


# Informed and Uninformed Naïve Assessment Constructors' Strategies for Item Selection

Journal of Teacher Education  
2017, Vol. 68(1) 85–101  
© 2016 American Association of  
Colleges for Teacher Education  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/0022487116668019  
jte.sagepub.com  


Helenrose Fives<sup>1</sup> and Nicole Barnes<sup>1</sup>

## Abstract

We present a descriptive analysis of 53 naïve assessment constructors' explanations for selecting test items to include on a summative assessment. We randomly assigned participants to an informed and uninformed condition (i.e., informed participants read an article describing a Table of Specifications). Through recursive thematic analyses of participants' explanations, we identified 14 distinct strategies that coalesced into three families of strategies: Alignment, Item Evaluation, and Affective Evaluation. We describe the nature of the strategies and the degree to which participants used strategies with frequency and effect size analysis. Results can inform teacher education on assessment construction through explicit instruction in the three families of strategies identified.

## Keywords

assessment, teacher education preparation, preservice teacher education, strategy use, test construction

Teachers use classroom level assessments to inform instruction and to make evaluations regarding student learning and progress. Investigations of preservice and practicing teachers' assessment knowledge, practice, and strategies have focused on their general conceptions of assessment literacy (e.g., Siegel & Wissehr, 2011; Volante & Fazio, 2007), understanding of measurement principles (e.g., Gotch & French, 2013), and use of varied assessment types (DeLuca, Chavez, & Cao, 2013). Moreover, this work suggests a lack of general assessment knowledge and skill for how to apply and integrate that knowledge into teaching practices. It could be that the general level at which assessment is taught in courses and addressed in research fails to transfer to meaningful practice when teachers are asked to generate and implement their own strategic processes when constructing or selecting assessments. The lack of connection between measurement theory and measurement construction (see DeLuca & Bellara, 2013; Schafer & Lissitz, 1987) may lead preservice and practicing teachers to rely on less effective or limiting assessment construction strategies instead of those that could provide sound information for making valid inferences about student learning.

## Relevant Research

In this section, we describe current expectations and research on preservice teachers' knowledge of classroom assessment. We examine the task of test construction, as one component of the assessment-related knowledge preservice teachers must develop. We suggest that test construction be considered

a complex cognitive task and as such argue that investigations of strategic processes for test construction could inform preservice preparation in this area.

## *Preservice Teachers' Classroom Assessment Knowledge and Practices*

Brookhart (2011) identified assessment-related knowledge and skills teachers need, including the ability to (a) construct and communicate learning objectives; (b) design, use, draw inferences from, and provide feedback to students on a range of assessments; (c) administer, interpret, and communicate results of external assessments; and (d) help students use assessment results to inform their decisions. Moreover, researchers have argued for the importance of planning for assessment (Fives, Barnes, Dacey, & Gillis, 2016), providing corrective feedback (Hattie & Timperley, 2007), and using formative assessment (Black & Wiliam, 2009) as strategies for improving student learning.

To respond to the demand for teachers to have greater assessment literacy, many educator preparation programs require preservice teachers to receive some coursework in educational assessment (DeLuca & Klinger, 2010). However,

<sup>1</sup>Montclair State University, NJ, USA

### Corresponding Author:

Helenrose Fives, Department of Educational Foundations, Montclair State University, 1 Normal Avenue, Montclair, NJ 07043, USA.  
Email: fivesh@mail.montclair.edu

assessment is typically taught as a single course offered in one semester, which limits the number of assessment-related issues that can be explored and provides little time for preservice teachers to practice integrating assessment knowledge into their instructional practices (DeLuca & Klinger, 2010; P. Graham, 2005). Textbooks crafted to prepare teachers for classroom assessment typically include recommendations for the construction of specific types of assessments; however, there is wide variation in the depth of coverage with little focus on “how to” apply assessment principles to practice (Campbell & Collins, 2007; Fives et al., 2016). Thus, it is not surprising that preservice teachers continue to report that they feel ill-prepared in their understanding of assessment and how to use assessment to improve teaching and student performance (e.g., Campbell & Evans, 2000; Maclellan, 2004; Volante & Fazio, 2007).

Volante and Fazio (2007) found that despite efforts to teach elementary preservice teachers a range of assessment techniques related to observation and evaluation, they continued to rely on the same assessment practices and had difficulty creating assessment systems that included a range and variety of assessments. After reviewing 65 preservice teachers’ lesson plans, Campbell and Evans (2000) reported that out of the 309 lesson plans reviewed, only 213 included information about *both* instructional goals and assessment, and in only 53 of them were instructional goals correctly aligned with assessment tasks (i.e., evidence of validity; p. 353). Thus, despite efforts to increase preservice teachers’ assessment-related knowledge and practices, they continue to remain weak in this area.

### The Task of Test Construction

Valid judgments about students’ knowledge, skills, and learning are based on the quality of information teachers obtain from the assessments they design and use with their students (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). The most recent edition of the *Standards for Educational and Psychological Tests* (AERA, APA, & NCME, 2014) details the kinds of evidence needed to develop a validity argument in support of using a particular test or measure. The *Standards* articulate five sources of evidence based on test content, response process, internal structure, relations to other variables, and related consequences. Scholars and researchers in classroom assessment, test construction, and use of assessment(s) in the classroom have argued that standards required for large-scale test producers and researchers need to be modified and made accessible and reasonable for classroom teachers (McMillan, 2003; Wolming & Wikstrom, 2010). For example, in his articulation of classroom assessment literacy for teachers, Popham (2009) underscored the importance of three types of validity evidence necessary for making appropriate inferences about

students; we refer to these evidence types as test content, response process, and relations to other variables.

When designing a *particular* assessment, the systematic alignment of instructional goals, learning activities, and assessment items are the hallmarks of valid evaluations of student learning and progress (Chappuis & Stiggins, 2008). A Table of Specifications (TOS), sometimes called a test blueprint, is a table that helps teachers map a test or other assessment to their instructional objectives and instructional priorities for a given segment of study (Fives & DiDonato-Barnes, 2013; see Appendix for an example). As such, the TOS is a planning tool intended to help teachers align objectives, instruction, and assessment practices so that content-related and response process-related evidence become the focus during test construction. The TOS places issues of validity evidence central to decision making during test construction as it requires teachers to consider the underlying purpose and alignment of their assessment tasks. The TOS also offers a bounded framework for discussing validity by focusing on a small segment of content in a very concrete way that may be more accessible to preservice and novice teachers.

In our previous work with the TOS as a test construction tool, we found group differences between preservice teachers (i.e., naïve assessment constructors: individuals with limited, if any, formal preparation in constructing classroom assessments) who received instruction in using a TOS (informed;  $n = 28$ ) and those who did not (uninformed;  $n = 25$ ; DiDonato-Barnes, Fives, & Krause, 2013). Specifically, preservice teachers exposed to the TOS tool constructed tests with higher test content evidence (TCE) but not response process evidence (RPE) scores<sup>1</sup> than those without TOS exposure. Our findings echoed early work by Carter (1984) who asked 310 practicing teachers to evaluate multiple choice items for the reading comprehension skills of detail, main idea, inference, and prediction. She found that while about half the teachers could identify items assessing detail and main idea, only a third or fewer could do so for items at the higher cognitive levels of inference and prediction. This work points to the cognitive complexity test writers face when they must consider the response processes expected of test takers.

In our TOS study we also found that some preservice teachers in the uninformed group constructed tests with high TCE, suggesting that they engaged in some form of strategic process rather than a random selection of items (DiDonato-Barnes et al., 2013). Similarly, the lack of difference with respect to RPE between the two groups indicated that informed participants may have employed strategies other than those supported by the TOS during test construction. Wise, Lukin, and Roos (1991) examined practicing teachers’ perspectives on the sources of their knowledge of measurement. Descriptive analysis of responses from practicing teachers indicated that the majority of teachers who had not had courses in measurement (59%) felt that “learning by trial and error in one’s classes” had the greatest effect on their testing and measurement knowledge (Wise et al., 1991, p. 39).

The data from these two studies suggest that in the absence of formal preparation, teachers (both preservice and practicing) will develop their own assessment strategies through trial and error, which may be both time consuming and disadvantageous to the Pre-K–12 children who are exposed to this experimentation.

### *Test Construction as a Complex Cognitive Task*

Complex cognitive tasks “require the integration of skills, knowledge, and attitudes and the extensive coordination of constituent skills in new problem situations” (van Merriënboer, Kirschner, & Kester, 2003, p. 6). Tasks that are more complex necessitate broad, deep knowledge and a variety of strategic processes for successful and fluid task completion (van Merriënboer et al., 2003). We argue that test construction is a complex cognitive task, as it requires test writers to consider the subject matter (declarative content as well as content relevant concepts, practices, and strategies), the test takers’ cognitive processes, test planning, and item construction techniques. By conceptualizing test construction as a complex cognitive task rather than a series of routine procedures or techniques, the importance of cognitive strategies for test construction becomes salient.

