categories: *symbols*, *tables*, *graphs*, *numbers*, and *verbalizations*.

Development of the Exam Characterization Framework

Our development of the *Exam Characterization Framework* (ECF) required that we first familiarize ourselves with our sample of 150 Calculus I final exams. In doing so we identified items that we conjectured might be difficult to characterize relative to the level of cognition required for a correct response. Since we were most interested in characterizing the cognitive demand of exam items, our initial coding attempts involved coding these items according to the level of cognitive demand required by an item. We chose to call this dimension *item orientation*.

Our initial attempt to apply the three cognitive requirements in Li's (2000) framework (mathematical feature, contextual feature, and performance requirements) in our coding revealed that these constructs were too coarsely defined and resulted in qualitative variation of items within each category with respect to the cognitive demand needed to respond to the task. When attempting to use the various reasoning practices defined in Lithner's (2004) framework, we found that without knowledge of students' prior experiences, the boundaries of each classification lacked the specificity required to code items with a high degree of reliability. However, we did adopt Lithner's approach to classifying items based on the cognitive demand that was *necessary* for providing a correct response to the item. Attending to the cognitive demand required by a task was necessary to achieve a reliable characterization of exam items since we did not have knowledge of students' experiences in their calculus courses.

We became aware of Mesa et al.'s framework after having developed the current draft of the ECF, but decided to consider her constructs for the purpose of refining our framework, especially since her dimensions had some similarities with ECF. We attempted to code select exam items using the cognitive demand dimension of Mesa et al.'s (2012) framework, but found their characterization of the cognitive demand of exam items more related to desired understandings instead of the understandings that a task demands. Moreover, we had difficulty

agreeing on precisely what constitutes a connection for an exam item. While discerning whether or not exam items evoked connections, we were unable to reliably categorize items within the “procedures without connections” and “procedures with connections” categories of the cognitive demand dimension of Mesa et al.’s framework.

The categories of Smith et al.’s (1996) framework, which evolved from Bloom’s original taxonomy (Bloom et al., 1956), appealed to us, but because they were designed to assist instructors in writing exam items, their categories lacked the hierarchical structure that a diagnostic taxonomy typically requires. Therefore, Anderson and Krathwohl’s (2001) modification of the six intellectual behaviors in the cognitive domain of Bloom’s taxonomy was the most helpful for informing our approach to coding the cognitive demand of exam items. Anderson and Krathwohl (2001) developed their revision of Bloom’s taxonomy to reflect developments in cognitive psychology occurring after the publication of Bloom’s original taxonomy, and to make the taxonomy more suited to the purposes of classroom teaching (e.g., defining learning objectives). They characterize the levels of their taxonomy using the verbs: *remembering, understanding, applying, analyzing, evaluating, and creating*.

Since we were evaluating tasks rather than actual behaviors, our definitions needed to focus exclusively on the nature of the exam items, leading us to consider the categories: *remember, understand, apply, analyze, evaluate, and create*. Coding the cognitive demand of calculus exam items using these constructs revealed that some items (e.g., determine the derivative of f) required students to use procedural skills while requiring no understanding of the concept(s) on which the skill is based. This required a different cognitive behavior than simply recalling information, and led to our introducing the construct *recall and apply procedure* to Anderson and Krathwohl’s levels. Similarly, our preliminary coding revealed that one could

demonstrate an understanding of a concept without applying the understanding to achieve a goal or solve a problem, resulting in our introducing the level *apply understanding*.

To classify exam items relative to the cognitive demand they elicit, we have adapted the six intellectual behaviors in the conceptual knowledge dimension of Anderson and Krathwohl's modified Bloom's taxonomy to (1) reflect the variety of cognitive behaviors elicited by mathematics tasks, and (2) accurately characterize the cognitive demand of mathematics tasks in the absence of information about students' prior knowledge for whom the tasks were intended. These constructs, and our use of them, are described in more detail in the next section.

In the process of coding items, we found it relatively straightforward to specify the representation in which mathematics items are asked. Consistent with Li (2000) and Mesa et al. (2012), we also found it useful to code both the representation of the problem statement and the representation of the solution. We noted a representation type in the solution only if it was *necessary* to solve the problem or complete the task.

The mathematical feature dimension of Li's framework led us to consider the coding dimension *item density*. We defined a *dense item* as one that necessitates the incorporation of multiple concepts, procedures, and skills. However, since calculus tasks are often subject to being solved in a variety of ways, we were unable to specify how many procedures and concepts were needed to solve a specific calculus problem. Moreover, the grain-size of "procedure" has a wide range of interpretations, thus making coding problematic for more complicated tasks. We thus determined that our attempts to code items according to the number of procedures and concepts that are needed to solve a problem was less important than the overall cognitive demand of the task. As a result, we abandoned the item density dimension early in our framework development.

We also considered the fourth dimension of Mesa et al.'s framework that identifies the strategies needed for students to employ *control* (i.e., verify the correctness and appropriateness of a solution) during the process of completing a task. Our efforts to code for control led us to recognize that control is exercised idiosyncratically since it is initiated by a student's assessment of the validity and appropriateness of their *own* work. For example, one student may take the time to reflect on the reasonableness of computations when working a multi-step applied problem, while another student may continue with a sequence of memorized steps and pay no attention to computations that produce unreasonable results. There are, however, instances in which students are more likely to exhibit control. Two such instances include when students are asked to provide an explanation or justification for their work. Mesa et al. refer to such actions as *further elaboration*. We conjecture that the act of providing an explanation or justification may require one to reflect upon and assess the correctness of his or her work. However, if the teacher had the student practice justifying a response or solution, in homework or during class, it is possible that the student did not exercise control when responding to the item on an exam. Therefore, instead of coding for a mental process that may or may not be exercised when responding to an item, we have decided to not code for control, but rather to add a dimension in the *item format* category for noting instances in which exam items require students to explain or justify their response.

Exam Characterization Framework

The Exam Characterization Framework characterizes exam items according to three distinct item attributes: (a) *item orientation*, (b) *item representation*, and (c) *item format*. The first two authors independently coded five Calculus I exams while making note of each exam item's characteristics. We then met to compare and refine our characterizations so as to ensure

that our meanings were consistent. After at least ten cycles of coding and meeting to compare and refine our characterizations, we emerged with a stable framework that included categories that we found to be most useful for characterizing Calculus I final exam items. Two other coders coded ten randomly selected exams for the purpose of establishing the readability of our construct descriptions. After we refined the framework, the lead author independently coded all 150 exams in the sample.

The item orientation dimension of the framework includes the following seven categories of intellectual behaviors needed to respond to an exam item: *remember*, *recall and apply procedure*, *understand*, *apply understanding*, *analyze*, *evaluate*, and *create*. When coding items for their representation type, we coded representations present in the question or problem statement. We also coded for the representations required when responding to a question or working a problem. This dimension of our framework includes the categories: *applied/modeling*, *symbolic*, *tabular*, *graphical*, *definition/theorem*, *proof*, *example/counterexample*, and *explanation*. The third dimension of our framework is item format, with the categories *multiple choice*, *short answer*, and *broad open-ended*. When coding the format of the item, we also noted whether the item required students to provide a justification or explanation.

