Adapting an Exam Classification Framework Beyond Calculus

Brian P Katz
Augustana College (IL)

Sandra Laursen
University of Colorado Boulder

This paper reports a methods-building project that seeks to make inferences about mathematics instructors' teaching practices from their exams. We adapt and revise a framework by Tallman et al. (2016) and expand its applicability across the undergraduate curriculum, beginning with a sample of seven exams from early-career mathematics instructors. We describe the rationale for the adaptation process and patterns of differences between exam sets. Future work includes coordinating this analysis with results from other data sets from the same instructors.

Keywords: Exams, Teaching Practice, Methods-building

This paper is part of a larger research project that seeks to detect changes in the teaching of individual instructors. To accomplish that, we are working to develop and coordinate methods that capture aspects of instructors' teaching from diverse data sets including syllabi, classroom video, instructor and student surveys, and classroom artifacts such as assessments. We expect that these data sources capture different aspects of instructors' teaching, and they vary in how invasive or expensive they are to collect and analyze. Rather than hoping to argue that one of these data sources and methods is best, we seek to understand the affordances of each method and the kinds of questions that are appropriate for each for the purposes of detecting change through professional development. Additionally, we hope to help clarify the desired outcomes of professional development, the change trajectories of participants, and the kinds of evidence that demonstrate that change is occurring. The results of this project could be used to help assess and improve professional development for mathematics instructors, which may in turn support the shift towards active and inquiry-based pedagogies advocated by professional organizations for collegiate mathematics instructors (CBMS, 2016).

This paper focuses on course exams to ask the following question: What can an instructor's exams tell us about that person's teaching? Exams are generally high-stakes assessments, so presumably they represent the values, beliefs, practices, and theory-in-use (Argyris, 1976) of the instructor who authored them (Black & Wiliam, 1998). However, timed exams also put constraints on the kinds of activities that are possible for students, so we do not expect exams to capture all of an instructor's perspective in general. Exams are easy to collect; they are also authentic artifacts in the sense that they are created as part of the course, for the students. Thus, we articulate the detailed research question:

- Can we build or adapt a coding scheme that detects patterns and differences among instructors' exams in order to support inferences about their teaching?

This phase of the project is methods-building, so this paper emphasizes the development of a scheme for coding exam items from undergraduate mathematics courses across the curriculum. We include some results from coding of a small study sample as evidence that the resulting scheme captures patterns and differences among instructors' exams that in turn may offer evidence about their instructional choices.

Later stages of this program could develop a theoretical perspective on the aspects of teaching and their hypothesized relationship to professional development, but we are not yet at the stage where we can articulate such a framework. Instead, we seek to develop methods that focus our attention on aspects of instructors' teaching, seek patterns and connections within or

across these data and methods, and use both our theoretical sensitivity as researchers and our experience as teachers and professional developers to identify potentially meaningful observations.

Literature Review

We are building a scheme for coding exams to learn about instructors' perspectives on teaching, so we focus on the requirements that they make of students in exam items. Subsequent to this paper, we will coordinate this scheme with analyses of other aspects of these instructors' practice, so our scheme must be independent of specific knowledge of other elements of the course or the students' backgrounds, though it can depend on a coder's more generic knowledge of undergraduate mathematics.

We draw on work of Tallman et al. (2016) to summarize some prior research that has examined individual mathematical tasks. Li (2000) built a three-dimensional coding scheme to assess whether the item required one or more mathematical procedures, whether the item was purely abstract or set in an illustrative context, and what format and cognitive demand were required for a response. It is difficult to determine the grain size of a single procedure without information about the specific course context, but the other dimensions of this scheme align with our goals. Lithner (2004) focused on the potential student strategies for seeking a solution to an exam problem; our project focuses on what is expected of all students in common rather than on potential differences. Smith et al. (1996) and Anderson and Krathwohl (2001) produced coding frameworks that modify and update Bloom's taxonomy. Bloom's taxonomy has been critiqued because the actual cognitive demand of any task depends on the individual student's prior experience, but we accept that an instructor can have a well-defined intended cognitive demand for a task, and that a coder with mathematical expertise could assess this intent from the exam. Mesa et al. (2012) used Charalambous et al. (2010) to incorporate information about representations and metacognition in their coding framework. These dimensions align with our goals, but we focus on how they are required by the instructor rather than on possible student understandings and approaches they support.