Cognitive strategies are a form of procedural knowledge that individuals consciously employ when engaged in goal-directed tasks such as identifying the main idea in a short story, organizing an essay, or solving a physics problem (Alexander, Grossnickle, Dumas, & Hattan, in press; MacArthur, 2012). MacArthur (2012) conceived of cognitive strategies as either domain general (e.g., rehearsal, summarization) or domain specific (e.g., mnemonic for the order of operations in mathematics, in English: Please Excuse My Dear Aunt Sally). Strategies can also be organized by function, for instance, Duckworth, Gendler, and Gross (2014) framed students’ self-control strategies into five families, each of which have an underlying goal in common. For example, one family of self-control strategies is *situation selection* and strategies in this family reflect ways that learners choose situations that enable them to experience success in self-control, such as “studying in the library rather than at home to avoid distraction” (Duckworth et al., 2014, p. 206).

Researchers have identified strategies for learning and engagement in academic domains through (a) the close review of expert performance (e.g., Pressley & Afflerbach, 1995) and (b) observing learners in the domain (e.g., Siegler, 1996). Strategies consciously used by experts or learners can be understood and explicitly taught to learners who demonstrate less competence (MacArthur, 2012; Siegler, 1996). For example, research in reading (National Reading Panel, 2000), writing (S. Graham, 2006), and mathematics (Laski et al., 2013) has found positive learning and performance outcomes because of strategy training. Cognitive Strategy Instruction (CSI) is used to describe a range of strategy development models (e.g., Self-Regulated Strategy Development [SRSD];

Harris et al., 2012) that all use a combination of explicit and guided instruction to help students acquire and practice new strategies (S. Graham & Harris, 2009; Krawec & Montague, 2012). Researchers investigating the effectiveness of CSI models have found that with continued practice and effective teacher support, learners exhibited increased knowledge and application of learned cognitive strategies and overall improvements in problem-solving performance (Case, Harris, & Graham, 1992; Montague, 2008; Montague, Enders, & Dietz, 2011). The foundation of CSI rests on the identification of relevant domain-specific strategies that can facilitate cognitive processing, problem solving, and task completion.

If one considers test construction a complex cognitive task, as we do, then it follows that we need to identify relevant strategies to support this task so that they can be taught to current and future teachers who need support and development in this activity. Specifically, domain-specific strategies that facilitate the coordination of the multifaceted concerns teachers have when developing tests may enhance their test writing practices and provide a foundation for future development.

### **Rationale and Research Questions**

In this study, we examine the strategies reported by naïve assessment constructors. As mentioned earlier, naïve assessment constructors refer to those individuals with limited, if any, formal preparation for constructing classroom assessments. Studying the strategic processes of this group is informative from both research and practice perspectives. First, from a research perspective, studying this population can lead to the identification of emergent strategies that naïve assessment constructors use. In doing so, this investigation may highlight the power of the apprenticeship of observation on all aspects of learning to teach, including assessment practices (Lortie, 1975). Second, as individuals develop competence in a field, their use of strategies becomes more effective, flexible, and elegant, frequently shifting from an intentional cognitive strategy to an automated skill, thereby making it difficult to access the depth and complexity of strategies used. In contrast, strategies are highly salient and subsequently more accessible for observation when performed by individuals new to the field who have less experience and preparation for targeted tasks (Alexander, Graham, & Harris, 1998). The National Reading Panel (2000), for example, found that studying beginning readers was a more effective way to identify learners’ decoding strategies since expert readers engaged in these processes automatically and unconsciously. From a practice perspective, the exploration of naïve assessment constructors provides insight for teacher educators into the intuitive strategic processes of potential learners. Thus, examining naïve assessment constructors’ explanations for item selection may provide greater insight into their decision-making strategies as they engaged in this task.

Therefore, the purpose of this investigation was to answer the following research questions:

**Research Question 1:** What strategies do naïve assessment constructors use in the selection of items to include on an end-of-unit test?

**Research Question 2:** How does strategy use differ between naïve assessment constructors who received instruction on the TOS (i.e., the informed group) versus those who did not (i.e., the uninformed group)?

## Method

We used inductive thematic qualitative analysis (Miles, Huberman, & Saldaña, 2014) to examine open-ended written rationales provided by two groups of naïve assessment constructors who we asked to select items for an end of unit test. Using this inductive approach allowed us to develop an explanation of the phenomenon of strategy use for assessment construction situated in the data gathered. The analyses provided here represent an extension of our previous quantitative study with these participants to ascertain the emergent strategies for assessment construction employed by these participants (DiDonato-Barnes et al., 2013). To achieve this goal, we examined the data both holistically (all data from all participants) and by condition using three analytical approaches: (a) thematic analysis to identify strategy families and strategies used, (b) frequency counts and percentages to describe numerically overall strategy use and variations in strategy by condition, and (c) effect size analysis to identify the magnitude of differences in strategy family used by condition. Onwuegbuzie (2003) argued for the inclusion of frequency and intensity counts (percentages) as well as effect size analyses when reporting thematic analyses of qualitative data to provide a level of “*empirical precision*” along with the “*descriptive precision*” inherent in the qualitative findings (p. 396, emphasis in original). Together these analyses allowed us to develop a multidimensional representation of the assessment construction strategies used by naïve assessment constructors.

## Participants

Fifty-three students (28, informed; 25, uninformed), from five undergraduate educational psychology classes, agreed to participate and provided complete data for this investigation. This 200-level course fulfilled a general education requirement in social sciences for students who are typically in their first or second year of study. Most students enrolled in this course because they intend to apply to the teacher education program for subject area certification, and this course is a prerequisite for that program. Participants ranged in age from 18 to 33 years, were predominantly female (75%), and the majority (92%) indicated that they intended to become teachers. Participants described themselves as White (66%), Hispanic (15%), Other (8%), African American (6%), and Asian American (6%).

## Procedures

We drew data for this study from a larger investigation of test construction practices (DiDonato-Barnes et al., 2013). In this investigation, we asked participants to construct an end-of-unit test for a fictitious fifth-grade class by selecting 10 items (seven multiple choice and three short answer) from a test bank we prepared. We gave participants unit materials (on the Southern Colonies of the United States) and a 44-item test bank. The unit plan included 11 objectives over 4 days of instruction, with details as to the amount of class time given to lesson activities for each objective. The test bank included four items per objective that varied in terms of type (multiple choice or short answer) and cognitive level (low or high, according to the Revised Bloom’s Taxonomy; Anderson et al., 2001). We organized items in the test bank by type; first, multiple choice items followed by short answer items. Within each section, we randomly rather than sequentially presented the items. In addition to selecting items, we asked participants to “explain why” they chose to include or exclude each item in the test bank. Figure 1 provides an excerpt from the study task test bank that illustrates the task and provides sample items.

We used a self-selection sampling strategy where participants volunteered to be part of this study in exchange for extra credit offered by their course instructors. We visited class sessions and interested students completed a sign-up sheet for scheduled research sessions held outside of class time in a university conference room. We contacted interested participants via email to remind them of upcoming sessions. Prior to each session, we compiled study packets that included all of the materials needed to participate in the investigation for each condition and arranged them randomly. As participants arrived to the room, we gave them a packet of materials thereby ensuring that each participant had an equal chance of being assigned to each condition. Thus, we randomly assigned participants to either an informed or uninformed condition. In our previous study, a one-way analysis of variance indicated that both groups were equivalent on pretest knowledge of assessment,  $F(1, 51) = .265, p = .609$  (DiDonato-Barnes et al., 2013, p. 101).

We gave participants in the informed condition a short article that explained the TOS strategy and a partially complete TOS tool that included learning objectives, time spent on each task, and the percent of time spent on each topic from overall time spent on the unit (see Appendix). Informed participants completed the TOS tool by determining the number of items (out of 10) to ask about each objective and the cognitive level the items should assess. We devised the unit and TOS such that of the 11 objectives, three should not be assessed and two should be assessed using two items.

## Analyses

We analyzed data for this investigation holistically (all data from all participants) and by condition. Our holistic analyses



<b>Test Bank</b>		
<i>Guidelines: Choose 7 multiple choice questions and 3 short answer questions from this test bank. Please list the item numbers that you have chosen to be a part of this study at the end of each section.</i>		
Items	Include on test?	Explain Why
<b>15. People in the southern colonies were most likely to make their living by growing</b> a. cranberries. b. peanuts. c. soy beans. d. tobacco.	<input type="checkbox"/> Yes  <input type="checkbox"/> No	
<b>21. Which of the following colonists would have made the best living by settling in the Southern Colonies?</b> a. Herbert Miln, a boot maker b. George Mitchim, a Protestant Minister c. Simon Cowell, a music producer d. Michael Warren, a planter e. Timothy Calhoune, a rancher	<input type="checkbox"/> Yes  <input type="checkbox"/> No	
<b>44. In 1-2 sentences describe at least three ways that people in the southern colonies made a living.</b>	<input type="checkbox"/> Yes  <input type="checkbox"/> No	
<b>26. Pretend you are a colonist during the 1700's. Which colony would you settle in? Why? Describe the type of work you and your spouse would do. Explain why this work would be suited for this region.</b>	<input type="checkbox"/> Yes  <input type="checkbox"/> No	

**Figure 1.** Test bank excerpt with sample low-level (#15, 44) and high-level (#21, 26) items for one lesson objective.

included thematic analysis and frequency analysis (i.e., counts and percentages of strategies identified in the thematic analysis). Our comparative analyses by condition included effect size analysis at the strategy family level and frequency analysis at the strategy level.

