Assessment in Education: Principles, Policy & Practice, 2013
http://dx.doi.org/10.1080/0969594X.2013.808173

Routledge
Taylor & Francis Group

Using a Table of Specifications to improve teacher-constructed traditional tests: an experimental design

Nicole DiDonato-Barnes*, Helenrose Fives and Emily S. Krause

Educational Foundations, Montclair State University, Montclair, NJ, USA

(*Received 31 August 2012; final version received 17 May 2013*)

We investigated if instruction on a Table of Specifications (TOS) would influence the quality of classroom test construction. Results should prove informative for educational researchers, teacher educators, and practising teachers interested in evidenced-based strategies that may improve assessment-related practices. Fifty-three college undergraduates were randomly assigned to an experimental (exposed to the TOS strategy) and a comparison condition (no specific strategy support) and given materials for an instructional unit to use to construct a classroom test. Results of a multivariate analysis of covariance suggested that students exposed to the TOS strategy constructed a test with higher test content evidence but not response process evidence scores. Furthermore, we found that treatment participants were able to accurately complete the TOS tool and choose items that reflected the subject matter specified in the TOS tool. However, they experienced difficulty selecting items at the cognitive level specified in the TOS tool.

Keywords: validity; summative assessments; teacher-made tests

When teachers engage in high-quality formative, summative, or diagnostic assessment practices they are able to derive accurate inferences about students' knowledge and skills, and can depend on valid and reliable data to guide future instruction (Brookhart, 1999). Despite this awareness, little attention has been given to ensure that teachers can collect 'good' data (data that provide accurate and sufficient evidence to make decisions) and interpret it in light of their professional needs (e.g. to guide instruction and assign grades). It is surprising that in the current climate of evidence-based practice, we continue to offer teachers 'rules of thumb' for test construction (e.g. Haladyna, Downing, & Rodriguez, 2002) and theoretically guided approaches to aligning assessment with instruction (e.g. Table of Specifications TOS; Notar, Zuelke, Wilson, & Yunker, 2004) rather than empirically supported strategies.

Theoretical framework

Quality test construction

Evaluation of test construction rests on the validity of the evaluations made based on the data gathered. The concept of validity has a long and complex history in the field of educational measurement (e.g. Lissitz & Samuelsen, 2007). The 1999 *Standards for Educational and Psychological Testing* (American Educational

*Corresponding author. Email: didonaton@mail.montclair.edu

Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) define validity as, 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' (p. 9). Early conceptions of validity suggested multiple types of validity (e.g. content, criterion, and construct) for which evidence should be gathered. However, the 1999 standards, based on Messick's (1989) conceptualisation of the construct, established a unitary conception of validity that can be evaluated through several kinds of evidence. A chain metaphor for evaluation of validity was offered suggesting a movement through eight stages or links (i.e. administration, scoring, aggregation, generalisation, extrapolation, evaluation, decision and impact) in which any link may involve a breakdown in validity evidence (Crooks, Kane, & Cohen, 1996). The combination of the nature of several kinds of validity evidence and the chain of assessment activities can provide test constructors with multiple lenses for evaluating their assessment activities for validity evidence within each stage or link. Importantly, when this theoretical conception of validity is presented to teachers, attention must be given to their context and professional needs in understanding this conception (McMillan, 2003). Thus, a more pragmatic understanding of validity for a classroom teacher is to conceive of validity as the degree to which the evaluations or judgements teachers make about their students can be trusted based on the quality of evidence gathered (Wolming & Wikstrom, 2010). Classroom teachers are faced with the challenge of creating classroom tests that demonstrate validity evidence based on test content (AERA, APA, & NCME, 1999): '… including items, tasks, formats, wording, and processes required of examinees' (Goodwin & Leech, 2003, p. 183). For classroom teachers who must rely on their own expertise, this evidence needs to ensure that their assessment or test items adequately assess the subject matter that was taught (i.e. test content evidence [TCE]) as well as the appropriate level of cognitive processing, or level of difficulty, to match how the subject matter was taught (i.e. response process evidence [RPE]). This level of cognitive processing is frequently identified using a taxonomy of cognitive processes such as those developed by Bloom, Engelhart, Furst, Hill, and Krathwohl (1956), Biggs and Collis (1982), Anderson et al. (2001) or Marzano (2001).

In the USA, the American Federation of Teachers, National Council on Measurement in Education and the National Education Association developed seven *Standards for Teacher Competence in Educational Assessment of Students* (1990). Plake, Impara, and Fager (1993) developed and administered a 35-item multiple-choice test evaluating knowledge for each of these seven standards to 555 teachers and 286 administrators from 38 of the 50 states in the USA. Participants scored highest on the items related to administering, scoring and interpreting test results and poorest on items about communicating results. Participants who had training in measurement tended to do better on the test overall.

Using the same measurement tool, in the Midwest USA, Mertler (2004) compared preservice and practising secondary teachers' assessment literacy. Comparisons between preservice and practising teachers revealed that on five of the seven subscales, practising teachers scored higher than preservice teachers (i.e. choosing appropriate assessment methods; developing appropriate assessment methods; administering, scoring and interpreting the results of assessments; using assessment results to make decisions; and recognising unethical or illegal practices). Further, for both preservice and practising teachers, the scale that received the lowest scores

assessed the standard of developing valid grading procedures. Across these two studies, we see that practising teachers demonstrated the greatest knowledge for administering, scoring and interpreting the results of the assessments. However, knowledge of developing valid grading procedures and communicating assessment results lagged behind. It could well be that these two bodies of knowledge are related; that is, if you are unable to develop a sound grading procedure, then communicating your findings may be difficult. In fact, we would argue that some level of expertise is needed across the seven standards in order for teachers to engage in assessment practices that would provide them with information to make valid judgements about their students.