Item Orientation

The six classifications of the item orientation taxonomy are hierarchical, as is Bloom's taxonomy and Anderson and Krathwohl's (2001) modification of Bloom's taxonomy. The lowest level requires students to remember information and the highest level requires students to make connections (see Table 1). When coding items using the item orientation categories, we classified an item at a specific level only when this cognitive behavior was *necessary* for responding to the item. Even if an item may have been designed to assess a student's

understanding of an idea, if a student is able to solve the problem by applying a memorized procedure, then we classified the item as “recall and apply a procedure.” As emphasized above, since we had no information about students’ experiences in their calculus courses, we attended only to the cognitive behavior that a task *required*. We also made a distinction between items that require students to understand an idea and ones that require students to apply their understanding of an idea or concept (e.g., derivative) to solve the problem. The intellectual behaviors in the item orientation taxonomy of the ECF are described in more detail in Table 1.

Table 1

Adaptation of six intellectual behaviors from Anderson & Krathwohl (2001)

Cognitive Behavior	Description
Remember	Students are prompted to retrieve knowledge from long-term memory (e.g., write the definition of the derivative).
Recall and apply procedure	Students must recognize what procedures to recall when directly prompted to do so in the context of a problem (e.g., find the derivative/limit/antiderivative of f).
Understand	Students are prompted to make interpretations, provide explanations, make comparisons or make inferences that require an understanding of a mathematics concept.
Apply understanding	Students must recognize when to apply a concept when responding to a question or when working a problem. To recognize the need to apply, execute or implement a concept in the context of working a problem requires an understanding of the concept.
Analyze	Students are prompted to break material into constituent parts and

	determine how parts relate to one another and to an overall structure or purpose. Differentiating, organizing, and attributing are characteristic cognitive processes at this level (Krathwohl, 2002, p. 215).
Evaluate	Students are prompted to make judgments based on criteria and standards. Checking and critiquing are characteristic cognitive processes at this level (Krathwohl, 2002, p. 215).
Create	Students are prompted to put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure. Generating, planning, and producing are characteristic cognitive processes at this level (Krathwohl, 2002, p. 215).

Characterizing the Item Orientation Dimensions

This subsection presents examples of Calculus I items that require the various cognitive behaviors within the item orientation taxonomy of the ECF. We provide examples from the first five categories of the item orientation taxonomy only, since none of the Calculus I final exam items we coded met the criteria for eliciting the cognitive behaviors “evaluate” or “create.” We chose these examples to illustrate some of the more difficult interpretations we made during coding. These descriptions should also help to clarify our taxonomy categories.

Exam items that elicit the cognitive behavior of remembering prompt the student to recall factual information, but do not require the student to apply a procedure or demonstrate understanding. As one example, a test question that is classified as remember prompted students to recall a part of the statement of the Mean Value Theorem (Figure 1).

Which of the following completes the statement of the Mean Value Theorem:

Let f be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b) .

Then there exists some c in (a, b) such that

- a. $f(c) = 0$ b. $f'(c) = 0$ c. $\frac{f(b) - f(a)}{b - a} = f'(c)$ d. $\frac{f'(b) - f'(a)}{b - a} = f(c)$

Figure 1. Sample item requiring the cognitive behavior of remembering.

In this task, students are not required to demonstrate that they understand the Mean Value Theorem, nor are they asked to apply the Mean Value Theorem in a problem context. Simply recalling the statement of the theorem itself is sufficient to respond to the task successfully.

Test items that require students to recall and apply a procedure typically contain verbal or symbolic cues that prompt the use of a definition, rule, theorem, algorithm, or procedure to determine an answer. These items commonly appear in a section of the test where students are directed to differentiate or integrate a function, or evaluate a limit. Such test items require students to apply differentiation rules, integration techniques, or limit properties in an algebraic context but do not require students to understand the rationale for their actions. For the purpose of contrasting the “recall and apply procedure” category with the “remember” category, consider the following problem requiring the use of the Mean Value Theorem:

Show that the function $f(x) = x^2$ satisfies the hypotheses of the Mean Value Theorem on the interval $[0, 4]$ and find a solution c to the equation,

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

on this interval.

Figure 2. Sample item requiring the cognitive behavior of recalling and applying a procedure.

The task (Figure 2) prompts the student to employ a procedure using the conclusion of the Mean Value Theorem without requiring that the student demonstrate an understanding of the Mean Value Theorem. That is, the task does not require the student to recognize that the Mean Value Theorem ensures that, for a function f that is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b) , there exists a point c on (a, b) such that the average rate of change of f over $[a, b]$ is equal to the instantaneous rate of change (or derivative) of f at c . Moreover, the task does not require the student to justify why the statement of the Mean Value Theorem is true; nor, does it require interpretation of the solution's meaning.

We view understanding in a way compatible with Piaget (1968). To Piaget, understanding and assimilating to a scheme were synonymous (Thompson, 2013). Hence, in our view, test items that prompt students to demonstrate they have assimilated a concept into an appropriate scheme are classified as “understand.” Tasks that require understanding may require students to make inferences about the concept by providing examples, making interpretations or comparisons, or providing explanations. For example, the item in Figure 3 requires students to interpret the meaning of the derivative at a point, the average rate of change, and the definite integral in context.

Let $r(x)$ represent the total revenue obtained by Company A from selling x items. Interpret the meaning of the following:

$$\text{a. } r'(4,597) \qquad \text{b. } \frac{r(998) - r(351)}{998 - 351} \qquad \text{c. } \int_{2,485}^{10,219} r(x) dx$$

Figure 3. Sample item requiring the cognitive behavior of understanding.

This task requires students to demonstrate an understanding by asking them to interpret the meaning of various expressions. Part (a) requires students to interpret the symbols that represent

a derivative at a point, part (b) requires students to explain the meaning of average rate of change on a specified interval of a function's domain, and part (c) requires students to interpret integral notation and the meaning of a definite integral. That students must interpret these expressions in the context of a function that defines the relationship between quantities (Thompson, 2011) additionally requires understanding.

We should caution, however, that tasks may elicit understandings we do not consider desirable. For instance, it is likely that a task that asks students to explain the meaning of $\int_a^b f(x)dx$ intends students to say something about the limiting value of the sum of the area of rectangles. Even though this understanding does not reveal a clear connection between differentiation and integration, it is an understanding nonetheless and an item of this type would be coded as such.³ Therefore, coding items relative to the understanding the task elicits is not influenced by our assessment of the utility of that understanding.

The intellectual behavior of applying understanding involves applying the product of an understanding as opposed to applying a rehearsed procedure. If a student is applying concepts or procedures that are brought to the task, the student must then *recall* those concepts or procedures and apply them. Since we are not able to assess the understandings that students bring to the task, we must ask, "Does the task present the student with an *opportunity* to demonstrate understandings, and further, does the task necessitate the application of these understandings in order to successfully complete the task?" If so, the item orientation is "apply understanding."

³ We prefer to speak of definite integration as representing the *accumulation* of some quantity as detailed in Thompson & Silverman (2008). Fleshing out the nature of this understanding, however, is beyond the scope of this paper.

Tasks that require students to apply understanding often require them to transfer knowledge to a less familiar domain in which surface features of the problem statement cannot serve as cues that trigger the enactment of a rehearsed problem-solving procedure. For example, consider the task in Figure 4. This task requires students to first identify relationships between relevant varying and constant quantities for the purpose of defining a function to describe how an angle measure and distance are varying together. Students then need to recognize that determining the derivative of the distance with respect to the angle measure will produce the a new function that describes the rate of change of the distance with respect to the angle measure. Knowing to take the derivative requires that students have an understanding of the usefulness of the derivative in determining the speed of the balloon at a moment in time.