As part of a project to determine characteristics of successful programs in post-secondary calculus, Tallman et al. (2016) developed a scheme for coding individual items on Calculus I final exams, called the Exam Characterization Framework (ECF). The ECF has three dimensions: *item orientation*, which captures the cognitive demand required to respond successfully to the item; *item representation*, which captures representations and other objects in both the task and required response to the item; and *item format*, which captures the structure and scope of the expected response to the task. Consistent with their critique of prior work, we observe that the ECF aligns with our own approach except for its exclusive focus on calculus. They applied the ECF to a large corpus of exams from 2010/11 to develop a summary of the expectations of calculus courses, and they contrasted these exams with a sample from 1986/87 to describe the impact of 25 years of reform efforts. Based on this work, we determined to start our scheme-building process by trying to adapt or generalize the ECF to a broader context.

Methods

Our data set is the exams (or mastery quizzes) from seven instructors who had completed a professional development program for early-career mathematics instructors. We expect this population to exhibit a range of teaching behaviors, styles, and skill levels. This sample is small because we collected multiple other kinds of data, including classroom video, from the same instructors (not discussed here). These seven instructors are teaching abstract algebra, discrete

mathematics or introduction to proofs, content courses for future elementary teachers, introductory statistics, or calculus II/III. The first author, who is the main coder, has the credentials to teach all of these courses and has experience teaching courses similar to 6 of them. The data set includes 208 items from 13 distinct assessments from these seven courses.

The first author familiarized himself thoroughly with the ECF as described in Tallman et al. (2016) and then attempted to use his understanding of this framework to code the seven sets of exams, along the way adapting and revising the ECF into a new but related scheme. The goal was to develop a coding scheme that was applicable across undergraduate mathematics courses, that captured all aspects of exam items that seemed to speak to larger patterns in the instructor's teaching, and that was articulated in an internally coherent way that supported reliable coding and distinction between codes. As he coded, the first author noted items for which his current understanding of the framework was not sufficient to assign definitive codes; he also noted aspects of items that were not captured by the codes. He later repeated this process and then compared the codes and comments as an indicator of intra-coder reliability. He then revised his interpretation of existing codes and defined new codes based on repeated comments; these revisions required overt articulations or re-articulations of the hierarchical structure of the codes in each dimension. Finally, he repeated this process of coding and revision until the framework and its interpretation stabilized.

For two of these cycles and for the stable framework, the first author presented examples of coded items, rationales for changing the framework, and descriptions of the hierarchical structure of each dimension of the scheme to the second author as sense-making checks; these checks were an initial effort to establish face validity in our study context. These discussions emphasized consistency in interpreting individual code definitions and coherence and discrimination across the framework components.

The Item Characterization Framework

The resulting framework, which we call the Item Characterization Framework (ICF), contains three broad dimensions: item orientation, item format, and item components. These dimensions are analogous to those in the ECF, but include new categories and codes (Table 1).

Table 1: Dimensions, Categories, and Codes in the Item Characterization Framework

Item Orientation			**Item Format**				**Item Component**
Cognitive Demand	Familiarity	Certainty	Breadth	Format	Formality	Other Support	Task/Response
Remember	Recreate	Low	Single	Multiple choice/TF	No support	Neither	Applied/Modeling context
Recall and apply Procedure	Adapt	Medium	Forked	Fill in the blank	Informal support	Interpretation/ Context	Symbolic representation
Recall and reproduce argument	New	High	Delineated	Short answer	Formal support	Control/ Evaluation	Verbal representation
Understand			Open	Long answer	Unclear	Both	Graphical representation
Apply understanding							Tabular representation
Analyze							Statement (Thm/Dfn)
Evaluate							Claim (Conj/Arg)
Create							(Counter-) Example

Item Orientation

This dimension captures the assumed cognitive demand of producing a successful response to the item. The category *Cognitive Demand* uses an expanded version of Bloom's Taxonomy. *Recall and reproduce argument* is the only novel code here; this code is analogous to *Recall and apply procedure* but applicable to the context of proof construction.