Based on the 1990 standards for teacher competence in educational assessment and the 20 years of intervening research, Brookhart (2011) offered 11 areas of assessment knowledge and skills for teachers. Notable for quality test construction are items:

- II-Teachers should be able to articulate clear learning intentions that are congruent with both the content and depth of thinking implied by standards and curriculum goals, in such a way that they are attainable and assessable (Brookhart, 2011, p. 7).
- V-Teachers should have the skills to analyse classroom questions, test items and performance assessment tasks to ascertain the specific knowledge and thinking skills required for students to do them (Brookhart, 2011, p. 8).

These two sets of knowledge and skills are fundamental to teachers' ability to construct tests that meet the basic expectations for validity evidence based on test content. Teachers need to be aware of the connections to assessment during instructional planning and classroom interactions (McMillan, 2003). Further, they need to be skilled in ascertaining the cognitive processing level of their instruction and subsequent assessment in order to provide an aligned curriculum (Anderson et al., 2001).

Concerns about teacher-made tests

Stiggins (1999) estimated that US teachers spend up to half of their professional time on assessment-related activities. Further, teachers often rely largely on data from their own assessments to make decisions about students' knowledge and skills despite the findings that these measures often lack good diagnostic properties (Baird, 2010). Despite the increased use of alternative assessments, teachers in the USA have reported that they continue to rely on 'traditional' paper and pencil tests as their primary data for determining course grades (Frey & Schmitt, 2010). Since a number of academic decisions (e.g. entrance into advanced placement courses and college admittance) are determined to a certain extent by course grades in addition to standardised test results, recommendations, etc., it is reasonable to consider the quality of teacher-made or teacher-selected tests.

A survey of 272 Canadian secondary school teachers by Leighton, Gokiert, Cor, and Heffernan (2010) found that these teachers believed that their own classroom tests, in comparison to large-scale assessments, were more informative in terms of student learning processes and were more likely to lead to student learning. Thus, these teachers relied on their own classroom assessments to guide future instruction and ascertain student grades. It therefore seems essential that teachers use well-grounded assessment methods in their classroom practice. However, studies of

teacher-made assessments suggest that there are concerns with the quality of the tests teachers create. For instance, Marso and Pigge (1991) analysed 175 teacher-made tests from elementary and high school teachers in the midwestern USA. They found that teachers were more likely to use multiple-choice, matching and short answer response items, and that most items measured low-level cognitive processes such as remembering and understanding. Billeh's (1974) study examined seventh tenth grade science teachers' tests from 18 secondary schools in Beirut, Lebanon. Teachers' instruction for one unit of science was recorded and each was asked to construct a one-hour examination of the unit taught. Three reviewers evaluated the test items in light of Bloom et al.'s taxonomy (1956) and the content taught. Findings indicated that the majority of these teachers' tests assessed the lowest cognitive levels (i.e. knowledge 72%; comprehension 20%), only 7% of the test assessed application, and there were no items assessing the highest levels of analysis, synthesis and evaluation.

McMillan, Myran, and Workman (2002) evaluated the responses of nearly 900 elementary teachers in the USA to questionnaires about the types of assessment practices they employed. Descriptive analyses of these data revealed that the three major assessments used by these teachers were projects, essays and presentations; objective assessments; and teacher-made exams. Further, teachers seemed to rely heavily on published assessment materials in addition to the assessments they created. This led these authors to conclude that teachers need training in methods for evaluating assessment materials, whether self-generated or provided by textbook publishers.

Finally, Oescher and Kirby (1990) surveyed a mixed sample of 35 rural and urban US teachers about the importance of classroom tests and the quality of these measures. Teachers in their sample reported that teacher-made tests were most important and that they were highly confident in their ability to construct assessments, despite the fact that an analysis of these same teachers' tests indicated poor validity and overall low quality in terms of directions, item construction, and cognitive levels assessed (Oescher & Kirby, 1990). Thus, teachers seem to value the tests they use in the classroom despite evidence that suggests these tests are of limited value for making sound educational decisions. Teachers' classroom tests need to provide quality information so that valid judgements about student learning, thinking, and achievement can be made and used for future instruction and placement decisions.

Reasons for poor quality of teacher-made tests

Teacher-made tests are problematic (e.g. Billeh, 1974; Broekkamp, Van Hout-Wolters, Van den Bergh, & Rijlaarsdam, 2004; Jetton & Alexander, 1997; Marso & Pigge, 1991; Oescher & Kirby, 1990). Several empirical and theoretical reasons for this poor quality have been offered. We summarise a few of these possible reasons below.

Teachers seem to be ill-prepared to construct quality classroom assessments

One can look to the curriculum of teacher education programmes and certification requirements in the USA and notice that most of these do not require future teachers to complete a course on assessment (e.g. Stiggins, 1991, 2001, 2002; Wise, Lukin, & Roos, 1991). Similarly, in Canada, Deluca and McEwen (2007) reported that only 3 of 10 bachelor's programmes in teacher education in Ontario included a required assessment course.

Examinations of preservice teacher knowledge of assessment (e.g. Maclellan, 2004) indicated that future teachers demonstrate a lack of understanding of assessment and feel ill-prepared in this area of their professional knowledge base. Maclellan (2004) performed a content analysis on 30 preservice Scottish teachers' written responses to a prompt on assessment. These participants were 30 weeks into a 36-week postgraduate certificate programme in elementary education and were about to apply for licensure. This analysis revealed some fundamental gaps in these future teachers' understandings. For instance, while they were able to articulate what the 'purpose' of assessment was, they could not recognise the relationship between the purpose of assessment and type in terms of normative vs. criterion-referenced scoring. Also consistent among these participants' responses was a lack of understanding of issues of reliability and validity with respect to classroom assessment. When participants referred to tests, they did not (1) offer criteria for evaluating such assessments, (2) describe expectations for item construction, (3) address how to effectively administer a test, or (4) consider issues of score interpretation. Thus, it seems that these preservice teachers held naive and disconnected conceptions of assessment.