A hot air balloon rising straight up from a level field is tracked by a range finder 500 feet from the lift-off point. At the moment the range finder's elevation angle is $\pi/4$, the angle is increasing at the rate of 0.14 radians per minute. How fast is the balloon rising at that moment? (Finney et al., 2007, p. 247).

Figure 4. Sample items illustrating the contrast between the “recall and apply procedure” and “apply understanding” levels of the item orientation taxonomy.

Test items that elicit the cognitive behavior of *analyzing* require one to determine the relationships between constituent components of a mathematical concept, and then establish how these components contribute to the overall structure or purpose of the concept. For instance, consider the task in Figure 5.

Write a one-page essay explaining why the concept of limit is a central theme of the course. Your essay should describe what it means to understand this concept and explain how the concept is related to other ideas in the course.

Figure 5. Sample item requiring the cognitive behavior of analyzing.

This task (Figure 5) prompts students to explain the meaning of the limit concept and describe why it is a central idea in the study of calculus; thereby requiring students to determine the relationship between the limit concept and other central concepts such as differentiation and integration. We claim that a task of this type requires a strong understanding of limit in addition to more advanced cognitive structures since students must understand and explain how complex ideas are connected.

Item Representation

Characterizing the representation of exam items involved classifying both the representation(s) used in the stated task and the representation(s) that is needed when constructing the solution. Table 2 describes these classifications relative to the task statement and the solution. We also note that both the task statement and the solicited solution can involve multiple representations.

Table 2

Descriptions of item representation categories

Item representation	Task statement	Solicited solution
Applied/modeling	The task presents a physical or contextual situation.	The task requires students to define relationships between quantities. The task may also prompt students to define or use a mathematical model to describe information about a physical or contextual situation.

Symbolic	The task conveys information in the form of symbols.	The task requires the manipulation, interpretation, or representation of symbols.
Tabular	The task provides information in the form of a table.	The task requires students to organize data in a table.
Graphical	The task presents a graph.	The task requires students to generate a graph or illustrate a concept graphically (e.g., draw a tangent line or draw a Riemann sum).
Definition/theorem	The task asks the student to state or interpret a definition or theorem, or presents/cites a definition or theorem.	The task requires a statement of a definition or theorem, or an interpretation of a definition or theorem.
Proof	The task presents a conjecture or proposition.	The task requires students to demonstrate the truth of a conjecture or proposition by reasoning deductively.
Example/ counterexample	The task presents a proposition or statement with the expectation that an example or counterexample is provided.	The task requires students to produce an example or counterexample.
Explanation	Not applicable. This code is	The task requires students to explain

particular to what is expected in the the meaning of a statement.
students' solution.

It is important to note that since many tasks can be solved in a variety of ways and with consideration of multiple representations, we coded for the representation requested in the solution by considering only what the task requires for an answer. For example, a problem that asks students to calculate the slope of a tangent line only requires a student to do symbolic work. Accordingly, we would not code “graphical” as a representation of the solution since reasoning graphically is not necessary to solve the problem, even though the problem has graphical meaning.

Item Format

The third and final strand of the Exam Characterization Framework is item format. The most general distinction of an item's format is whether it is multiple-choice or open-ended. However, open-ended tasks vary in terms of how they are posed. For instance, the statement of an open-ended task may prompt the student to respond to one question that has one correct answer. Such an item is similar to a multiple-choice item without the choices and is therefore classified as *short answer*. In contrast, a *broad open-ended task* elicits various responses, with each response typically supported by some explanation. The form of the solution in a broad open-ended item is not immediately recognizable when reading the task. In addition to coding tasks as short answer or broad open-ended, we also note instances in which a task is presented in the form of a word problem. Also, tasks that require students to explain their reasoning or justify their solution can be supplements of short answer or broad open-ended items. We distinguish between explanations and justifications in that explanations are presented in the form of narrative

descriptions, using words, and justifications are presented mathematically (e.g. requiring further symbolic or computational work to demonstrate the validity of some previous result).

Descriptions of the item format codes are given in Table 3.

Table 3

Descriptions of item format categories

Item format	Description
Multiple Choice	One question is posed and one answer in a list of choices is correct. The student is prompted to select the correct answer among the choices.
Short answer	The item asks the student to respond to one question that has one correct answer. The student can anticipate the form of the solution merely by examining the task—this is similar to a multiple-choice item without the choices.
Broad open-ended	There are multiple ways of expressing the answer. The form of the solution is also not immediately recognizable upon immediate inspection of the task.
Word problem	A word problem is posed in a contextual setting using words, and prompts students to create an algebraic, tabular and/or graphical model to relate specified quantities in the problem, and may also prompt students to make inferences about the quantities in the context using the model. Note that a word problem can be posed as either short answer or broad open-ended or multiple choice. Hence, we code a task as a word problem in addition to identifying

it as either short answer or broad open-ended.

Explain (*E*) and justify (*J*) are subcodes of Item Format.

The categories of the three dimensions of the Exam Characterization Framework are provided in Table 4.

Table 4

Summary of Exam Characterization Framework dimensions

Exam Characterization Dimension	Categories		
Item orientation	Procedural	Remember	
		Recall and apply procedure	
		Understand	
		Apply understanding	
		Analyze	
		Evaluate	
		Conceptual	
		Create	
	Item representation		Applied/modeling
			Symbolic
		Tabular	
		Graphical	
		Definition/Theorem	
		Proof	
		Example/counterexample	
		Explanation	

	Multiple choice		
			Short answer
Item format	Open-ended	Word Problem	
		Broad open-ended	

The first author trained two graduate students in mathematics to use the Exam Characterization Framework. The graduate students then coded ten randomly selected exams (containing 226 items) so that we could assess inter-rater reliability. The first author and the two graduate students achieved a reliability of 89.4% for the item orientation codes, 73.2% for the item representation codes, and 92.7% for the item format codes. We note that the item representation codes were more difficult since both the representation of the task statement and the representation of the intended solution were coded and that either could receive multiple representation codes.

Even though we have evidence that our coding framework is highly reliable if used by knowledgeable coders, we do not advocate that others attempt to use this framework to code Calculus I exams, or exams in another mathematics content area in mathematics (e.g., precalculus) without establishing consistency among coders. We believe the process of familiarizing oneself with the items to be coded and the constructs in our coding framework is necessary, and that in any setting a high level of consistency among coders should be established prior to generating coded results for use in analyses. In our case, we had established consistency between the first two authors of this article for the purpose of refining the constructs and providing clearly stated descriptions of our meaning of each construct; however, we assured consistency in coding since the first author coded all items on all 150 exams in our sample. It

was also important that the first author had a clear understanding of the intended meaning of all constructs in our framework and repeatedly coded items that were particularly challenging to code.