*Thematic analysis.* Research assistants transcribed all data into a spreadsheet. We conducted a thematic analysis of these data and engaged in recursive emergent coding (Miles et al., 2014) to answer Research Question 1 (i.e., What strategies do naïve assessment constructors use in the selection of items to include on an end-of-unit test?). In our inductive analysis, we first generated idea-unit codes for each strategy or rationale given by participants for selecting or rejecting a particular test item presented in the text bank. Idea units reflected each independent idea or thought. We identified initial codes during the transcription of the data. At that time, the co-authors discussed potential terms and collective definitions

for common response patterns. Code development was exhaustive and recursive. To best represent the data, we added new codes as needed. We initially used several codes that we later collapsed or separated. After all data were coded, we closely reviewed all responses coded with terms we felt required deeper analysis, namely those coded as “should,” “other,” and “quality.” We jointly reviewed these responses and came to agreement about the nature of the code. We maintained data memos throughout this process. This process allowed us to fully explicate the underlying conceptual meaning of each code.

After we completed the coding process and determined that idea-unit codes had reached saturation, we conducted a final review of the themes. In some instances, we divided larger themes into smaller groupings that allowed for a better and more descriptive identification of the data. The reader should note that, when appropriate, we often coded a single statement from a participant for multiple strategies. Based on these data,

we identified three overarching categories or strategy families in response to Research Question 1 that participants used to make decisions about item selection: (a) Alignment, (b) Item Evaluation, and (c) Affective Evaluation. We discuss the nature of the specific strategies by family in the results section.

**Frequency analyses.** To provide additional empirical precision to our findings for both Research Questions 1 and 2, we conducted counts and percentages at the family and strategy level. For Research Question 1, we calculated percentages of the number of times participants used strategies by and within families. Then, within each family, we calculated the percentage of times participants reported each individual strategy. In addition, within families, we counted the number of participants who used each individual strategy at least once and calculated the frequency of strategy use by participant.

For Research Question 2, we calculated the percentage of strategies used within each family for each condition (informed and uninformed). This allowed us to descriptively see differences in specific strategy use by participants in each condition. To get a sense of the intensity of strategy use by participants, we calculated the percentage of participants from each condition who used each individual strategy at least once.

**Effect size analysis.** In addition to frequency analysis, we also used effect size analysis to address Research Question 2. According to Onwuegbuzie (2003), "There are many instances in which effect sizes provide a thicker description of underlying qualitative data" (p. 394). Although our study is qualitative in nature, the use of effect size calculations allowed us to "avoid underrating or overrating the importance frequencies associated with qualitative categories" (Aulls & Ibrahim, 2012, p. 124) and instead provided further support for the descriptive results we observed. Based on these data, we were able to explore how strategy use differed between naïve assessment constructors who received instruction on the TOS and those who did not, as well as the magnitude of those differences.

## Results

### Description of Strategy Families: Holistic Content and Frequency Analyses

Evidence from the rationales participants provided and the range and frequency of the use of these rationales suggested that participants engaged in a systematic process when selecting items for a classroom test, thus they were strategic (Alexander et al., 1998). We organized the strategic choices of these participants into three overarching families of strategies to guide item selection or rejection: (a) Alignment, (b) Item Evaluation, and (c) Affective Evaluation.

**Holistic thematic analysis.** Table 1 is organized by family and provides a description of the specific strategies within each

family and sample responses from our participants. We describe the nature of the specific strategies within each family in the sections that follow.

**Alignment.** The first family, *Alignment*, included strategies that sought to provide alignment between the classroom experience and test items. Specific strategies included in this theme were content coverage, cognitive level, class time, learning objectives, and TOS. When using *content coverage*, participants focused on whether or not the item reflected content covered during the lesson. Responses indicated this strategy referred to the subject matter in general (e.g., geography) or the specific topic (e.g., indigo dye) or class resource (e.g., class presentation slides) that was included in the materials provided. Participants using this strategy sought to ensure that the items on the test were reflective of the material taught in class. Variation in the use of this strategy ranged from a direct alignment with class materials (e.g., "Learned the definition from the Powerpoint," id 4-10-2-1\_U) to a more nuanced consideration of relevance of the content in context of the unit (e.g., "This fact was only mentioned once throughout the unit and did not hold as much meaning to the overall unit," 4-10-2-8\_U).

Participants employed the *cognitive level* strategy when they considered the kind of thinking and level of cognitive processing needed for the fictitious student to complete each item. Responses associated with cognitive level illustrated that participants were concerned with the quality and complexity of the thinking required by each item, but that they did not necessarily consider these things in relation to the learning objectives. For instance, one participant reasoned "I like #18 better than this question about Georgia because this question is just about *spitting out facts, it doesn't tell us if he/she understands why*" (id 4-11-1-8\_U, emphasis added).

Participants also used *class time* as a strategy for selecting or rejecting test items. That is, participants reasoned that the amount of time spent on the topic each item assessed (as indicated in the lesson plans) was sufficient (or not) for the item to be included on the test. Some references to class time were explicit with the participants actually using the language "class time" and indicating a relationship between the amount of time spent on the information assessed in the item and whether the item should be included. In some instances, the reference to time was more implicit. For instance one participant wrote, "not enough of the lesson put into Maryland" (id 4-18-1-1\_I), which we interpreted to mean that not enough time in the lesson addressed this topic.

Alignment with the *learning objective(s)* described in the lesson plans was another strategy that emerged to guide decisions about item selection. Responses that included the term "objective" explicitly were included in this theme. This suggested that the participants were considering more than the content presented in lesson materials; they were actively considering the instructional objectives for the unit.

**Table 1.** Family, Strategy, Definition, and Sample Response.

Family: Strategy		Meaning	Sample response	
Alignment	Content	Referred to subject matter (e.g., geography), teaching materials (e.g., lesson plan), or specific content (e.g., Eliza Pickney)	Because a woman discovered something in the 18th century (id 4-25-1-1_U)	
	Cognitive level	Referenced the cognitive level (low or high) required to complete the item	Because this asks a question based on what students know but also forces them to think critically rather than just repeat a definition (id 4-10-2-5_U)	
	Class time	Time spent in class, according to the lesson plan, on the instruction of the subject matter was described either explicitly or implicitly	Not enough of the lesson put into Maryland (id 4-18-1-1_U)	
	Learning objective	Described the alignment of the item to a learning objective from the lesson as a rationale. Used the word “objective”	That objective wasn’t in my choice questions (id 4-10-2-4_U)	
	Table of Specifications	Response referenced the TOS either explicitly or as inferred by us	I didn’t choose this for a question on the TOS (id 4-25-1-1_U)	
	Item Evaluation	Quality	Referred to the quality of the item itself as a reason for using. Addressed issues such as wording, image quality, or inferred interpretation of the item	I didn’t choose this because it said “most.” I wanted to base the test on facts (id 4-25-1-1_U)
		Item to test	Referred to other items selected for the test as part of the decision-making process	Used a fact and opinion question in short answers (id 4-20-2-4_U)
		Type	Referred to the type of item (mc or short answer) as the reason for inclusion/exclusion	I think this would be better as a short answer (id 4-10-2-4_U)
Developmental level		Made reference to the capabilities of fifth-grade students or the students taking the test as a reason for accepting or rejecting item	Fifth-grade students most likely would like to see scenarios to help them understand concept (id 4-20-1-6_U)	
Affective Response	Importance	Referred to the perceived importance of the item content either for the unit or for the field in general	It is important for students to know historical figures and where things originated (id 4-20-1-6_U)	
	Self-referencing	Participant referred to own knowledge base or experience as a rationale for their decision	Questions that have words like “most” always tended to confuse or second guess myself, I think a more straight forward question would be better (4-11-1-5_U)	
	Motivation	Indicated that the item would motivate or demotivate the student	This is a fun, exciting method of testing that would excite students to become engaged while allowing them to use knowledge learned in class (id 4-11-1-2_C)	
	Epistemic cognition	Referenced the nature of the knowledge to be assessed	Not as important as history questions (id 4-11-1-1_U)	
	Should	Indicated that students should know the content but it was unclear if it should be known because it was <i>important</i> or because the <i>content</i> was covered in <i>class time</i> allotted	Students should be able to know this (id 4-19-1-2_U)	

TOS as a strategic response referred to instances in which participants explicitly referred to using the TOS provided to the informed group when making a decision about item selection. For some participants this was a simple use of the tool (e.g., “not on TOS” id 4-24-12-1-1\_U) and others provided more elaboration on their thinking about using the TOS (e.g., “In the TOS, Column B, Day 1, Row A asks to identify Southern Colonies; therefore, this question can help in terms of recalling—lower level thinking” id 4-20-2-5\_U). However, if a participant from the informed condition did not explicitly refer to the TOS tool but seemed to be

considering aspects included on the TOS (e.g., “Follows the objective and is a lower level question” id 4-5-2-2\_U), we did not assume that the TOS strategy was in use.