Another concern regarding the preparation of preservice teachers with regard to assessment is that much of this practical preparation is left in the hands of cooperating teachers rather than in teacher education programmes. Thus, for many future teachers, their exposure to classroom assessment is guided by the practical experiences encountered during student teaching. However, cooperating teachers may not be well-versed in the conceptual basis of classroom assessment either. Black, Harrison, Hodgen, Marshall, and Serret (2010) worked with practising teachers in Great Britain and found that 'the concept of validity was not a salient feature in [teachers'] approach to pedagogy' (p. 227). In this intensive qualitative research and intervention investigation, they found that teachers initially did not think about validity as part of their teaching practice and felt hindered in doing so by the importance of required accountability testing. This study supports the concern that practising teachers may not have the knowledge needed to help student or novice teachers learn and implement foundational understandings of assessment. Thus, it is not surprising that the quality of teacher-made tests continues to be poor.

Lack of empirically supported strategies

A second salient reason for the poor quality of teacher-made tests may rest on the lack of quality strategies for test construction that teachers can employ. The vast majority of test-construction strategies offered to teachers are based on 'rules of thumb' and theory rather than empirically tested evidence. For instance, Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) performed a content analysis of 20 educational assessment textbooks written in English for guidelines to writing objective items. In their work, they briefly acknowledge that few of these rules have been tested experimentally, yet do not address the concerns related to this lack of evidence. Given that many of these recommendations were offered by Lefever (1933) over 72 years earlier, it is ill-conceived to expect teachers to follow test-writing conventions based on the way it has always been done in the light of current expectations for evidence-based practice. While Frey et al. (2005) do indicate that these item-writing rules attempt to address validity concerns related to (1) confusing or ambiguous wording, (2) student guessing, (3) test-taking efficiency and (4) testwiseness, this article underscores the need for empirical evidence to support recommended assessment strategies for teachers.

The emphasis in published strategies for classroom assessment seems to be on the use of formative assessment (Black & Wiliam, 2003; Dunn & Mulvenon, 2009; Shute, 2012), or item construction (e.g. Frey et al., 2005; Haladyna et al., 2002), with little attention given to overall assessment or test planning (e.g. Gareis & Grant, 2008; Notar et al., 2004). Moreover, when attention is given to test planning, it is often an afterthought and recommendations are made based on personal experience rather than empirical evidence. For instance, Carroll and Moody (2006) wrote:

> [b]ased on our experience, we recommend that 60–70% of the questions on an exam be based on easy, single-concept content … Twenty to thirty percent of exam questions should be designed with more difficulty … The final 5–10% is reserved for questions that involve frequent misconceptions or multi-step problem solving. (p. 6)

This type of recommendation ignores both assessment theory in terms of validity evidence and the need for empirical evidence to support this advice for generalisable use. Teachers need well-supported strategies to inform their instruction and assessment practices, and researchers need to move beyond 'rules of thumb' and personal experience and provide evidence for the strategies we recommend in preservice and practising teacher education.

Table of Specifications: a strategy for improving teacher-made tests

A TOS, sometimes called a test blueprint, is a table that helps teachers map a test onto their instructional objectives for a given segment of study (see Grondlund, 2006; Notar et al., 2004; Reynolds, Livingston, & Wilson, 2006) and is endorsed by experts in educational measurement who develop large-scale standardised tests as a tool to address validity evidence based on test content (e.g. Lissitz & Samuelsen, 2007). We argue that a TOS can be used as a planning tool intended to help teachers align objectives, instruction, and assessment. This strategy can be used for a variety of assessment methods but is most commonly associated with constructing traditional summative tests. From our perspective, the primary goal of a TOS is to improve the validity of a teacher's evaluations based on a given assessment. As discussed earlier, valid judgements are based on the quality of information teachers get from the assessments they design and give to their students. The TOS places issues of validity as the central rationale for using the strategy, and it requires teachers to consider the underlying purpose and quality 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. When constructing a test, teachers need to be concerned that the test measures an adequate sampling of the class content at the cognitive level that the material was taught. Thus, in contrast to Carroll and Moody's (2006) arbitrary allocation of difficulty levels for test items, the theory of validity recommends that tests be designed to assess student learning as a result of the instruction given. The TOS can help teachers map the amount of class time spent on each objective with the cognitive level at which each objective was taught, thereby helping teachers to identify the types of items they need to include on their tests. A sample TOS used in this study can be found in Figure 1.

The use of a TOS to guide classroom test construction currently lacks empirical support. However, in a cursory review of 13 classroom assessment textbooks, we found that eight of the texts recommended the use of a TOS for classroom test

Table of Specifications

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 Test Items to Include	Type of Item to Include
Day 1	1. Identify the Southern Colonies on a map.	5	3.0%	.3	0	Lower order MC or SA
	2. Identify who colonized Maryland and explain why people colonized Maryland	5	3.0%	.3	0	Lower order MC or SA
	3. Explain why people colonized the Carolinas and describe how Eliza Lucas Pinckney's discovery impacted the crop industry.	15	9.1%	.91	1	Lower order MC or SA
	4. Explain why people colonized Georgia.	15	9.1%	.91	1	Lower order MC or SA
Day 2	5. Predict how did people in each of the Southern Colonies made a living.	15	9.1%	.91	1	Higher order MC or SA
	6. Describe the difference between fact and opinion.	15	9.1%	.91	1	Lower order MC or SA
	7. Analyze information and determining whether it is fact or opinion.	15	9.1%	.91	1	Lower order MC or SA
Day 3	8. Apply geographic tools, including legends and symbols, to collect, analyze, and interpret data.	30	18.2%	1.82	2	Higher order MC or SA
	9. Explain the geographic factors that influenced the development of plantations in the Southern Colonies.	5	3.0%	.3	0	Lower order MC or SA
Day 4	10. Compare and contrast the life of a slave and a planter.	30	18.2%	1.82	2	Higher order MC or SA
	11. Identify the characteristics of an indentured servant.	15	9.1%	.91	1	Lower order MC or SA
		120	**100.00%**	**10**		