Limitations

There are several limitations to our framework. The first is a limitation shared by some of the existing frameworks in the literature. Identifying the cognitive demand that exam items elicit in students is problematic since we have no knowledge of the instruction and experiences of the students taking any particular exam. As a result, a seemingly rote task has the potential to be highly novel for students who have not been exposed to the procedure or method for solving the task. Conversely, tasks that may appear to require an understanding of a particular concept are amenable to being proceduralized if the solution method is practiced repetitiously. We spoke earlier of our intention to code for the minimal reasoning practices and solution representations that a task requires. To do this with a reasonable degree of validity suggests that we assess “minimal reasoning practices and solution representations” relative to our model of the epistemic collegiate calculus student⁴ and are informed by our knowledge of those concepts in the calculus curriculum that are generally proceduralized (e.g., differentiation, integration, limit evaluation). The first two authors have taught the entire calculus sequence multiple times and have researched student learning in Calculus I. These common experiences may have resulted in the authors reaching a common understanding of what constituted viable models of the epistemic calculus student and the general calculus curriculum.

⁴ “Epistemic collegiate calculus student” is used in the Piagetian sense and is characterized by our idealized abstraction of the college calculus student.

A second limitation is that we do not have data indicating how the exams were graded. Thus, our results are presented as percentages of exam items within respective categories of the Exam Characterization Framework. This method of computing results does not account for the reality that particular exam items may represent a higher portion of the exam grade, and thus be valued by the instructor to a greater extent.

A third limitation is that we coded each part of an item as an individual item. For instance, a task that has parts (a), (b), and (c) would be coded as three distinct items. This is a limitation in the sense that some instructors have a tendency to sub-divide tasks while others do not, and our Exam Characterization Framework does not identify the number of mathematical procedures or understandings needed to complete a task. We believe, however, that our large sample size mediates this concern.

Results

Data analysis in this study consisted of three phases. In the first phase, we developed the Exam Characterization Framework discussed above and used it to code the 150 randomly selected exams in our sample. In the second phase we compared the coded exam data with data obtained from a post-term instructor survey with the intention of determining the extent to which our characterization of the exams corresponds with instructors' perceptions of their exams relative to their conceptual orientation. In the third phase, we coded 13 Calculus I final exams administered in U.S. colleges and universities in 1986/87 and compared these results to those of our random sample of 150 exams.

Characteristics of Post-Secondary Calculus I Exams

The results from coding the 150 Calculus I final exams using the item orientation taxonomy revealed that only 14.72% of the 3,735 exam items required students to demonstrate

an understanding of an idea or procedure. More specifically, 6.51% of the items required students to employ the cognitive behavior of remembering, 78.70% of the items required students to recall and apply a rehearsed procedure, 4.42% required students to demonstrate an understanding, and 10.30% of the exam items required students to apply their understanding while solving a problem. These results indicate that the vast majority of exam items (85.21%) could be solved by simply retrieving rote knowledge from memory, or recalling and applying a procedure, requiring no understanding of an idea or why a procedure is valid.

Of the 150 exams coded, 90% of them had at least 70% of the items coded as “remember” or “recall and apply procedure” levels of the item orientation taxonomy. Additionally, only 2.67% of exams had 40% or more of the items requiring students to demonstrate or apply understanding. The coding results for the item orientation taxonomy are given in Table 5.

Table 5

Percentage of items within each item orientation category

Item Orientation	%
Remember	6.51
Recall and Apply Procedure	78.70
Understand	4.42
Apply Understanding	10.30
Analyze	0.11
Evaluate	0
Create	0

The box plots (Figure 6) provide a summary of the percentage of exam items per exam within each category of the item orientation taxonomy. Consistent with the aggregate data in Table 5, we observe that the highest percentage of exam items per exam required the cognitive behavior of recalling and applying a procedure with consistent distributions for the “remember”, “understand”, and “apply understanding” categories of the taxonomy. We also observe that the minimum value for the distribution of the percentage of exam items requiring students to recall and apply a procedure exceeds the maxima of all other distributions in the plot—representing that any single exam contained more items requiring students to recall and apply a procedure than demonstrate any other cognitive behavior.

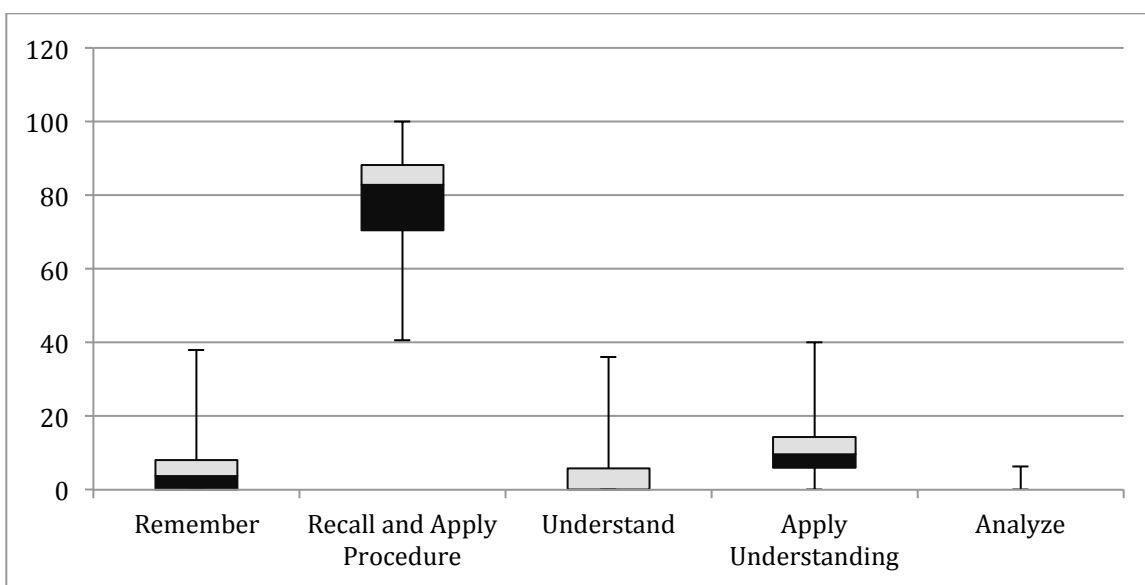


Figure 6. Box plots representing the percentage of items per exam within each category of the item orientation taxonomy.

We calculated the percentage of items within each category of the item orientation taxonomy for each post-secondary institution type and the results are provided in Table 6. The highest percentage of items requiring the cognitive behavior of recalling and applying procedures were those from community colleges (84.4%) while the lowest were from national liberal arts

colleges (69.6%)—those that emphasize undergraduate education and award at least half of their degrees in the liberal arts fields of study. It is also noteworthy that national universities contained the highest percentage of items requiring students to apply their understanding (10.8%) while the lowest were from regional colleges (3.8%)—those that focus on undergraduate education, awarding less than half of their degrees in liberal arts fields of study. Moreover, national liberal arts colleges contained the highest percentage of items at the “understand” level of the item orientation taxonomy or higher (21%) while regional colleges contained the highest percentage of items at the “remember” or “recall and apply procedure” levels (96.26%).