*Item Evaluation.* The second strategy family, *Item Evaluation*, included strategies that focused on evaluating individual items to determine if they were appropriate for inclusion on the test. This theme included four specific strategies that required item-level evaluations to be made as part of the item selection/rejection process: quality, item-to-test, type, and developmental level. The *quality* strategy involved

evaluating the item in terms of the wording, image quality, or perceived “trickiness” of the item. For instance, one participant reported that he/she chose an item because it was “Visual, straight to the point” (id 4-18-1-1\_U). Other participants were concerned about the messiness of the map items provided and the generality or specificity of items. Alternatively, participants evaluated some items as being very good or clear. Thus, participants used their own quality evaluations as reasons to include or exclude test items.

The *item-to-test* strategy referred to participants’ intentional comparison of individual items to the test as a whole or to other items selected. Participants using this strategy seemed to be aware of the content representation on the test they were constructing in terms of the number of items by content and the format used. For example, one participant kept track of items he/she already selected related to the same content (id 4-20-2-4\_U). A special form of Item Evaluation focused on the appropriateness of *item type* for assessing content topics. Participants employed this item type strategy when they considered the nature of the item type, multiple choice or short answer, as part of their item selection process.

Developmental level emerged as a strategy for item selection as well. When using *developmental level*, participants referred to their perceptions of what would be appropriate for fifth-grade students. For instance, several participants using this strategy thought that students at this age would like scenarios or were concerned that the items were asking too much of students in this developmental stage. Note, we distinguish the *developmental level* strategy from the *cognitive level* strategy in the Alignment family of strategies, by nature of the perspective taken. In Item Evaluation, participants were making judgments on specific items in terms of their perspective on how the average fifth-grade student would respond (or not) to the item. In contrast, in the Alignment family, participants considered *cognitive level* in relation to the material covered in the unit.

**Affective Evaluation.** The final family, *Affective Evaluation*, included strategies that relied on the affective experience of or perceptions about the items by participants. As described in Table 1, this family included five distinct strategies used to guide item selection: importance, self-reference, motivation, epistemic cognition, and should. Participants referred to the *importance* of the item in terms of information for the field at large (i.e., history) or for the specific unit presented (i.e., Southern Colonies). Further, participants often cited “not important” or “not relevant” without further explanation as a reason to exclude an item on a test. Participants used this strategy in conjunction with *item-to-test* as a means to support a decision about a particular item (e.g., “#1 is not as important as #2/#3” id 4-20-1-6\_U).

*Self-referencing* emerged as a strategy when participants referred to their own knowledge or experience as a reason for including or excluding items. For instance one participant rejected an item and wrote “Because I didn’t know who

James Oglethorpe was” (id 4-25-1-1\_U). They also indicated preferences for or against particular kinds of items, as seen in the self-reference example in Table 1, where the participant described his/her frustration with items relying on “most” (id 4-11-1-5\_U). Other uses of self-reference included personal responses to the nature of an item (“Hate Venn diagrams. Kindergarteners use them” id 4-3-1-1\_U). Thus, participants used self-referencing in item selection in terms of the content value and aspects of the item format or quality.

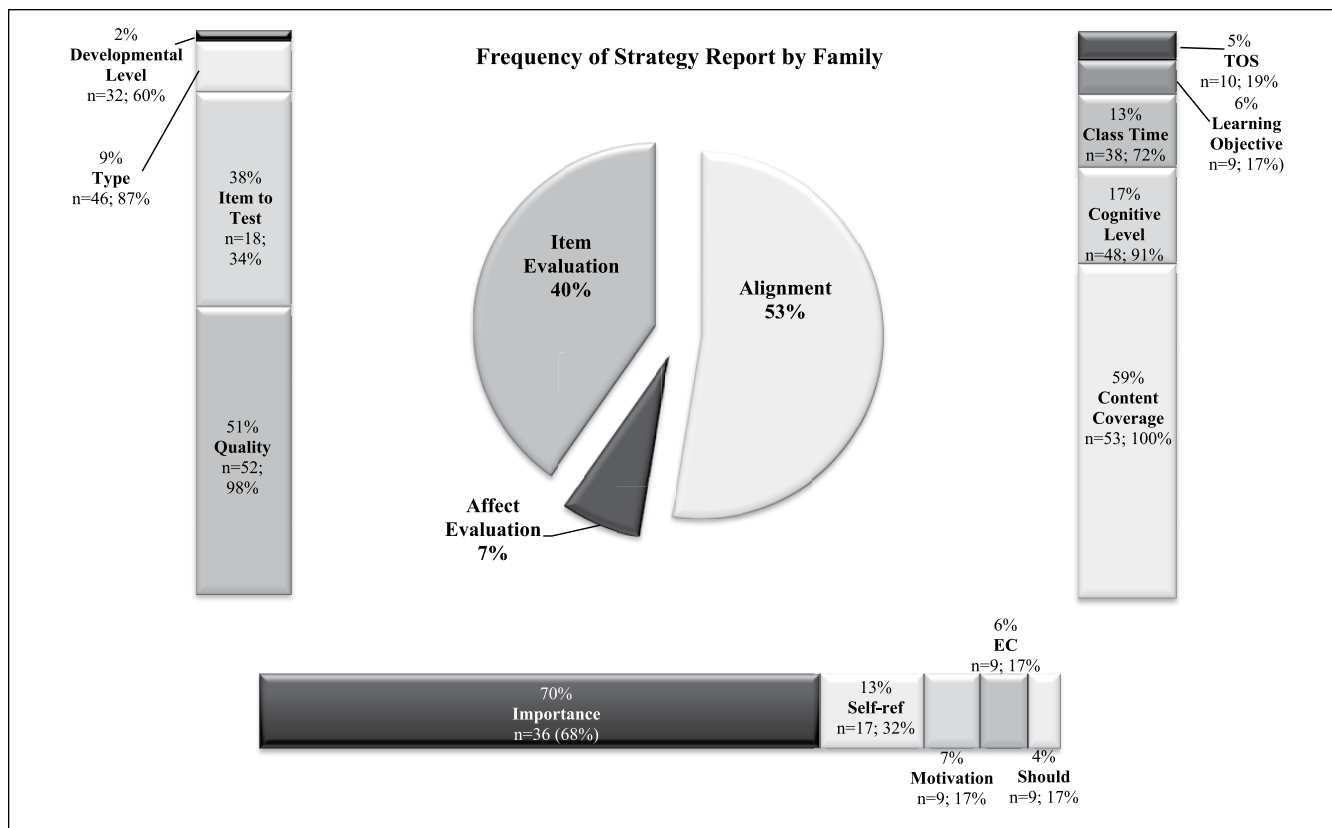
When participants used terms like overwhelming, boring, fun, interesting, and uninteresting to describe the items and the potential response from the students who will take the test, we saw this as employing *motivation* as a strategy for item selection. Positive affect typically led to the decision to include the item whereas negative affect, even if the participant deemed the item to be qualitatively “good” led to the rejection of the item. For instance, one student rejected an item stating, “Good question, but not the most exciting/creative question in the selection for this topic” (id 4-11-1-2\_U). Included with this strategy were inferences made about whether or not students would like or want to respond to a particular item (e.g., “Too long for a fifth-grade student to want to answer” id 4-11-1-1\_U). Some participants using this strategy considered how a particular item might serve to motivate or support students completing this test. For instance, one participant responded to an item this way, “It is a simple question to ease into the test, make the student confident and relaxed while completing the rest of the test” (id 4-20-1-5\_U).

Participants evoked epistemic cognition when individuals made judgments about inclusion of content within domains or across domains based on the nature of the knowledge included in the content. *Epistemic cognition* is described

as the cognitive process in which people engage while considering the nature and the justification of knowledge . . . As such, it refers to something that people do when they are prompted to reflect on the nature of what they regard as knowledge and on the [underlying] warrants. (Maggioni & Parkinson, 2008, pp. 446-447)

Recently, Buehl and Fives (2016) presented a model of teachers’ epistemic cognition, which highlighted the understanding that teachers must think about the nature of knowledge and knowing not only for themselves but also for their students. This seemed to emerge in the use of *epistemic cognition* as a strategy for item selection among these participants. For instance, two responses that focused on the nature of knowledge and knowing included “Students should not have to memorize information/definitions” (id 4-20-1-6\_U) and “[this item] Puts student in the past. A good starting point for great knowledge” (id 4-11-1-1\_U). These two rationales illustrate that these participants were considering the nature of knowledge and its relevance for the students who may take this test. This consideration was also seen in several decisions to reject items because they were too simple and





**Figure 2.** Frequency of strategy reports by and within families.

Note. On the within family bar graphs, the percentage above strategy indicates the frequency that strategy was used within the strategy family; the number below the strategy indicates the number of participants who used the strategy.

reflected common knowledge (e.g., “Too simple of a question, common knowledge” [id 4-3-11\_U]).