Figure 1. Expert TOS.

construction. This approach to test construction seems to be recommended by authors of texts on classroom assessment (e.g. Grondlund, 2006; Reynolds et al., 2006) as a 'rule of thumb' (e.g. Frey et al., 2005; Haladyna et al., 2002) rather than based on empirical investigations of the use of this strategy for classroom test construction. Given the importance of teacher-made tests, concerns about the quality of these tests, and the lack of empirically supported assessment strategies available for teachers to implement in their practice, we identified the TOS as a potential strategy that could help teachers to construct better quality tests providing the strategy itself bears empirical support. Thus, the purpose of this study was to ascertain the effectiveness of the TOS as a strategy for improving test quality based on evidence that would support claims of validity, namely alignment of TCE and RPE used during instruction with summative tests.

Research questions

We pursued the following research questions:

(1) Do participants exposed to the TOS strategy create a classroom test that allows for evaluations with greater TCE scores compared to students not exposed to any specific strategy?
(2) Do participants exposed to the TOS strategy create a classroom test that allows for evaluations with greater RPE scores compared to students not exposed to any specific strategy?

We hypothesised that participants exposed to the TOS would develop a classroom test that would allow for evaluations with greater TCE and RPE scores because they received instruction on test planning and item selection.

Methods

All students from five sections of an undergraduate educational psychology course (not taught by the researchers) were offered extra credit for their voluntary participation in this study. This course is a prerequisite for the teacher certification programme at this mid-Atlantic, four-year state university in the USA. Fifty-three students (age range from 18 to 33) agreed to participate and were randomly assigned to treatment ($n = 28$, 52.8%) and comparison groups ($n = 25$, 47.2%). The majority of the sample in both conditions was female, Caucasian, and represented a wide variety of content majors. On average, participants had completed 2.75 (SD = 2.336) semesters of university coursework and 92.5% indicated an intention to become teachers (see Table 1).

Table 1. Participant demographic descriptions.

	Treatment (%)	Control (%)
Participants	52.8	47.2
Gender		
Male	21.4	28
Female	78.6	72
Race/ethnicity		
African American	0	12
Caucasian	71.4	60
Asian American	7.1	4
Hispanic	14.3	16
Other	7.1	8
Major		
Art education	3.6	4
Biology	7.1	8
English	14.3	28
Family and child studies	14.3	0
Foreign language	7.1	0
History	7.1	20
Linguistics	3.6	4
Math	10.7	8
Music education	3.6	4
Physical education/health	7.1	20
Psychology	10.7	0
Sociology	3.6	0
Undeclared	3.6	4
No response	3.6	0
Intend to become teachers		
Yes	89.3	96
No	10.7	4

Description of sample

	Treatment		Control	
	Mean	SD	Mean	SD
Age	20.25	1.81	20.48	2.96
Number of semesters completed	2.68	2.44	2.84	2.27
Number of history courses completed	1.07	1.84	.84	1.68
Number of education courses completed	2.11	1.27	2.44	.92

Procedures

Participants were randomly assigned to either the TOS condition or the comparison condition and were given one of two envelopes (treatment materials or comparison materials). After orienting them to the overall purpose of the study and obtaining informed consent, participants in both conditions completed a pre-test to assess their knowledge of classroom assessment and the Southern Colonies (the content of the lesson materials used in the study task; see description of these measures below). Participants in the treatment condition read an article (Fives & DiDonato-Barnes, 2013), written for this study, explaining the purpose of the TOS and how to use it in test construction. Treatment participants were also provided with a partially completed TOS for the study task.

The TOS article and tool (see Figure 1 for an example) described a modified TOS that was scaled down from those used by large-scale standardised test writers. The TOS was modified by the second author, a classroom teacher of six years, to be more user friendly and practical for classroom teachers. McMillan (2003) argued that measurement processes like using a TOS either need to be abandoned or modified for pragmatic use by teachers. Typically, a TOS maps out each of the cognitive, metacognitive, affective and knowledge levels (depending on the taxonomy used, e.g. Anderson et al., 2001; Bloom et al., 1956; Marzano, 2001). In our TOS, we modified Bloom et al.'s (1956) taxonomy of the cognitive domain to two levels of cognitive processing: low-level processing (knowledge and understanding) and higher levels of processing (application, analysis, synthesis and evaluation; Kastberg, 2003). Using this broader classification ameliorates the philosophical criticisms about the hierarchical nature of the taxonomy and the distinction among the categories (Kastberg, 2003). Furthermore, this framework may help teachers to organise and clarify objectives, allowing them to plan better instruction and assessments, and to align instructions, assessments and objectives (Anderson et al., 2001). Finally, we chose to use Bloom's taxonomy as a framework in this study because it is the most commonly taught taxonomy in the teacher education programme that participants in this study would apply for; this increased the benefits for our study participants.

The study task asked participants in both conditions to review the materials of a fictitious fifth-grade unit on the Southern Colonies of what is now the USA and select items for an end-of-unit test. The Southern Colonies refer to the British territories of South Carolina, North Carolina, Maryland, Virginia, and Georgia during the sixteenth and seventeenth centuries. This is common content for fifth-grade students in the USA and is typically taught during the first half of the fifth-grade academic year. The study task materials included learning objectives reflecting the range of topics examined in the unit and cognitive levels.