Table 6

Percentage of items within each item orientation by post-secondary institution classification

Item Orientation	National				
	National University	Regional University	Community College	Liberal Arts College	Regional College
Remember	5.53%	9.33%	4.25%	8.77%	13.46%
Recall and Apply					
Procedure	78.9%	77.6%	84.4%	69.6%	78.8%
Understand	4.6%	3.0%	3.7%	10.5%	3.8%
Apply Understanding	10.8%	10.0%	7.6%	10.5%	3.8%
Analyze	0.1%	0.1%	0.0%	0.0%	0.0%
Evaluate	0.0%	0.0%	0.0%	0.0%	0.0%
Create	0.0%	0.0%	0.0%	0.0%	0.0%

In terms of item representation, exam items were predominantly stated symbolically (73.70%) or required a symbolic solution (89.4%) while few items prompted students for information in the form of a table (1.02%), presented a proposition or statement with the expectation that an example or counterexample be provided (0.59%), or presented a conjecture or proposition with the expectation that a proof be provided (1.29%). The percentages of exam items within each of the item representation categories are provided in Table 7.

Table 7

Percentage of items within item representation categories

Item Representation		Item Representation	
(Task)	%	(Solution)	%
Applied/Modeling	13.20	Applied/Modeling	6.96
Symbolic	73.70	Symbolic	89.40
Tabular	1.02	Tabular	0.19
Graphical	10.40	Graphical	5.70
Definition/Theorem	3.51	Definition/Theorem	4.36
Proof	1.29	Proof	1.53
Example/Counterexample	0.59	Example/Counterexample	0.59
Explanation	2.36	Explanation	2.36

The vast majority of Calculus I final exam items in our sample were stated symbolically and required a symbolic solution (65.45%). Moreover, only 6.94% of the items presented a physical or contextual situation and solicited a solution in which students were required to define relationships between quantities or use a mathematical model to describe information about the

physical or contextual situation. Taking into consideration the increased emphasis in calculus on making connections among multiple representations (i.e., words, symbols, tables, and graphs), it is noteworthy that only 9.59% of exam items were stated exclusively as “applied/modeling,” “symbolic,” “tabular,” or “graphical” and required the solution to be represented differently. Table 8 outlines the most prevalent combinations of item representations in our sample.⁵

Table 8

Percentage of exam items within item representation categories

Item representation (task)	Item representation (solution)	%
Symbolic	Symbolic	65.45
Applied/modeling	Applied/modeling; Symbolic	6.86
Graphical	Symbolic	6.72
Applied/modeling	Symbolic	4.90
Symbolic	Graphical	2.01
Definition	Definition	1.90

Results also indicate that introductory calculus final exams seldom include tasks that are stated in the context of a real-world situation. Our coding revealed that 38.67% of the coded exams had less than 5% of the items classified as word problems, in either the “short answer” or “broad open-ended” format categories. Further, only 22% of the coded exams had more than 10% of the exam’s test items classified as a word problem in the “short answer” or “broad open-

⁵ By “most prevalent” we refer those item representations that accounted for more than 1% of items in our sample.

ended” format categories. It is also noteworthy that 18% of the exams contained no word problems. The coding results from the item format strand are given in Table 9.

Table 9

Percentage of items within each item format category

Item Format	%
Multiple Choice	11.70
Multiple Choice (Explain)	0.59
Multiple Choice (Justify)	0.19
Multiple Choice (Word Problem)	0.40
Short Answer	76.10
Short Answer (Explain)	2.38
Short Answer (Justify)	1.04
Short Answer (Word Problem)	6.05
Broad Open-Ended	1.23
Broad Open-Ended (Explain)	0.08
Broad Open-Ended (Justify)	0
Broad Open-Ended (Word Problem)	0.03

In summary, results from coding the 150 randomly selected exams with the Exam Characterization Framework indicate that the Calculus I final exams in this study require low levels of cognitive demand, contain few problems stated in a real-world context, rarely elicit explanation, and do not require students to demonstrate or apply their understanding of the course’s central ideas.

Contrast Between Instructors' Beliefs and Coding Results

To address our second research question, we coordinated the results of the exam codes and instructors' responses to a post-term instructor survey. We devoted particular attention to identifying inconsistencies between our findings about the characteristics of the Calculus I final exams and instructors' perceptions of their exams.

Figure 7 provides the distribution of instructors' survey responses to the survey question, "How frequently did you require students to explain their thinking on exams?" Responses ranged from 1 (not at all) to 6 (very often). We note that an exam item requiring explanation would very likely have been coded as either "understand" or "apply understanding" in the item orientation taxonomy. Results from the item format codes indicate that a total of 3.05% of all items coded ($N = 3,735$) required an explanation. Further, only 14.72% of all exam items were coded as "understand" or "apply understanding" in the item orientation taxonomy. However, 68.18% of all instructors who submitted the exams that were a part of our sample selected either 4, 5, or 6 on this survey item, indicating that these instructors claim to frequently require that their student explain their thinking on exams. These data reveal that these instructors' Calculus I final exams do not align with their perceptions of their exams relative to the extent to which students are required to explain their thinking.

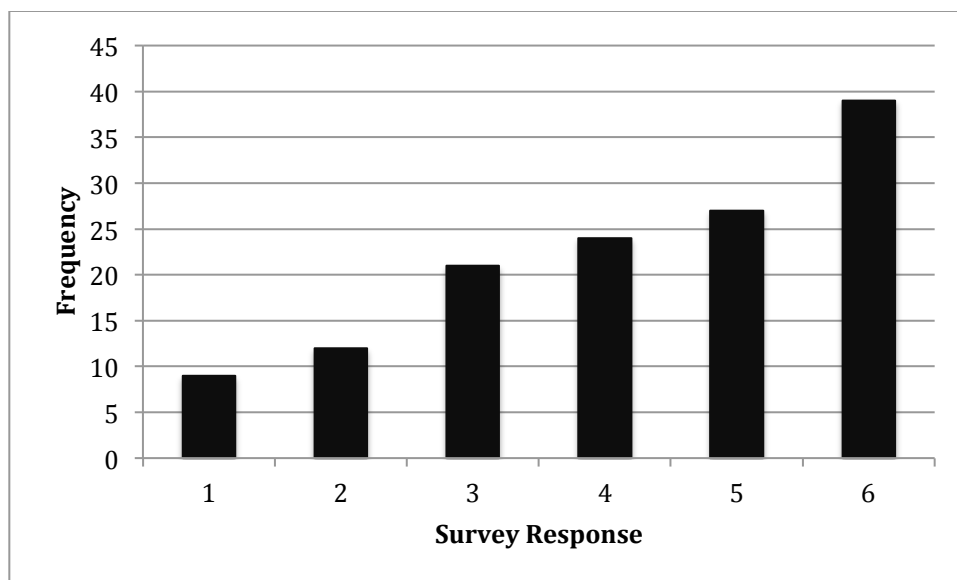


Figure 7. Instructors' response to the question, "How frequently did you require students to explain their thinking on exams?"

Similarly, there was also discrepancy between our characterization of the final exams and survey responses with respect to the percentage of exam items that emphasized skills and methods for executing computations. Figure 8 indicates the distribution of instructors' responses to the survey question, "On a typical exam, what percentage of the points focused on skills and methods for carrying out computations?" The median response was 50%. Our coding results, however, indicate that 78.7% of exam items require students to recall and apply a procedure. Additionally, 89.4% of all exam items required students to perform symbolic computation.