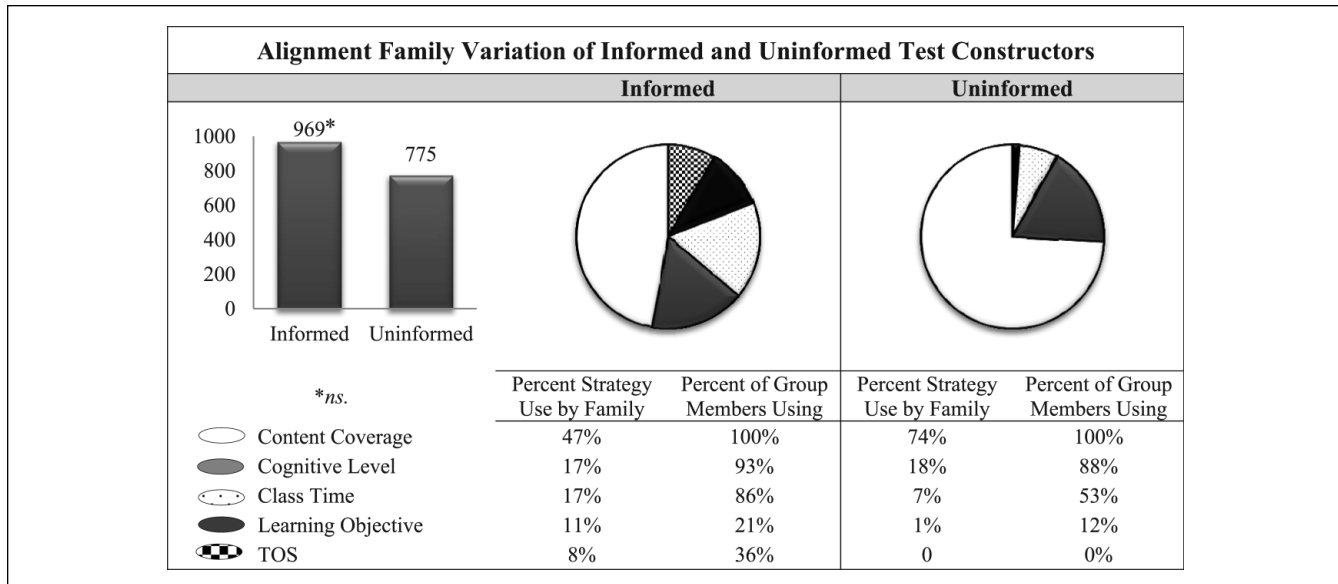
We coded responses as *should* whenever participants used the word “should,” and we could not infer if this meant the students should know the content because it was important (therefore coded as *importance*) or because it was so well covered in the lesson (therefore coded as *content* or *class time* depending on the specificity of the response). Thus, when we could not determine what the participant meant by “should,” we left them in this theme.

**Holistic frequency of strategy use by family.** The center of Figure 2 illustrates the frequency of strategy use by family and the three stacked bars depict the distribution of strategies in each family. In each bar the number above the strategy indicates the percentage of times each strategy was reported within that family and the numbers below the strategy refers to the number and percentage of participants who used the strategy. This information provides perspective on both the frequency of use as well as the degree to which these strategies spanned multiple users.

**Alignment.** Alignment was the most widely used strategy family among participants with 53% of all strategies reported falling into this family (center of Figure 2). Within Alignment,

content coverage was the most frequently used strategy in this family (59% of strategies reported) and was used by every participant in this investigation (right stacked bar in Figure 2). The majority of participants (91%) used cognitive level and this accounted for 17% of the overall Alignment strategies. More than half of participants (72%) used class time to make decisions about item selection. Of all the strategies in the Alignment family, explicit reference to use of the TOS was used the least in this category (5% of the total Alignment strategy responses) by 19% of the participants.

**Item Evaluation.** This was the second most frequent strategy family including 40% of all strategies reported in this investigation (center of Figure 2). Quality was the most frequently used strategy in this family and 98% of the participants used it (left stacked bar in Figure 2). Although the use of the item type strategy was limited, only 9% of all strategies in the Item Evaluation family, it was used by a majority of our sample (87%) at some point while engaged in the test construction task. It is worth noting that we purposefully generated the test bank to include four items for each objective, two multiple choice and two short answer, and within each type a high-level and low-level item was provided. Thus, the preference for item type related to some topics seemed based



**Figure 3.** Comparison of within Alignment family variation of informed and uninformed test constructors.

on participants' prior experiences with tests rather than an analysis of the specific items provided. Lastly, participants commented infrequently (2% of the Item Evaluation family) that the item was or was not appropriate for fifth-grade students (i.e., developmental level) for a variety of reasons.

**Affective Evaluation.** Affective Evaluation was used least frequently (7% of all strategies reported, see center of Figure 2). As illustrated in the bottom stacked bar in Figure 2, Importance was the most frequently used strategy in this family and was employed by a majority of the participants ( $n = 68\%$ ). Thirty-two percent of participants engaged in self-referencing when making decisions about item inclusion whereas 17% of participants considered the motivational aspects of items when making their selections. A portion of our participants (17%) engaged in the active consideration of the nature of knowledge in terms of whether topics count as "knowledge," the nature of knowledge in terms of its specificity or simplicity, and its stability. The final strategy, *should*, was used the least frequently in this family (4%) by 17% of participants.

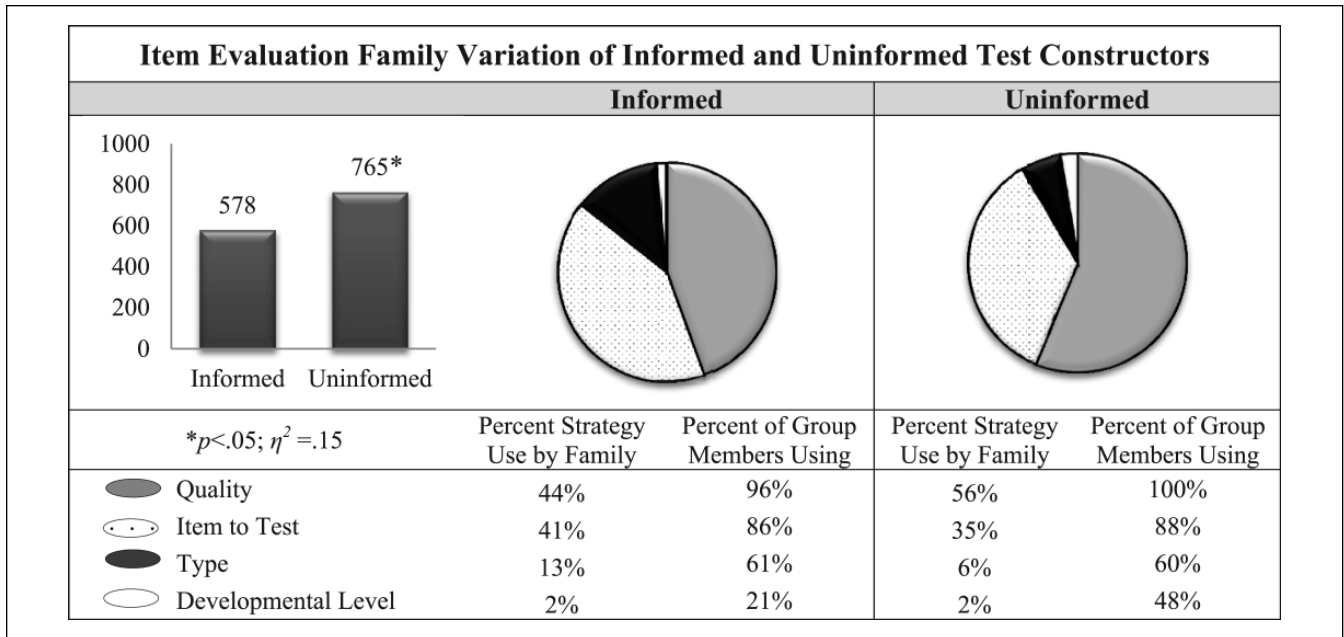
### Comparison of Informed and Uninformed Strategy Use Within Family

We explored differences in the frequency of strategy use by participants in each condition (informed and uninformed) within each family. Figures 3, 4, and 5 illustrate the comparative strategy reports for each strategy family, by group, indicating both the frequency of strategy use, the percent of strategy use by family, and the percentage of participants in each group reporting that strategy. We used the  $t$ -statistic, derived from running independent-samples  $t$  tests, to calculate effect sizes to determine the magnitude of strategy family use (i.e., Alignment, Item Evaluation, and Affective