For each of the 11 objectives in the unit, the test bank included four items. Two of these items (one multiple choice and one short answer) measured higher thinking skills (i.e. application, analysis, evaluation and synthesis) and two items (one multiple choice and one short answer) measured lower-level thinking skills (i.e. knowledge and comprehension). See Figure 2 for examples. The test bank included 44 items grouped by item type (all multiple choice and then all short answer items). Within each grouping, the items were sorted randomly so that they did not directly align with the order of the objectives or unit materials. For each of the 44 items, participants indicated (1) if they would include the item on the test (yes/no) and (2) the reason for their decision.

Objective 2: Identify who colonized Maryland and explain why people colonized Maryland. (**Low level objective**)	
Low level multiple choice	Maryland was settled as a/an a. area to grow rice and cotton. b. safe place for English debtors. c. colony for indentured servants. d. refuge for Roman Catholics.
High level multiple choice	Which of the following people would **most** want to settle in Maryland? a. A Catholic from southern England. b. A debtor from an English Prison. c. A tobacco planter. d. A French trapper.
Low level short answer	State one reason why people colonized Maryland.
High level short answer	Use a Venn Diagram to compare and contrast the reasons people colonized Maryland and Georgia.

Figure 2. Sample test bank items for objective 2.

It should be noted that the item type (multiple choice or short answer) had no bearing from our perspective on the quality of test created, however through personal experiences with students in the courses taught by the first two authors, we had a strong suspicion that participants might assume that open-ended items were more cognitively demanding than multiple-choice items. For this reason, we wanted to ensure that participants had the option to select items written at both high and low levels in each format.

After completing the study task, participants completed the same demonstrated knowledge tests received as pre-tests. Upon completion of the study, participants returned the completed materials to the researcher, who provided the student with a proof of participation letter.

Measures

Demographics

Participants completed a demographic questionnaire that elicited information about their sex, ethnicity, age, semester in university, academic major, intention to become a teacher (and if so, what level and content area), and the number of education courses and history courses the student had completed.

Demonstrated knowledge

We assessed participants' knowledge of assessment and the Southern Colonies at pre- and post-test using identical measures that followed a similar format (i.e. 'Please jot down words, phrases, and sentences that tell me what you know about each of the following statements'). To assess knowledge of assessment, we provided

three prompts: (1) Bloom's taxonomy, (2) Validity in classroom assessments, and (3) TOS. We also used three prompts to assess participants' knowledge of the Southern Colonies: (1) Plantation Life, (2) Reasons for settling in the Southern Colonies, and (3) Resource and Product Map. Scores were determined based on the accuracy and elaboration of ideas offered using a rubric modified from one used by Alexander, Fives, Buehl, and Mulhern (2002). The four-point rubric is detailed in Table 2 with sample responses. Two members of the research team scored 20% of the data for each test and an inter-rater reliability score of .996 was found for the knowledge of assessment test and .977 was found for the knowledge of Southern Colonies test. After discussion, agreement was reached on all codes.

Table 2. Scoring rubric and sample participant responses to open-ended knowledge items.

Score	Number of correct ideas	Incorrect ideas	Sample responses	
			Assessment: 1. Bloom's taxonomy	Southern Colonies: 2. Reasons for settling in the Colonial South
0	0	Present/ no response	*Tests. Validity.* (ID: 4_10_2_6)	*Industrialisation*. (ID: 4_5_2_1)
1	1–3	Present	In terms of questioning, has *eight levels* that relate to cognitive ability required to answer, with less ability needed for lower levels and higher ability needed for high levels. (ID: 4_11_1_9)	You could make a lot of money owning a plantation and farming. *More natural resources than the north.* More land to be settled on. (ID: 4_10_1_1)
2	1–3	Not present	Levels of thinking (ID: 4_11_1_4)	Warmer weather provided perfect conditions for crops, especially tobacco. Some colonists wanted religious freedom as well. (ID: 4_10_2_7)
3	4+	Present	*Bloom's taxonomy was created for the teachers to have a basis on how to create questions for students.* From simple questions such as 'What color is Goldilocks' Hair?' to more complex and analytical questions, as to 'Explain what is the moral of Goldilocks?' (ID: 4_11_1_8)	Tobacco, *slaves*, new beginning, freedom for former indentured servants, *dye*, resources. (ID: 4_19_1_1)
4	4+	Not present	There are six levels of learning. They can be divided into two groups, the higher and lower levels. The higher level = more analysing and explanation. The lower level = recalling information. (ID: 4_10_2_7)	Some settled for crop cultivation, others for religious freedom, others still for escaping debts from England as indentured servants. (ID: 4_3_1_1)

Notes: Underline indicates each correct idea unit counted and italics indicate each incorrect idea identified. Irrelevant information or restatement of the prompt was ignored.

Test content evidence (TCE) and response process evidence (RPE) scores

The scores used to assess TCE and RPE were determined by examining the items participants selected from the test bank for inclusion on the end-of-unit test, hereafter referred to as the participants' tests. An expert TOS (Figure 1) for the unit was constructed in order to evaluate each item selected for inclusion on a test. This expert TOS was developed following the model presented in the article read by the treatment group (Fives & DiDonato-Barnes, 2013).

Participants were assigned a TCE score by awarding points for selecting items that accurately reflected the subject matter that should be included on the exam in comparison to the expert TOS (e.g. column F of Figure 1, maximum 10 points). Participants received zero points if they selected items that assessed objectives that should not have been included on the test (objectives 1, 2 and 9). For items selected that mapped onto objectives that should be assessed, they received one point for each correct selection (regardless of the cognitive level of the items). However, if they selected two items for an objective that should have had only one item selected (objectives 3 7 and 10), they only received one point, therefore forgoing the opportunity to earn that point on another objective. There were two objectives for which participants should have selected two items (objectives 8 and 10). For these two objectives, participants had the potential to earn two points, one point for each correct selection.