The cognitive demand of a task depends heavily on the student's past experience with the task (Anderson & Krathwohl, 2001). Both ECF and ICF assume that the coder holds an understanding of the generic undergraduate curriculum and student; in ICF, these assumptions are made explicit by coding how novel the coder believes the task to be for the intended student in *Familiarity* and their confidence in this assessment (and thus of *Cognitive Demand*), in *Certainty*.

Item Format

This dimension captures the structure of a required response. *Breadth of Responses* captures whether there is a single or multiple acceptable form(s) for a successful response, as well as whether that form is overt in the task statement for the student. *Format of Responses* captures the extent to which the structure of a successful response is provided for the student. *Formality of Response* captures the explicit requirements for justification and support for a successful response. *Other Support* captures the extent to which the item requires the student to corroborate conclusions with secondary evidence or metacognition (e.g., checking work).

The ECF also has a dimension called *Item Format* that is significantly revised in the ICF. The ECF *Item Format* codes blend ideas of breadth, format, and support; additionally the ECF *Item Representation* code *Explanation* contained ideas that blended formality of support and other support with item components. Splitting and rearranging these ideas in this fashion represents the largest revision from ECF to ICF. The shift is from questions about the overt structure of the task and response to questions about how much of the structure of the task and response is made explicit or unknown for a student.

Item Components

This dimension captures the representations, objects, and statements in the item. These codes apply separately to both the task statement and the required response. For example, *Statement* is applied to tasks that contain a statement, such as a theorem or definition, whose truth-value is (framed as) known, while the same code applied to a response means that the student is required to produce such a statement. The *Claim* code is applied to tasks containing statements with unknown truth-value and to responses that require students to decide on the truth-value of a statement or to generate a statement with unknown status, such as a conjecture.

Data and Results

Tables 2 and 3 summarize the frequencies and ranges seen for the exams in this data set. Averages are computed from the percentages of each exam set, rather than from the total collection of items, to give the same weight to each participant.

Comparing Courses

Item Orientation and Item Format. To compare exams, we label each as average (within 10% of the group average), or otherwise high/low frequency for each code. For example, P1 and P3 have low frequencies of items with single answers; P1 has correspondingly high frequency on questions with forked responses, while P3 is high on delineated and open items. Similarly, P1

and P3 are both low in items asking students to remember declarative facts, but P1 is high in terms of asking students to reproduce arguments, and P3 is high in tasks that ask students to apply understanding or analyze.

Table 2: Observed frequencies and ranges for Item Orientation and Item Format codes

Item Orientation			Item Format		
Code	Average	Min-Max	Code	Average	Min-Max
Remember	14%	0% - 36%	Single	64%	16% - 100%
Recall and apply proc	30%	0% - 66%	Forked	22%	0% - 79%
Recall and reproduce argument	12%	0% - 47%	Delineated	7%	0% - 36%
Understand	0%	0%	Open	7%	0% - 18%
Apply understanding	38%	19% - 69%	Multiple choice/ TF	16%	0% - 38%
Analyze	2%	0% - 8%	Fill in the blank	11%	0% - 39%
Evaluate	5%	0% - 20%	Short answer	44%	11% - 62%
Create	0%	0%	Long answer	28%	2% - 74%
Recreate	37%	7% - 79%	No support	40%	0% - 93%
Adapt	60%	21% - 93%	Informal support	33%	3% - 57%
New	3%	0% - 19%	Formal support	27%	0% - 74%
Low certainty	3%	0% - 14%	Unclear	0%	0% - 3%
Medium certainty	24%	2% - 46%	Neither	87%	62% - 100%
High certainty	73%	54% - 98%	Interpretation/ Context	1%	0% - 5%
			Control/Evaluation	9%	0% - 38%
			Both	2%	0% - 14%