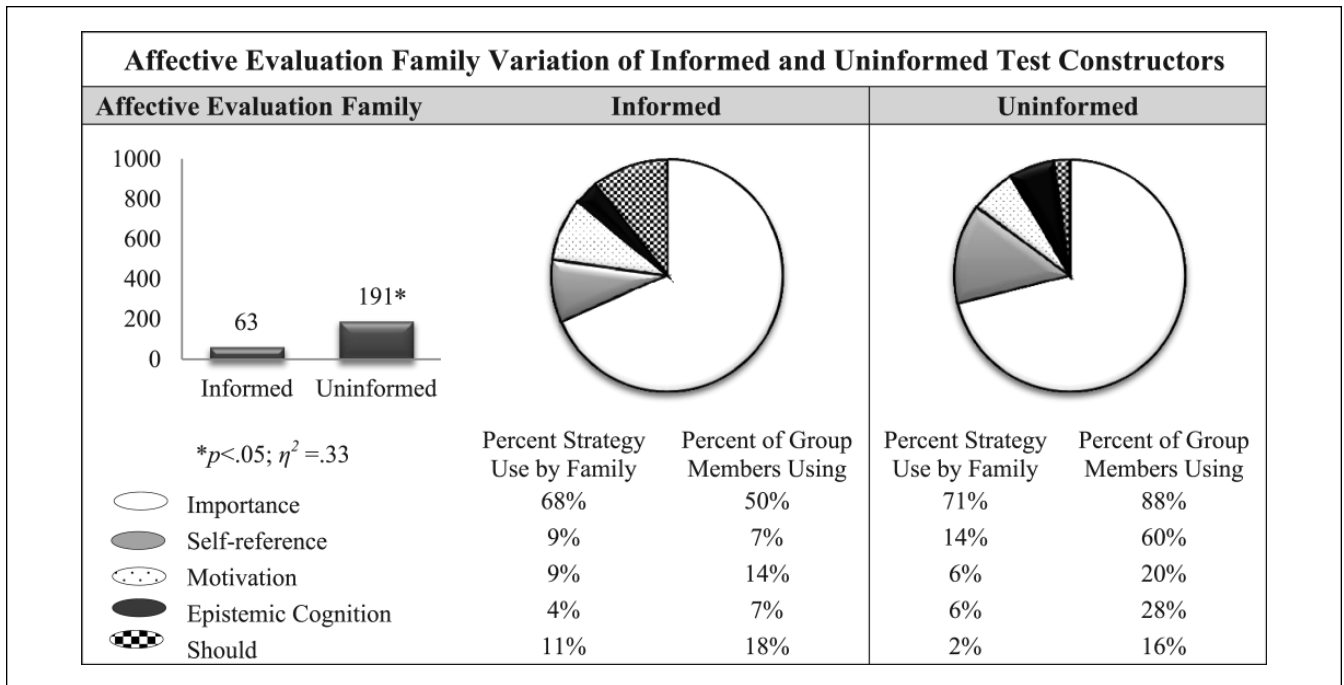
Evaluation) between participants in the informed and uninformed groups.

**Alignment.** With regard to the use of Alignment strategies, there was no significant difference in strategy use between the two groups,  $t(42) = -1.05, p > .05$ . Although participants in both groups used Alignment strategies for item selection, a review of the frequency in which they used particular strategies reveals nuanced differences between the types of Alignment strategies they relied on to make decisions. For example, all participants in this study used *content coverage* (Figure 3); however, the uninformed participants reported this strategy more frequently (U: 74% compared with I: 47%). With respect to *cognitive level*, the frequency of use across the two groups was similar (U: 18% compared with I: 17%), although this strategy was used by more of the participants in the informed group (93%) than by members of the uninformed group (88%). In other words, a greater number of informed participants used this strategy somewhat less frequently, and lesser number of uninformed participants used the strategy more frequently. The somewhat greater use of this strategy among informed participants is unsurprising as the instructional materials they received reviewed Bloom's Taxonomy (Anderson et al., 2001) and divided the lower levels (remember, understand) from the higher levels (apply, analyze, create, evaluate). The large percentage of uninformed participants evoking this strategy with consistency suggests that this may be a parameter of tests that they have noticed in their past experiences.

The remaining three Alignment strategies were more frequently used by a larger number of participants in the informed group. Thus, participants who received information on test construction via a TOS used *class time* (I: 17%; U: 7%), *learning objectives* (I: 11%; U: 1%), and the TOS



**Figure 4.** Comparison of within Item Evaluation family variation of informed and uninformed test constructors.



**Figure 5.** Comparison of within Affective Evaluation family variation of informed and uninformed test constructors.

(I: 8%; U: 0%) to guide their decision making about item inclusion more so than did participants in the uninformed group.

**Item Evaluation.** Participants in the informed and uninformed groups differed in their use of Item Evaluation strategies,  $t(51) = 3.00, p < .05$ , with the uninformed group ( $M = 30.60, SD = 13.03$ ) scoring higher than the informed

group ( $M = 20.64, SD = 11.09$ ) and the magnitude of this difference was large ( $\eta^2 = .15$ ). Figure 4 illustrates these differences. An examination of strategy use within this family suggested that all of the participants in the uninformed group used *quality* as a strategy for selecting items, and they did so more often (U: 56%; I: 44%) than the participants in the informed group. Conversely, the informed participants used *item-to-test* (I: 41%; U: 35%) and *item*

type (I: 13%; U: 6%) more frequently than uninformed participants. Participants in the both groups were equally likely to use *developmental level* as a strategy for item selection.

**Affective Evaluation.** Participants in the uninformed group were more likely to rely on affective strategies for item selection,  $t(34) = 5.12, p < .05, (M = 7.48, SD = 4.82)$  compared with participants in the informed group ( $M = 2.04, SD = 2.36$ ) and the magnitude this difference was also large ( $\eta^2 = .33$ ; Figure 5). Within the Alignment family, uninformed participants used *importance, self-reference, motivation, and epistemic cognition* to support their decisions to include or exclude items on the tests (see Figure 5). Most salient of these was to make decisions based on perceived *importance* of the content with 88% of the uninformed sample using this strategy more frequently (71%) and only 50% of the informed sample using it 68% of the time. A second trend to note in these responses is the use of *self-referencing*, which was used by more than half of the uninformed participants ( $n = 60\%$ ) and only a few informed participants ( $n = 7\%$ ). This difference may suggest that in the absence of assessment-related strategies, these participants instead relied on personal preferences to make their decisions.

## Discussion

A thematic analysis of naïve test constructors' rationales for selecting items for an end-of-unit fifth-grade Social Studies unit revealed that participants used a number of test construction strategies to select or reject items. We organized these strategies into three families: Alignment, Item Evaluation, and Affective Evaluation. A frequency analysis suggested that although participants systematically used strategies from all three strategy families, the majority of reported strategies came from the Alignment family, then Item Evaluation, and the fewest from Affective Evaluations. This indicates that these may be preliminary or core strategies that are needed during test construction. If that is the case, then future teachers may benefit from explicit instruction in how to use these techniques to develop checklists or other tracking systems to facilitate these cognitive processes while engaged in test construction. Initial strategies that may facilitate test construction include (a) the development and use of the TOS, (b) identification of cognitive levels, and (c) consideration of the item-to-test relationship. However, only use of the TOS currently has empirical evidence to support its use as a test construction tool (DiDonato-Barnes et al., 2013). The other strategies need explicit testing to determine the nature of their influence in constructing quality tests.

Next, we examined the strategy use of participants in the informed and uninformed groups. Participants in both groups consistently relied on rationales that reflected attention to issues related to Alignment for item selection. Because both

groups used Alignment strategies relatively equally, it suggests that some strategies, especially attention to *content*, may emerge for some future teachers in the context of test construction. Because the informed group was not statistically more likely to use Alignment strategies compared with the uninformed group, it may be that in addition to explicit instruction in the strategy, future teachers may also need scaffolded practice and feedback to facilitate strategy use.

Other researchers studying strategy development have noted similar findings. For example, in a qualitative study of teachers in a 4-year professional development program, Duffy (1993) argued that it takes repeated practice for learners to acquire the skills to seamlessly employ strategies to fit the task requirements. Therefore, preservice teachers may need an opportunity to practice the use of the TOS and the strategies that support that use through experiences that provide them with feedback and opportunities to observe others engaged in a similar task. Driscoll (2000) suggested that there are external and internal conditions that affect strategic development. For example, learners are more likely to enact strategies when the strategies are described or demonstrated for them, when they have opportunities to practice using the strategy, and when they are given feedback about their performance. Internal conditions that affect strategy acquisition include "prior knowledge of the simple concepts and rules that make up highly general strategies (such as breaking larger problems into subparts) or task-relevant concepts, rules, and information" (p. 360).