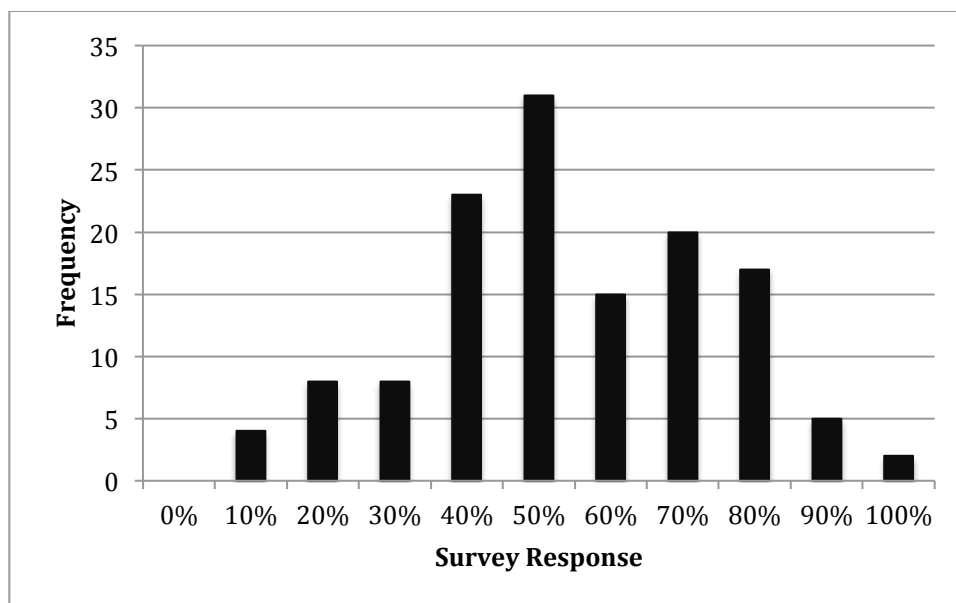


Figure 8. Instructors’ response to the question, “On a typical exam, what percentage of the points focused on skills and methods for carrying our computations?”

These data reveal that there is a misalignment between our characterization of Calculus I final exams and instructors’ perceptions of their exams relative to the extent to which students are asked to explain their thinking and the percentage of exam items that focus on skills and methods for carrying out computations.

Correlating Item Orientation with Representation and Format

In order to determine if particular item representations or formats necessitated higher-order cognitive activity, we calculated the percentage of item representations and item format types within each item orientation category. Table 10 documents the percentage of the most common item representation types (used in the task and asked for in the solution) for the first four levels of the item orientation taxonomy.⁶

⁶ By “most common,” we refer to an exclusion of item representations that represented less than 5% of the items within a specific item orientation category.

Table 10

Percentage of item representations within each category of the item orientation taxonomy

Item orientation	Item representation (task)	Item representation (solution)	%
Remember	Symbolic	Symbolic	30.20
	Definition	Definition	27.76
	Graphical	Symbolic	21.63
Recall and apply procedure	Symbolic	Symbolic	78.98
	Applied/modeling	Symbolic	5.20
	Graphical	Symbolic	5.10
Understand	Graphical	Symbolic	20.00
	Symbolic	Explanation	13.33
	Symbolic	Symbolic	10.30
	Definition	Explanation	7.27
	Applied/modeling	Explanation	10.30
Apply understanding	Applied/modeling	Applied/modeling; Symbolic	65.18
	Applied/modeling	Symbolic	5.76
	Symbolic	Symbolic	5.50
	Symbolic	Definition	5.45

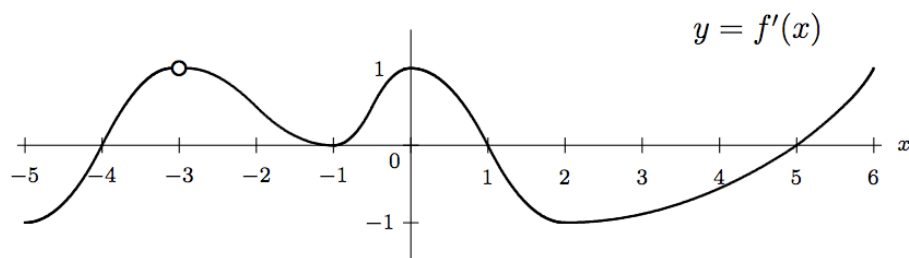
Items that were stated symbolically and required a symbolic solution were most prevalent among items that required students to remember (30.20%). The vast majority of tasks that

required students to recall and apply a procedure (78.98%) were stated symbolically and required a symbolic solution. We also observed that the percentage of items on an exam that were stated symbolically and solicited a symbolic solution decreased as tasks demanded higher levels of cognitive behavior (10.30% in the “understand” category, 5.50% in the “apply understanding” category, and 0% in the “analyze” category⁷).

It is also noteworthy that items presented in a contextual or physical situation (i.e., applied/modeling) were most prevalent among items that required students to apply understanding (70.94%). The applied/modeling items also represented a significant percentage of items that required students to demonstrate understanding (10.30%). Moreover, 20% of items that required students to demonstrate understanding presented information in the form of a graph and solicited symbolic work in the solution—although these items often required explanation or justification. Thus, in the majority of items that required students to demonstrate understanding or apply understanding, students were required to interpret information from a graphical representation or applied setting. As an example, consider the task in Figure 9 stated graphically.

The following graph represents the graph of the derivative f' of a function f that is defined on the interval $[-5, 6]$. Using this graph, answer the following questions about the function f and the function $\int f(x)dx$.

⁷ The percentages of item representations within the “analyze” category of the item orientation taxonomy are not provided in Table 11 as a result of the small number of items within this category ($N = 4$).



Note: the empty circle represents a hole in the graph.

- On what intervals is f decreasing?
- On what intervals is $\int f(x)dx$ increasing?
- On what intervals is f concave up?
- On what are the critical points of $\int f(x)dx$?
- For the critical points identified in part (d), determine whether it is a local maximum, local minimum, or neither and explain your response.

Figure 9. Item stated graphically that elicits the cognitive behavior or understanding.

All parts of the task in Figure 9 require students to demonstrate understanding since they are asked to infer the behavior of a function f and a function $\int f(x)dx$ by examining only the graphical behavior of the function f' ; thereby requiring students to attend to the meaning of the derivative function and what the indefinite integral represents.

Table 11 indicates the percentage of various item formats within the first four categories of the item orientation taxonomy.⁸ We find it notable that short answer items requiring students to provide an explanation assumed a much higher percentage of items eliciting the cognitive

⁸ We include only those item formats representing more than 1% of the items within the first four item orientation categories.

behavior of understanding (15.15%) than items requiring the recollection and application of a procedure (1.16%) or items simply prompting students to remember (0.41%). Moreover, 97.93% of word problems—whether they be stated in the form of multiple choice, short answer, or broad open-ended—provided students with the opportunity to apply their understanding. Finally, we observe that 6.50% of the total number of items prompting students to provide an explanation or justification required only the cognitive behavior of remembering whereas 27.64% required students to recall and apply a procedure, 31.71% required students to demonstrate understanding, and 34.15% provided students with the opportunity to apply their understanding.

These data suggest that items presented in the form of word problem or prompting students to provide an explanation or justification for their work elicit higher-order cognitive behavior.