Table 3: Observed frequencies and ranges for Item Component codes

Item Components	Task		Response	
Code	Average	Min-Max	Average	Min-Max
Applied	13%	0% - 46%	8%	0% - 32%
Symbolic	71%	12% - 100%	64%	16% - 100%
Verbal	15%	0% - 64%	12%	0% - 56%
Graphical	15%	0% - 40%	12%	0% - 56%
Tabular	9%	0% - 31%	6%	0% - 15%
Statement	14%	0% - 38%	9%	0% - 46%
Claim	33%	4% - 79%	33%	0% - 79%
Example	8%	0% - 31%	16%	0% - 44%

Of potential interest for detecting instructors' authentic instruction practices (Gulikers et al. 2004) through exams are the codes that capture uncertainty and open-ended tasks. In Table 4, we summarize the data by participant for four such codes or combinations: analyze, evaluate, and create (A+E+C); new; forked, delineated, and open tasks (F+D+O); and claim.

Table 4: Observed frequencies for combined uncertainty/open-ended codes by course

	P1	P2	P3	P4	P5	P6	P7
A+E+C	5%	8%	**28%**	0%	7%	0%	0%
New	0%	**19%**	0%	0%	0%	0%	0%
F+D+O	**84%**	0%	**80%**	31%	29%	13%	12%
Claim	**79%**	**54%**	**48%**	38%	4%	7%	36%

Item Component. We separate the task and response component codes. Table 5 shows that P3 is high frequency in five *Item Component* subcodes, which is more than the other exam sets, and is also the only course to be high frequency for more student response subcodes than task codes.

The ICF appears to capture the distinctive demands of teaching subfields of mathematics. P5 is the statistics course, and it has the highest frequency of applied components. P1, P2, and P4 are proof-based courses that have symbolic representations in every task and response; P2 and P4 are introductions to proof, with high frequencies of theorem and definition statements in tasks. P3 and P6 are courses for pre-service elementary teachers with lower than average use of symbolic representations and higher than average use of graphical and geometric representations. The high frequency of *Recall and apply procedure* in P5 and P7 may encode the fact that they are lower-division, computational courses.

Table 5: Item Component code frequencies (light/medium/dark represents low/average/high frequency)

Task	P1	P2	P3	P4	P5	P6	P7
Applied	0%	0%	20%	0%	46%	23%	0%
Symbolic	100%	100%	12%	100%	75%	23%	86%
Verbal	5%	0%	64%	0%	0%	20%	12%
Graphical	0%	0%	40%	0%	14%	36%	12%
Tabular	0%	3%	0%	31%	25%	7%	0%
Thm/Dfn	5%	38%	12%	38%	0%	0%	2%
Claim	79%	43%	40%	23%	4%	7%	36%
Example	0%	8%	20%	31%	0%	0%	0%
Response							
Applied	0%	0%	20%	0%	32%	2%	0%
Symbolic	100%	100%	16%	100%	43%	36%	52%
Verbal	0%	5%	56%	0%	21%	0%	2%
Graphical	0%	3%	56%	0%	0%	18%	5%
Tabular	11%	3%	4%	15%	0%	7%	0%
Thm/Dfn	0%	46%	0%	0%	18%	0%	0%
Claim	79%	54%	48%	38%	4%	0%	7%
Example	26%	0%	44%	23%	7%	11%	2%

Discussion

Our evidence tentatively supports the claim that the ICF also detects differences among teaching practices in similar courses. For example, P3 and P6 are both courses for future elementary teachers, but P3's exams include more higher-order and open-ended summary codes in Table 4, and more item components, especially those required in the response, while P6 is average or below in each of these indicators. We suggest that P3 is asking more, or perhaps more authentic mathematics, of students than P6, which could indicate that its instructor holds a more-developed teaching perspective. There are similar, if weaker, patterns of difference between P2 and P4 (introduction to proofs, discrete) and P5 and P7 (lower-division, computational, applied).

These analyses also highlight the ways P3, and to a lesser extent P1 and P2, are asking students to do mathematics that is potentially more authentic (Gulikers et al., 2004) on exams than P4, P5, P6, and P7. P1 accomplishes this by asking students to prove or disprove statements, P2 by asking students to work with new definitions, and P3 by asking students to evaluate arguments and to coordinate multiple goals simultaneously. The evidence and analysis above is consistent with the assertion that the ICF captures dimensions of teaching that are of interest to professional developers of mathematics instructors.