To assign a RPE score, we examined whether each item selected accurately reflected the cognitive level of the objective related to the test item (i.e. column G, Figure 1). Essentially, this score indicates if participants were able to select items at the correct cognitive level for each objective regardless of whether the objective should have been assessed on the end-of-unit test. For each test item selected at the accurate cognitive level, one point was assigned (10 points maximum). For example, if a participant selected two low-level test items that assessed objective 1, he/she would receive a score of two, and if a participant selected two high-level items to assess objective 1, he/she would receive a score of zero because the high-level items would lack RPE.

Data analysis and findings

Quantitative data analysis using SPSS was used to determine whether students who were exposed to the TOS strategy constructed tests with statistically significantly greater TCE and RPE scores compared to students in the comparison group. Descriptive statistics including means and standard deviations (Table 3) as well as a correlational analysis (Table 4) were used to initially describe the data and identify covariates to be included in subsequent analysis. A one-way analysis of variance was used to determine if both groups were equivalent in terms of pre-test scores. The analysis indicated that both groups were equivalent on pre-test knowledge of assessment (F (1, 51) = .265, p = .609), however the comparison group had greater knowledge of the Southern Colonies compared to the treatment group (F (1, 51) = 4.881, p = .032). A correlational analysis indicated that none of the remaining variables were significantly correlated with the dependent variables; therefore, a one-way multivariate analysis of covariance (MANCOVA) was used to determine whether the treatment condition had an effect on the dependent variables after removing the variance associated with knowledge of the Southern Colonies. MANCOVA was chosen as the primary statistical analysis in order to test for significant

Table 3. Means and standard deviations for the pre-test variables.

	Condition	*M*	SD	*N*
Pre-test assessment	Treatment	1.21	1.548	28
	Control	1.00	1.472	25
Pre-test Southern Colonies	Treatment	3.29	2.052	28
	Control	4.44	1.710	25
TOS: TCE score	Treatment	6.86	1.380	28
	Control	5.84	1.028	25
TOS: RPE score	Treatment	5.79	1.258	28
	Control	5.53	1.358	25

differences between group means when there are several dependent variables and it is necessary to control for a covariate(s) in a single experiment without inflating Type I error. A slight correlation between the dependent variables supported the use of MANCOVA over independent univariate tests (Grice & Iwasaki, 2007).

MANCOVA was performed to determine the effect of using the TOS strategy (experimental and comparison) on two dependent variables (TCE and RPE scores) while controlling for the covariate. Box's *M* and Levene's Test of Equality of Error Variances were non-significant, suggesting that the homogeneity of variance for both dependent variables and homogeneity of variance-covariance matrix assumption were not violated. Results of the MANCOVA analysis indicated significant differences between students in the experimental and comparison groups overall (Wilks' = .862, F (2, 49) = 3.92, p = .026 and d = .80), however there were no group differences on the covariate (Wilks' = .983, F (2, 50) = .420, p = .659 and d = .26). Power to detect the effects were .680 and .114, respectively. Univariate analysis of variance (ANOVA) tests were conducted controlling for Type I error by evaluating significance at the .025 level. The univariate main effect suggests group differences in TCE scores (F (1, 50) = 7.724, p = .008 and d = .79), but not RPE scores (F (1, 50) = .910, p = .345 and d = .27). Power to detect the effects were .778 and .155, respectively. When controlling the effect for pre-test knowledge of the Southern Colonies, we found no significant effect of the covariate.

In order to better understand the lack of significant difference between comparison and treatment groups on the RPE dependent variable, we performed a *post hoc* analysis on the data from the treatment group. Participants in the treatment condition were asked to complete a partially finished TOS (referred to as the TOS tool) that included the objectives, time spent on topic and percentage of time spent on topic (Columns A D on Figure 1) before engaging in item selection. Treatment participants needed to determine and record (1) the number of test items to include for each objective; (2) whether each objective should be assessed with a low or high level test item; and (3) the type of item they would use to assess the objective (i.e. multiple choice or short answer).

An examination of participants' TOS tool was conducted to investigate (1) how correctly they completed the TOS tool (an issue of accuracy); and (2) the extent to which they used the tool to guide test item selection (an issue of alignment). To successfully complete the TOS tool, participants needed to accurately indicate on the TOS tool (1) the correct number of items for each objective (i.e. *accuracy number*,

Table 4. Correlations for demographic variables, pre-test measures and dependent variables.

	Number of semesters completed	Number of history courses completed	Number of education courses completed	Pre-test assessment	Pre-test colonies	TOS: TCE score	TOS: RPE score
Number of semesters completed	1						
Number of history courses completed	.252	1					
Number of education courses completed	.309□	.008	1				
Pre-test assessment	.079	□.009	.175	1			
Pre-test colonies	□.051	□.064	.133	.403□□	1		
TOS: TCE score	.012	□.184	□.042	□.003	□.145	1	
TOS: RPE score	□.060	□.035	□.252	□.158	.082	.177	1

□$p < .05$; □□$p = .00$.

$M = 7.14$, maximum score = 10) which provides an indication of accurate identification of TCE; and (2) the accurate cognitive level corresponding to the level at which the objective was written for each objective selected (i.e. *accuracy level*, $M = 6.29$; maximum score = 10) which provides an indication of accurate identification of RPE. A *t*-test suggested that participants were equally as skilled at correctly identifying the number of test items on the TOS tool as they were at accurately selecting the correct cognitive level to measure each objective ($t(27) = 1.4$ and $p = .173$). This suggests that participants were able to *complete the tool* with similar accuracy for decisions related to both TCE and RPE.