Table 11

Percentages of item formats within each category of the item orientation taxonomy

Item orientation	Item Format	%
Remember	Short Answer	62.04
	Multiple Choice	32.65
	Multiple Choice (Explain)	3.27
Recall and apply procedure	Short Answer	86.87
	Multiple Choice	10.58
	Short Answer (Explain)	1.16
Understand	Short Answer	36.97
	Multiple Choice	24.85

	Short Answer (Explain)	15.15
	Broad Open-Ended	13.94
	Multiple Choice (Explain)	5.45
	Multiple Choice (Justify)	1.82
	Broad Open-Ended (Explain)	1.21
Apply	Short Answer (Word Problem)	
understanding		56.28
	Short Answer	22.25
	Short Answer (Explain)	7.59
	Multiple Choice (Word Problem)	3.93
	Broad Open-Ended	3.66
	Multiple Choice	2.09
	Short Answer (Justify)	1.83
	Short Answer (Word Problem/Justify)	1.57

Comparing the Cognitive Demand of Our Sample with Final Exams from 1986/87

The results presented thus far reveal that our sample of post-secondary Calculus I final exams require low levels of cognitive engagement, rarely make use of real-world contexts, seldom elicit explanations or justifications, and do not provide students with opportunities to demonstrate or apply their understanding. These results are particularly surprising considering the considerable amount of attention devoted to the conceptual teaching of calculus within the past 25 years. This led us to consider how contemporary Calculus I final exams compare with those administered prior to the calculus reform movement. We were able to achieve this

comparison by coding Calculus I final exams administered in 1986/87 to students enrolled in first semester calculus from a variety of post-secondary institutions across the United States, from large doctoral granting universities to small two-year colleges (Steen, 1988). The 13 Calculus I final exams (with 354 total items) we coded were included in the appendix of Steen (1988). The results from coding this sample of 1986/87 exams relative to item orientation are provided in Table 12.

Table 12

Percentage of 1986/87 exam items within each item orientation category

Item Orientation	%
Remember	3.67
Recall and Apply Procedure	87.01
Understand	1.13
Apply Understanding	8.19
Analyze	0
Evaluate	0
Create	0

Our analysis revealed that there is a statistically significant difference between the proportion of items requiring students to recall and apply a procedure in our sample of 150 calculus exams and the 13 Calculus I final exams from 1986/87. The difference between these two data sets in the proportion of items within the “apply understanding” category of the item orientation taxonomy, however, is not statistically significant. While there is significantly less emphasis on recalling and applying procedures in contemporary Calculus I exams, compared to

those administered 25 years ago, our results suggest that the proportion of exam items eliciting the cognitive behavior of applying understanding has not significantly changed. This finding is surprising considering the tremendous effort devoted to the conceptual teaching of calculus since the initiation of the calculus reform movement in the late 1980s.

Calculus reform initiatives have emphasized the importance of developing students' reasoning with multiple representations. One would expect, then, that more recent Calculus I exam items would require students to interpret information presented in a particular representation (e.g., applied/modeling) and translate it to another (e.g., graphical). Table 13 provides the percent of exam items from the Calculus I final exams from 1986/87 within various item representation categories.

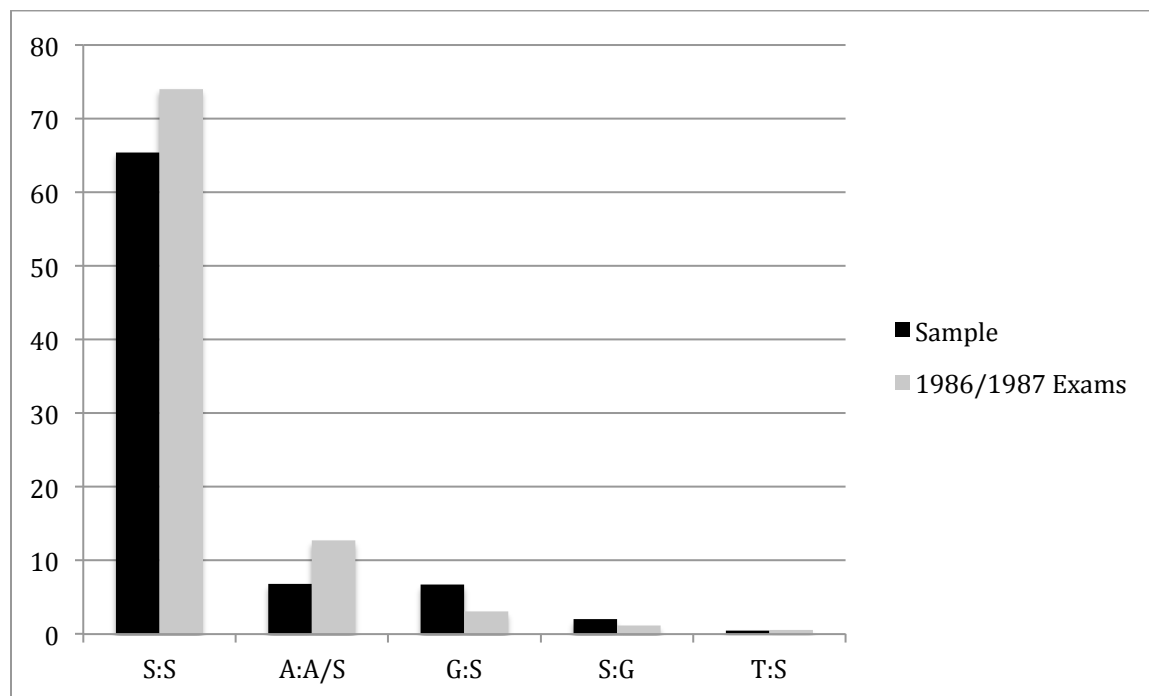
Table 13

Percent of 1986/87 exam items within each item representation category

Item representation (task)	Item representation (solution)	%
Symbolic	Symbolic	74.01
Applied/modeling	Applied/modeling; Symbolic	12.71
Symbolic	Symbolic; Graphical	4.80
Graphical	Symbolic	3.11
Definition	Definition	1.69
Symbolic	Graphical	1.13
Symbolic; Definition	Symbolic	0.56
Tabular	Symbolic	0.56
Symbolic	Applied/modeling; Symbolic	0.28
Symbolic; Tabular;	Symbolic	0.28

Graphical		
Graphical	Graphical	0.28
Symbolic	Symbolic; Tabular; Graphical	0.28
Symbolic	Symbolic; Definition	0.28

We contrast select percentages from Table 13 with those from our sample of 150 exams in Table 8. These data reveal that a higher percentage of items on the 1986/87 final exams were stated symbolically and required students to interpret, represent, or manipulate symbols in the solution, than items in our initial sample. However, a higher percentage of items on the exams from 1986/87 presented a physical or contextual situation and required students to define the relationship between quantities. We illustrate a comparison of item representations among exams from 1986/87 and our original sample in Figure 10.



S:S represents symbolic task representation and symbolic solution representation.

A:A/S represents applied/modeling task representation and applied/modeling and symbolic solution representation.

G:S represents graphical task representation and symbolic solution representation.

S:G represents symbolic task representation and graphical solution representation.

T:S represents tabular task representation and symbolic solution representation.

Figure 10. Comparison of item representations among exams from 1986/87 and our original sample.

Discussion

A review of student learning of ideas in introductory calculus (e.g., Carlson & Rasmussen, 2008) revealed that Calculus I students are generally not developing conceptual understanding of the central ideas of calculus. At the same time, the calculus reform movement has resulted in shifts in calculus curriculum to provide increased focus on student understanding of the course's key ideas (Ganter, 2001). With this increase in conceptual focus in the calculus curriculum, one might expect that calculus exams would include more items that assess students' understanding and ability to use the key concepts of calculus. Our analysis suggests this is indeed not the case.