Participants in the uninformed group relied on Item Evaluation and Affective Evaluation strategies more often than the informed group to guide item selection. Recall that Wise and colleagues (1991) found that without formal preparation in assessment, teachers (preservice and practicing) developed their own assessment strategies through trial and error. The findings from this study suggest preservice teachers who were not taught otherwise relied more on affective strategies and personal opinion. If preservice teachers are unaware of the strategies that they use intuitively and are not exposed to alternative (and possibly better) strategies in their preparation programs, then when they enter the classroom and begin to learn by trial and error they will most likely continue to rely on those intuitive strategies. Therefore, teacher educators need to elicit preservice teachers' intuitive strategies, and teach them how to determine when to rely on these strategies and when to utilize strategies that reflect more sound assessment practices.

## Limitations

A limitation to this investigation is the homogeneity of our sample of 53 undergraduate students from a single university in the United States. The cultural context of schools in the United States and the region of this university in particular may have informed these participants' sense of what tests should be. This may have limited the range and frequency of



strategies reported. Second, the study task (particularly for students in the informed group) was time consuming and appeared to be cognitively fatiguing for participants. Task fatigue may have influenced both the strategy use of these participants as well as their motivation to continue to explain their item selection decisions in writing. The nature of the data collection may have inhibited the number of strategies recorded. Participants completed this task independently and wrote explanations for their choices. Because we did not observe them individually, we may have missed additional strategies or techniques such as skimming through the test bank, shifting back and forth between items, or re-evaluating a selected item.

### Implications for Teacher Educators

The results of this investigation provide a strategy framework for instruction on assessment construction that can be used directly with preservice and practicing teachers to scaffold knowledge and skill development. Teacher educators can guide learners to more appropriate strategies (and away from less appropriate ones) within each family and facilitate deliberate practice on their use. In particular, deliberate practice with evolving difficulty related to the strategies that currently have sound theoretical and empirical support is warranted. For instance, in the Alignment family, instruction on how to use a TOS is supported by research to improve the quality of TCE (DiDonato-Barnes et al., 2013). Regarding Item Evaluation, extensive recommendations for practice on constructing test items and thereby evaluating the quality of each is supported in the field (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005).

In addition, the emergence of (a) three families of strategies, (b) the use of multiple strategies by item selected, and (c) the use of multiple strategies across items by individuals indicates that these naïve assessment constructors weighed multiple perspectives when completing the test construction task. Teacher educators can use these families of strategies to design learning experiences that help preservice and practicing teachers to actively integrate salient strategies during assessment construction and develop professional judgment about how aspects of alignment, item-level concerns, and the affective experience of the assessment should be valued when making decisions. Naïve assessment constructors (particularly those intending to become teachers) need to engage in meaningful reflection on and dialogue with the goals and purposes of assessment activities as related to the goals and purposes of the overall educational enterprise. This includes meeting the socio-emotional and developmental needs of each potential test taker along with ensuring that they can make valid evaluations about student learning. Instructional experiences such as ongoing reflection and ethical debates informed by the rich theory and evidence available from the fields of assessment, child development, motivation, and learning should be provided in preservice teacher preparation and ongoing professional development.

### Implications for Research

Continued calls for assessment literacy (e.g., Popham, 2009), data literacy (e.g., Mandinach & Gummer, 2013), and widespread use of formative assessments, performance assessments, and data-based decision making (e.g., Marsh, 2012) evidence the need for empirically supported strategies and techniques that preservice and practicing teachers can use to build their repertoire of practice. However, the field of classroom assessment seems bereft of empirically supported domain-specific strategies and techniques that can be adapted for use across classrooms. A next step for research in the area of assessment construction is to design investigations that assess the effect of the strategies identified in this investigation on the quality of tests constructed. In such investigations, test quality could include attention to validity evidence of test content, response process, and relations to other variables as these were three areas of validity evidence endorsed by Popham (2009) as salient for classroom teachers.

Common among the responses in our data were participants use of *multiple strategies* when they made decisions about items. For instance, the quote below illustrates how one participant employed five different strategies spanning the three themes identified:

Assuming the map would be better, [Item Evaluation-*quality*] I think it is important that the student learn [Affect-*importance*] and apply their knowledge [Alignment-*cognitive level*] with reading a resource and product map [Alignment-*content*]. I remember having numerous maps on tests throughout my years of schooling [Affect-*self-referencing*; id 4-20-1-5 C].

This suggests that it may be important to examine how different configurations of strategies can be used to influence the decisions made about test items.

In this investigation, we examined the strategies reported by naïve assessment constructors. These participants had limited formal knowledge (if any) in classroom assessment, the fifth-grade Social Studies curriculum, or the nature of fifth-grade learners. Thus, the range and quality of strategies identified is limited by the experiences of our participants. Future research exploring the test construction strategies of practicing teachers is warranted to identify the techniques used by individuals with experience in classroom assessment practice. Classroom teachers who have an understanding of content, context, learners, and assessment represent a unique form of assessment expertise whose voice is largely absent from this field of research. Such investigations could explore the classroom assessment craft knowledge held by practicing teachers and the ways that knowledge influences the strategies they evoke when constructing classroom tests.

Researchers need to engage in an exploration of the developmental nature of assessment knowledge and practice for preservice and practicing teachers to identify relevant learning trajectories that facilitate the integration of teachers' knowledge of subject area, assessment theory and

practice, students' developmental capabilities, motivational influences, socio-emotional effects, and their particular learning context. Prominent experts in the field of classroom assessment have articulated the necessary components of the knowledge base for classroom assessment (Brookhart, 2011) or assessment literacy (Popham, 2009). Such work is crucial for framing the field and identifying salient content needed for instruction. What is also needed is a more pragmatic identification of central or foundational knowledge that can be taught in the frequently limited time allocated to assessment in preservice preparation and used as a basis to facilitate ongoing professional development in this area.

## Conclusion

This investigation underscores the complex interplay of cognition, knowledge, and affect that are engaged when preservice teachers construct classroom assessments. The nature of this study involved selecting 10-items for an end-of-unit

assessment—a task that classroom teachers complete routinely. The range and depth of strategies employed by novices completing this task evidences the need for explicit instruction in supportive strategies that can become automated skills. Further, explicit connections from theory (e.g., validity) to classroom recommendations (e.g., align instruction with assessment) to specific strategies and techniques for practice (e.g., TOS) are necessary so that teachers can become fluid in their use and adaptations of theory, strategy, and technique in their specific contexts. Fives and Buehl (2014) referred to this approach to theory and practice as the *McGyver Mentality*, which, if fostered, promotes teachers' active use of knowledge of theory, context, and strategy (technique) to suitably adapt to meet the needs of their students. In the absence of such instruction, preservice and practicing teachers are left to rely on their own intuition and develop heuristics based more in belief (e.g., items with the word "most" should be avoided and Venn Diagrams are for Kindergarteners) than in sound evidence and theory of classroom assessment.

## Appendix

### Table of Specifications: The Southern Colonies

Fifth-Grade Social Studies

Chapter 7: The Southern Colonies

A	B Instructional objectives	C Time spent on topic (minutes)	D Percent of class time on topic	E Number of test items: 10	F		G	
					Number of lower levels items -Knowledge -Recall -Identification -Comprehension	Number of higher levels items -Application -Analysis -Evaluation -Synthesis		
Day 1	Identify the southern colonies on a map	5	3.0%	.03*	0	—	—	—
	Identify who colonized Maryland and explain why people colonized Maryland	5	3.0%	.03	0	—	—	—
Day 2	Explain why people colonized the Carolinas and describe how Eliza Lucas Pinckney's discovery impacted the crop industry	15	9.1%	.91	1	—	—	—
	Explain why people colonized Georgia	15	9.1%	.91	1	—	—	—
	Predict how did people in each of the southern colonies made a living	15	9.1%	.91	—	1	—	1
	Describe the difference between fact and opinion	15	9.1%	.91	1	—	—	—
Day 3	Analyze information and determining whether it is fact or opinion	15	9.1%	.91	1	—	—	—
	Apply geographic tools, including legends and symbols, to collect, analyze, and interpret data	30	18.2%	1.82	—	—	—	2
	Explain the geographic factors that influenced the development of plantations in the Southern Colonies	5	3.0%	.03	0	—	—	—

(continued)

## Appendix (continued)

## Fifth-Grade Social Studies

## Chapter 7: The Southern Colonies

A	B	C	D	E	F	G
	Instructional objectives	Time spent on topic (minutes)	Percent of class time on topic	Number of test items: 10	Number of lower levels items -Knowledge -Recall -Identification -Comprehension	Number of higher levels items -Application -Analysis -Evaluation -Synthesis
Day 4	Compare and contrast the life of a slave and a planter	30	18.2%	<i>1.82</i>	—	2
	Identify the characteristics of an indentured servant	15	9.1%	<i>.91</i>	<i>1</i>	—
	<b>Totals</b>	<b>165</b>	<b>100.0%</b>	<b>10</b>	<b>5</b>	<b>5</b>

Note. Columns A, B, C, and D were completed for participants who then filled in columns E, F, G, we have filled these columns in using italics, with the correct responses for each. For each objective only high-level or low-level item(s) should be selected; we indicated the appropriate cognitive level for each item by including a "0" in the cognitive level column for objectives that should not be assessed.