To determine how well students *used* the TOS tool to guide item selection, two alignment scores were calculated. That is, we compared the completed TOS tools to the actual test items selected to see if participants selected items from the test bank that were reflective of the test they designed on the TOS tool. To assess test content alignment, we awarded one point for each test item selected from the test bank that was reflective of an objective selected on the TOS tool (i.e. *alignment test content*, $M = 7.68$, maximum score = 10). The response process alignment score was calculated in a similar way; we awarded one point for each test item selected from the test bank that reflected an objective at the cognitive level indicated on the TOS tool (i.e. *alignment response process*, $M = 5.39$, maximum score = 10). A *t*-test suggested that participants were better at choosing items that reflected the subject matter in the selected objectives than they were at choosing items at the cognitive level they specified in the TOS tool ($t(27) = 8.6$ and $p = .000$). Therefore, although participants were able to accurately classify the cognitive level of objectives in the TOS tool, they were not skilled at choosing items from the test bank at the identified cognitive levels.

Discussion

We found significant differences in the quality of TCE scores (with the treatment group scoring higher than the comparison group) but no significant differences between groups on RPE scores after controlling for knowledge of the Southern Colonies. This finding provides empirical support that the TOS can help teachers choose test items that adequately assess the subject matter that was taught. This provides justification for the use of the TOS beyond 'rule of thumb' suggestions and offers preliminary evidence as to how this tool can improve the quality of tests teachers construct.

In order to better understand the lack of significant difference between comparison and treatment groups on the RPE dependent variable, we performed a *post hoc* analysis on the data from the treatment group. Results suggested that on average treatment participants were able to accurately complete the TOS tool and choose items that reflected the subject matter specified in the TOS tool. However, they experienced difficulty *selecting items* at the cognitive level specified in the TOS tool. Because this is a micro-process related to RPE, this difficulty might explain why the treatment group did no better than the comparison group on the RPE score. At this point, it is difficult to determine whether students were unable to correctly select items at the identified cognitive level or if they did not use the TOS tool with fidelity to guide item selection. For the former, this means they thought they were selecting the accurate level item, but did not. And for the latter, this means that they ignored the indicated cognitive level in favour of some other rationale for choosing the cognitive level.

There is some evidence to suggest that it is quite difficult for even practising teachers to select items at differing cognitive levels. Carter (1984) provided 310 teachers from the south central USA with 10 multiple-choice items measuring varying cognitive levels, and asked them to classify each item. Results suggested that participants were only 50% likely to correctly identify items measuring low-level cognitive skills and less than 30% likely to correctly identify items measuring higher-level cognitive skills. When asked to write objectives at differing levels, teachers struggled with interpreting the skill and writing objectives measuring higher-level cognitive skills. Thus, instruction aimed at helping preservice and practising teachers identify and write test items at varying levels may be needed prior to or in conjunction with introducing the TOS tool, especially since participants were able to correctly classify objectives by cognitive level in the TOS tool.

Another possibility is that participants did not use the TOS tool to guide item selection with much fidelity, and instead relied on their prior beliefs related to the nature and purpose of assessment to direct their decision-making. Bonner and Chen (2009) developed the Survey of Assessment Beliefs in order to measure preservice teachers' assessment-related beliefs before and after completing coursework in classroom assessment. Results suggested that students (from an urban university in the Northeast USA) reported only small changes in their beliefs from pre- to post-test. Consistent with these findings, McMillan and Nash (2000) found from conducting semi-structured interviews with 28 US Mathematics and English teachers from 12 schools that personal beliefs and values were the most significant influence on their assessment-related decision-making. Thus, it is possible that participants made decisions about items using other criteria and ignored the plan set forth in the TOS tool. Perhaps, participants chose items based on the perceived easiness or difficulty of the item regardless of the cognitive level indicated in the objective. Thus, in addition to any instruction on item interpretation, future and current teachers also need to develop a sound set of beliefs about the goals and purposes of educational assessment.

Limitations and future research

We have identified two limitations of this study. The first is related to the depth and duration of the treatment and the second is our sample size. First, in our treatment condition, students independently read a short article on how to complete a TOS with only two examples of items at different cognitive levels. Ideally, students would receive in-depth instruction about the use of the TOS, cognitive taxonomies, and would work with a variety of examples in order to hone these skills. Because the breakdown in our hypothesis is related to identifying accurate cognitive levels of items, it seems evident that a stronger intervention of the identification of items at differing cognitive levels is warranted. Future research will be designed to provide instruction on cognitive taxonomies and practice identifying items at varying cognitive levels as part of the intervention.

Second, the sample size was limited, which raises issues of generalisability to the overall population. Future research should attempt to replicate these findings using a larger sample. Without explicit instruction on assessment and opportunities to evaluate their own practices, teachers may assume greater confidence in their assessment practices than is warranted. Still, with this limited instructional intervention and small sample, significant differences were found.

Significance and implications

Results of this study should prove informative for many groups including educational researchers, teacher educators, school leaders and practising teachers. Researchers and teacher educators need more information about the usefulness of recommended strategies for implementation by classroom teachers. Similarly, school leaders and classroom teachers should find value in the results of this study for direct use in their professional settings. The TOS technique may seem time-consuming but may be worth that time if it provides better assessment and instructional planning. Given the current educational climate, such evidenced-based strategies may be particularly useful in helping teachers improve their practices and meet the needs of their students.

Notes on contributors

Nicole DiDonato-Barnes is an assistant professor and educational psychologist in the Department of Educational Foundations at Montclair State University, NJ, USA. Her research interests are in the areas of assessment practices and self-regulated learning. She has recently published work on students' self and co-regulated learning on an authentic performance assessment in Instructional Science.