We would predict that items that use different representations in their task and response would be more complex and demanding for students. Tallman et al. (2016) use statistical methods to determine if task components correlate with response components or other codes, but

this kind of analysis is not possible on our small sample. In a larger sample, this might highlight another aspect of more-developed teaching practice.

The next phase of the project will involve coordinating the analyses of these seven participants' teaching using other data sets, including their syllabi, video recordings of their classrooms, and surveys of both the participants and their students, for which method development and coding have proceeded independently. Initial conversation indicates that these different data sets generally highlight a similar subset of courses as exhibiting valued teaching aspects, but these data will also highlight different aspects of their teaching, such as espoused theory (Argyris, 1976) from instructor surveys to contrast with theory-in-use from exams.

We have not yet tried to code and test for inter-rater reliability. Thus far, reliability rests on three points of content validity. The first author re-coded the data repeatedly until the framework and codes stabilized, and justified codes and changes to the framework to the other researchers. Restructuring the item format dimension around epistemological questions in particular helped the team agree on their understanding of codes. Finally, the *Certainty* codes serve as a measure of confidence in the coding. The majority of items were coded as medium (24%) or high (73%) certainty. If higher certainty is desired, items coded as medium certainty could be resolved either by scanning the course textbook to see if the question was familiar or asking the instructor to complete a simple survey declaring the familiarity of each item on their exams. These approaches are both more invasive than simply collecting exams, but could be ways to gather this information easily in future rounds of data collection. Next steps for this project must include reliability testing across multiple raters.

Thus far, claims about the utility of the ICF rest on the analysis of a small sample, which is intertwined with the researchers' experience with professional development of mathematics instructors, including advocacy for active and inquiry-based pedagogies. The local goal is to develop a method for coding exams so that we can understand whether and how analyzing exams may be helpful to characterizing teaching. If this method proves useful, the larger goal is to detect change in the instructors who participate in these kinds of professional development. The target teaching outcomes for this kind of professional development are often broad; developing a measure that is focused on assessments and that can be applied widely across course topics may contribute to detecting change in dimensions not currently studied in other ways. We do not claim that an ideal exam is entirely higher-order cognitive tasks, but we do believe that high quality teaching would include requiring students to engage *some* higher-order tasks on exams and that ask students to work in uncertainty. We do not think that an ideal exam completely avoids symbolic representations, but we have valued those exams that avoid using only symbolic representations and that ask students to reason with multiple representations. We also need to connect the ICF to existing research to solidify these utility claims.

Future research could explore these questions, some of which are analogous to those explored by Tallman et al. (2016). Are exams for courses other than calculus changing across time in the discipline? What correlations exist among the codes in the ICF (in a sample large enough for statistical analyses), and how do these correlations depend on course level/domain? To what extent are mathematics students asked to use and translate between multiple representations in their (high stakes, timed) assessments? With additional instructor data, how do instructors' perspectives about their exams related to researcher analyses of the items, and what are the relationships between instructors' stated values and their assessment practices? Using a coordinated and longitudinal data set from professional development, do changes in exams lead or trail other teaching changes in response to professional development?

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.

Argyris, C. (1976). Theories of action that inhibit individual learning. *American Psychologist*, *31*(9), 638.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81-90.

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.

Conference Board of the Mathematical Sciences (2016). *Active Learning in Post-Secondary Mathematics Education*. Retrieved from http://www.cbmsweb.org/Statements/Active_Learning_Statement.pdf

Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational technology research and development*, *52*(3), 67-86.

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.

Mesa, V., Suh, H., Blake, T., & Whittemore, T. (2012). Examples in college algebra textbooks: opportunities for students' learning. *Problems, Resources, and Issues in Mathematics Undergraduate Studies*, 23(1), 76–105.

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.

Tallman, M. A., Carlson, M. P., Bressoud, D. M., & Pearson, M. (2016). A characterization of calculus I final exams in US colleges and universities. *International Journal of Research in Undergraduate Mathematics Education*, *2*(1), 105-133.