## Authors' Note

A previous version of this paper was presented at the 2014 Annual Meeting of the American Educational Research Association. The authors contributed equally to the completion of this research.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Note

1. Scores were determined by comparing the items participants selected from the test bank to an expert Table of Specifications that was developed following the model presented in the article read by the treatment group. We awarded points for selecting items that accurately reflected the subject matter (test content evidence) and the cognitive level of the objective (response process evidence) related to the test item. See DiDonato-Barnes, Fives, and Krause (2013) for additional details.

## References

- Alexander, P. A., Graham, S., & Harris, K. R. (1998). A perspective on strategy research: Progress and prospects. *Educational Psychology Review*, *10*(2), 129-154. doi:10.1023/A:1022185502996
- Alexander, P. A., Grossnickle, E. M., Dumas, D., & Hattan, C. (in press). A retrospective and prospective examination of cognitive strategies and academic development: Where have we come in twenty-five years? In A. O'Donnell (Ed.), *Handbook of educational psychology*. Oxford, UK: Oxford University Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., . . . Wittrock, M. C. (2001). *Taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Needham Heights, MA: Allyn & Bacon.
- Aulls, M., & Ibrahim, A. (2012). Pre-service teachers' perceptions of effective inquiry instruction: Are effective instruction and effective inquiry instruction essentially the same? *Instructional Science*, *40*, 119-139. doi:10.1007/s11251-010-9164-z
- Black, P., & Wiliam, D. (2009). Developing a theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5-31.
- Brookhart, S. M. (2011). Tailoring feedback. *Education Digest*, *76*(9), 33-36.
- Buehl, M. M., & Fives, H. (2016). The role of epistemic cognition in teacher learning and praxis. In J. A. Greene, W. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 247-264). New York, NY: Routledge.
- Campbell, C., & Collins, V. (2007). Identifying essential topics in general and special education introductory assessment textbooks. *Educational Measurement: Issues and Practice*, *26*(1), 9-18. doi:10.1111/j.1745-3992.2007.00084.x
- Campbell, C., & Evans, J. A. (2000). Investigation of preservice teachers' classroom assessment practices during student teaching. *The Journal of Educational Research*, *93*(6), 350-355. doi:10.1080/00220670009598729
- Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education*, *35*(6), 57-60.
- Case, L. P., Harris, K. R., & Graham, S. (1992). Improving the mathematical problem-solving skills of students with learning disabilities: Self-regulated strategy development. *The Journal*

- of *Special Education*, 26(1), 1-19. doi:10.1177/002246699202600101
- Chappuis, S., & Stiggins, R. (2008). Finding balance: Assessment in the middle school classroom. *Middle Ground*, 12(2), 12-15.
- DeLuca, C., & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education*, 64(4), 356-372. doi:10.1177/0022487113488144
- DeLuca, C., Chavez, T., & Cao, C. (2013). Establishing a foundation for valid teacher judgment on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice*, 20(1), 107-126. doi:10.1080/0969594X.2012.668870
- DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, 17(1), 419-438. doi:10.1080/0969594X.2010.516643
- DiDonato-Barnes, N. C., Fives, H., & Krause, E. (2013). Using a table of specifications to improve teacher constructed traditional tests: An experimental design. *Assessment in Education: Principles, Policy & Practice*, 21(1), 90-108. doi:10.1080/0969594X.2013.808173
- Driscoll, M. P. (2000). *Psychology of learning for instruction* (2nd ed.). Boston, MA: Allyn & Bacon.
- Duckworth, A. L., Gendler, T. S., & Gross, J. (2014). Self-control in school-age children. *Educational Psychologist*, 49(3), 199-217. doi:10.1080/00461520.2014.926225
- Duffy, G. G. (1993). Rethinking strategy instruction: Four teachers' development and their low achievers' understandings. *The Elementary School Journal*, 93(3), 231-247. doi:10.1086/461724
- Fives, H., Barnes, N., Dacey, C. M., & Gillis, A. (2016). Assessing assessment texts: Where is planning? *The Teacher Educator*, 51(1), 70-89. doi:10.1080/08878730.2015.1107442
- Fives, H., & Buehl, M. M. (2014). Exploring differences in practicing teachers' valuing of pedagogical knowledge based on teaching ability beliefs. *Journal of Teacher Education*, 65(5), 435-488. doi:10.1177/0022487114541813
- Fives, H., & DiDonato-Barnes, N. C. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research & Evaluation*, 18(1). Retrieved from <http://pareonline.net/pdf/v18n3.pdf>
- Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, 21(4), 357-364. doi:10.1016/j.tate.2005.01.008
- Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-efficacy for measurement concepts. *The Teacher Educator*, 48(1), 46-57. doi:10.1080/08878730.2012.740150
- Graham, P. (2005). Classroom-based assessment: Changing knowledge and practices through preservice teacher education. *Teaching and Teacher Education*, 21(6), 607-621. doi:10.1016/j.tate.2005.05.001
- Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187-207). New York, NY: Guilford.
- Graham, S., & Harris, K. R. (2009). Almost 30 years of writing research: Making sense of it all with The Wrath of Khan. *Learning Disabilities Research & Practice*, 24(2), 58-68. doi:10.1111/j.1540-5826.2009.01277.x
- Harris, K. R., Lane, K. L., Graham, S., Driscoll, S., Sandmel, K., Brindle, M., & Schatschneider, C. (2012). Practice-based professional development for self-regulated strategies development in writing: A randomized controlled study. *Journal of Teacher Education*, 63(2), 103-119.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. doi:10.3102/003465430298487
- Krawec, J., & Montague, M. (2012). Cognitive strategy instruction. *Current Practice Alerts*, 19, 1-4.
- Laski, E. V., Casey, B. M., Yu, Q., Dulaney, A., Heyman, M., & Dearing, E. (2013). Spatial skills as a predictor of first grade girls' use of higher level arithmetic strategies. *Learning and Individual Differences*, 23, 123-130. doi:10.1016/j.lindif.2012.08.001
- Lortie, D. C. (1975). *Schoolteacher*. Chicago, IL: The University of Chicago Press.
- MacArthur, C. A. (2012). Strategies instruction. In K. R. Harris, S. Graham, & T. Urdu (Eds.), *APA educational psychology handbook: Volume 3 application to learning and teaching* (pp. 379-401). Washington, DC: American Psychological Association.
- MacLellan, E. (2004). Initial knowledge states about assessment: Novice teachers' conceptualizations. *Teaching and Teacher Education*, 20(5), 525-535. doi:10.1016/j.tate.2004.04.008
- Maggioni, L., & Parkinson, M. M. (2008). The role of teacher epistemic cognition, epistemic beliefs, and calibration in instruction. *Educational Psychology Review*, 20(4), 445-461. doi:10.1007/s10648-008-9081-8
- Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, 42(1), 30-37. doi:10.3102/0013189X12459803
- Marsh, J. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, 114(11), 1-48.
- McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, 22(4), 34-43. doi:10.1111/j.1745-3992.2003.tb00142.x
- Miles, M., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis. A methods sourcebook*. Thousand Oaks, CA: SAGE.
- Montague, M. (2008). Self-regulation strategies to improve mathematical problem solving for students with learning disabilities. *Learning Disability Quarterly*, 31(1), 37-44. doi:10.2307/30035524
- Montague, M., Enders, C., & Dietz, S. (2011). Effects of cognitive strategy instruction on math problem solving of middle school students with learning disability. *Learning Disability Quarterly*, 34(4), 262-272. doi:10.1177/0731948711421762
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (Report No. 00-4769). Washington, DC: National Institute of Child Health and Human Development.
- Onwuegbuzie, A. J. (2003). Effect sizes in qualitative research: A prolegomenon. *Quality & Quantity*, 37(4), 393-409.



- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*, 4-11. doi:10.1080/00405840802577536
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.
- Schafer, W.D., & Lissitz, R.W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57-63. doi:10.1177/002248718703800312
- Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391. doi:10.1007/s10972-011-9231-6
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York, NY: Oxford University Press.
- van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner' mind: Instructional design for complex learning. *Educational Psychologist, 38*(1), 5-13. doi:10.1207/S15326985EP3801\_2
- Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*(3), 749-770. doi:10.2307/20466661
- Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37-42. doi:10.1177/002248719104200106
- Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice, 17*(2), 117-132. doi:10.1080/09695941003693856

### Author Biographies

**Helenrose Fives** is a professor of educational psychology in the Department of Educational Foundations at Montclair State University. Her research focuses on teachers' beliefs (i.e., beliefs about learning/ability, knowing/epistemic, teaching, assessment, and self-efficacy) and teachers' classroom assessment practices.

**Nicole Barnes** is an associate professor of educational psychology in the Department of Educational Foundations at Montclair State University. Her research focuses on empirically supported classroom assessment techniques and data use to improve teaching and learning, teachers' beliefs about assessment, and peer and teacher supports for self-regulated learning in middle school classrooms.