Helenrose Fives is an associate professor and educational psychologist in the department of Educational Foundations at Montclair State University, NJ, USA. Her research interests are in the areas of teachers' beliefs, knowledge, and practice, and classroom assessment. She has recently published work on teachers' beliefs in the *Handbook of Educational Psychology*.

Emily S. Krause is a master's student in the Community Counseling Program in the Department of Counseling and Educational Leadership Foundations at Montclair State University, NJ, USA. She is serving as an intern at Hudson River Care and Counseling in NJ and was recently accepted to Montclair State University's Advanced Counseling Certificate Program for the fall of 2013.

References

Alexander, P. A., Fives, H., Buehl, M. M., & Mulhern, J. (2002). Teaching as persuasion. *Teaching and Teacher Education, 18*, 795–813. doi:10.1016/S0742-051X(02)00044-6

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: National Council on Measurement in Education.

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.

Baird, J. (2010). Beliefs and practices in teacher assessment. *Assessment in Education: Principles, Policies, and Practice, 17*, 1–5. doi:10.1080/09695940903562682

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.

Billeh, V. H. (1974). An analysis of teacher-made science test items in light of the taxonomic objectives of education. *Science Education, 58*, 313–319. doi:10.1002/sce.3730580305

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*, 215–232. doi:10.1080/09695941003696016

Black, P., & Wiliam, D. (2003). In praise of educational research: Formative assessment. *British Educational Research Journal, 29*, 623–637. doi:10.1080/0141192032000133721

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York, NY: David McKay.

Bonner, S., & Chen, P. (2009). Teacher candidates' perceptions about grading and constructivist teaching. *Educational Assessment, 14*, 57–77. doi:10.1080/10627190903039411

Broekkamp, H., Van Hout-Wolters, B. H. A. M., Van den Bergh, H., & Rijlaarsdam, G. (2004). Teachers' task demands, students' test expectations, and actual test content. *British Journal of Educational Psychology, 74*, 205–220. doi:10.1348/000709904773839842

Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practices, 18*, 5–13. doi:10.1111/j.1745-3992.1999.tb00002.x

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practices, 30*, 3–12. doi:10.1111/j.1745-3992.1999.tb00002.x

Carroll, T., & Moody, L. (2006, September). Teacher-made tests. *Science Scope, 30*, 66–67.

Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education, 35*, 57–60. doi:10.1177/002248718403500613

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*, 265–285.

DeLuca, C., & McEwen, L. (2007, April). *Evaluating assessment curriculum in teacher education programs: An evaluation process paper*. Paper presented at the annual Edward F. Kelly Evaluation Conference, University of Ottawa, Ottawa, ON.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14,* 1–11. Retrieved from http://pareonline.net/getvn.asp?v=14&n=7

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research, and Evaluation, 18,* 1–7. Retrieved from http://pareonline.net/getvn.asp?v=18&n=3

Frey, B., & Schmitt, V. (2010). Teachers' classroom assessment practices. *Middle Grades Research Journal, 5*, 107–117. doi:10.1108/S2048-0458(2012)0000001010

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*, 357–364. doi:10.1016/j.tate.2005.01.008

Gareis, C. R., & Grant, L. W. (2008). *Teacher-made assessments: How to connect curriculum, instruction, and student learning*. Larchmont, NY: Eye on Education.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181–191.

Grice, J. W., & Iwasaki, M. (2007). A truly multivariate approach to MANOVA. *Applied Multivariate Research, 12*, 199–226.

Grondlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Pearson.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–334. doi:10.1207/S15324818AME1503_5

Jetton, T. L., & Alexander, P. A. (1997). Instructional importance: What teachers value and what students learn. *Reading Research Quarterly, 32*, 290–309.

Kastberg, S. E. (2003). Using Bloom's taxonomy as a framework for classroom assessment. *The Mathematics Teacher, 96*, 402–405.

Lefever, D. W. (1933). Dangers and values in teacher-made tests. *Education, 53*, 409–413.

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher views about the cognitive diagnostic merits of classroom versus large-scale assessments: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice, 17*, 7–21. doi:10.1080/09695940903565362

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448. doi:10.3102/0013189X07311286.

Maclellan, E. (2004). Initial knowledge states about assessment: Novice teachers' conceptualizations. *Teaching and Teacher Education, 20*, 525–535. doi:10.1016/j.tate.2004.04.008

Marso, R. N., & Pigge, F. L. (1991). An analysis of teacher-made tests: Item types, cognitive demands, and item construction errors. *Contemporary Educational Psychology, 16*, 279–286. doi:10.1016/0361-476X(91)90027-I

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives*. Thousand Oaks, CA: Corwin Press.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*, 34–43. doi:10.1111/j.1745-3992.2003.tb00142.x

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*, 203–213. doi:10.1080/00220670209596593

McMillan, J. H. & Nash, S. (2000, April 25–27). *Teachers' classroom assessment and grading decision making*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 31*, 49–64.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Old Tappan, NJ: Macmillan.

Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher-made tests. *Journal of Instructional Psychology, 31*, 115–129.

Oescher, J. & Kirby, P. C. (1990, April 17–19). *Assessing teacher-made tests in secondary math and science classrooms*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA. ERIC Document Reproduction Service No. 322 169.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*, 10–12, 39. doi:10.1111/j.1745-3992.1993.tb00548.x

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). *Measurement and assessment in education*. Boston, MA: Pearson.

Shute, V. J. (2012). Focus on formative feedback. *Review of Educational Research, 82*, 153–189. doi:10.3102/0034654307313795

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*, 7–12. doi:10.1111/j.1745-3992.1991.tb00171.x

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*, 23–27. doi:10.1111/j.1745-3992.1999.tb00004.x

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice, 20*, 5–15. doi:10.1111/j.1745-3992.2001.tb00065.x

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*, 78–83.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*, 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*, 117–132. doi:10.1080/09695941003693856