Our examination of 150 contemporary Calculus I final exams revealed that these exams are primarily assessing students' ability to recall and apply procedures. There is little focus on assessing student understanding, with 85.21% of the 3,735 exam items being solvable by simply retrieving rote knowledge from memory, or recalling and applying a procedure. The Calculus I final exams rarely make use of real-world contexts, seldom elicit explanation or justification, and provide few opportunities for students to demonstrate or apply their understanding. Of the 150 exams we coded, 90% had 70% or more of the exams' items coded as "remember" or "recall and

apply procedure.” In contrast, only 2.67% of the 150 exams had 40% or more of the items requiring students to demonstrate or apply understanding.

As a result of the low percentage of exam items at the “understand” or “apply understanding” levels of the item orientation taxonomy, we conclude that a large percentage of exam items failed to provide insight into how students understand the concepts on which their computational or procedural work is based. Hence, these results suggest that a large majority of Calculus I final exams being administered in colleges and universities in the United States encourage memorization of procedures for answering specific problem types and do not encourage students to understand or apply concepts of beginning calculus.

Our data further revealed that Calculus I final exams in U.S. colleges and universities have changed very little in the past 25 years relative to the percentage of exam items that require students to apply understanding of concept. Since final course exams commonly reflect the level of mastery and understanding students have attained at the end of the course, these data suggest that the calculus reform movement of the late 1980s has had little effect on what is being assessed in contemporary Calculus I courses in U.S. postsecondary institutions.

We also found that there exists inconsistency between our characterization of postsecondary Calculus I final exams and instructors’ perception of the nature of the exams they administer to their students. Instructors’ report that their exams require students to demonstrate and apply understandings, while our analyses revealed that a very high percentage of exam items focus on skills and methods for carrying out computations and a very low percentage of items prompt students to explain their thinking.

If exams are predominantly based on measures of low level cognitive behaviors such as “remember”, or “recall and apply procedure”, students are likely to engender perceptions about

mathematics as not being about understanding and using ideas to solve novel problems. This focus on procedures has been reported to be uninteresting to some more capable students who enjoy understanding and reasoning through novel problems (Thompson et al., 2007).

We encourage Calculus I instructors and mathematics departments to contemplate the role of their Calculus I final exams in developing students' mathematical abilities, and to consider whether their exams are supporting the development of students' mathematical understandings and thinking. In many cases this may require departments to examine their Calculus I curriculum and instruction. Some may also find it useful to consult literature (e.g., Carlson & Rasmussen, 2008) that discusses what is involved in understanding calculus ideas, and what studies have revealed about the process of understanding and learning to use these ideas to solve novel problems. We also encourage the use of our Exam Characterization Framework in pre-service teaching programs and graduate teaching assistant workshops as a didactic tool to assist teachers in constructing assessments that afford students the opportunity to demonstrate their understanding.

Acknowledgement: This research was supported in part by National Science Foundation Grant DRL-0910240. Any recommendations or conclusions stated here are those of the authors and do not necessarily reflect official positions of the NSF.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

- Bergqvist, E. (2007). Types of reasoning required in university exams in mathematics. *Journal of Mathematical Behavior*, 26, 348-370.
- Bloom, B. S., et al. (Eds.). (1956). *Taxonomy of educational objectives: Part I, cognitive domain*. New York: Longman Green.
- Boesen, J., Lithner, J., & Palm, T. (2006). *The relation between test task requirements and the reasoning used by students*. Umeå, Sweden: Department of Mathematics, Umeå University.
- Carlson, M. P., & Rasmussen, C. (Eds.). (2008). *Making the connection: Research and teaching in undergraduate mathematics education*. Washington, DC: Mathematical Association of America.
- Charalambous, C., Delaney, S., Hsu, A., & Mesa, V. (2010). The addition and subtraction of fractions in the textbooks of three countries: A comparative analysis. *Mathematical Thinking and Learning*, 12(2), 117-151.
- Finney, R. L., Demana, F. D., Waits, B. K., & Kennedy, D. (2007). *Calculus: Graphical, numerical, algebraic* (3rd ed.). Boston, MA: Pearson Prentice Hall.
- Ganter, S. L. (Ed.). (2001). *Changing calculus: A report on evaluation efforts and national impact from 1988-1998*. Washington DC: Mathematical Association of America.
- Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *The Journal of Educational Research*, 91(1), 26-32.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), p. 212-218.

- Li, Y. (2000). A comparison of problems that follow selected content presentations in American and Chinese mathematics textbooks. *Journal for Research in Mathematics Education*, 31(2), 234-241.
- Lithner, J. (2000). Mathematical reasoning in task solving. *Educational Studies in Mathematics*, 41(2), 165-190.
- Lithner, J. (2003). Students' mathematical reasoning in university textbook exercises. *Educational Studies in Mathematics*, 52(1), 29-55.
- Lithner, J. (2004). Mathematical reasoning in calculus textbook exercises. *The Journal of Mathematical Behavior*, 23, 405-427.
- Mesa, V., Suh, H., Blake, T., & Whittemore, T. (2012). Examples in college algebra textbooks: Opportunities for students' learning. *PRIMUS*, 23(1), 76-105.
- Mesa, V. (2010). Strategies for controlling the work in mathematics textbooks for introductory calculus. In F. Hitt, D. Holton & P. W. Thompson (Eds.), *CBMS Issues in Mathematics Education* (Vol. 16, pp. 235-261). Providence, RI: AMS.
- Palm, T., Boesen, J., & Lithner, J. (2006). *The requirements of mathematical reasoning in upper secondary level assessments*. Umeå, Sweden: Department of Mathematics, Umeå University.
- Piaget, J. (1968). *Six Psychological Studies*. New York: Vintage Books.
- Senk, S. L., Beckman, C. E., & Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education*, 28(2), 187-215.
- Smith, G., Wood, L., Coupland, M., Stephenson, B., Crawford, K., & Ball, G., (1996). Constructing mathematical examinations to assess a range of knowledge and skills.

International Journal of Mathematics Education in Science and Technology, 27(1), 65-77.

Steen, L. (Ed.) (1988). *Calculus for a new century: A pump, not a filter*. Washington, DC: The Mathematical Association of America.

Thompson, P. W. (2011). Quantitative reasoning and mathematical modeling. In S. Chamberlin, L. L. Hatfield, & S. Belbase (Eds.), *New perspectives and directions for collaborative research in mathematics education: Papers from a planning conference for WISDOMÉ*. Laramie, WY: University of Wyoming.

Thompson, P. W. (2013). In the absence of meaning. In K. Leatham (Ed.), *Vital directions for research in mathematics education*, p. 57-93. New York: Springer.

Thompson, P. W., & Silverman, J. (2008). The concept of accumulation in calculus. In M. Carlson & C. Rasmussen (Eds.), *Making the connection: Research and teaching in undergraduate mathematics* (pp. 117-131). Washington, DC: Mathematical Association of America.

Thompson, P. W., Castillo-Chavez, C., Culbertson, R. J., Flores, A., Greely, R., Haag, S., et al. (2007). *Failing the future: Problems of persistence and retention in science, technology, engineering, and mathematics majors at Arizona State University*. Tempe, AZ: Office of the Provost.