ITEM & TEST DESIGN CONSIDERING INSTRUCTIONAL SENSITIVITY

Michael C. Rodriguez
University of Minnesota
April, 2017

Paper presented at the annual meeting of the
National Council of Measurement in Education,
San Antonio, TX.

Abstract

Understanding and investigating opportunity to learn (OTL) and the related notion of instructional sensitivity (IS) have long been hallmarks of test score interpretation and use, and as some argue, keys to effective education (Scheerens, 2017). These are core validity concerns, as they affect interpretations regarding achievement at the individual and group levels. OTL is considered to be the basic assumption underlying most achievement tests. This clearly has item-writing and test-design implications. In a general sense, many education researchers have argued about the important connections between instruction and testing (Airasian & Madaus, 1983; Baker et al., 2001; Ebel, 1951; Haladyna & Rodriguez, 2013; Monroe & Clark, 1924). For decades, measurement specialists have attended to the challenges of item analysis to support item selection and test design for criterion-referenced tests intended to have IS (Haladyna, 1976; Popham, 1969). Researchers have examined multiple approaches to detecting or estimating IS (Baker, 2008; Chen & Kingston, 2012; Li, Ruiz-Primo, & Wills, 2012; Muthén, Kao, & Burstein, 1991; Polikoff, 2010; Popham & Ryan, 2012; Wiliam, 2007). Others have examined OTL and IS and their implications for score interpretation (Burstein, 1989; D'Agostino, Welsh, & Corson, 2007; Ing, 2008; Popham et al., 2014; Yoon & Resnick, 1998). Finally, a few have also attempted to develop a set of principles regarding the role of IS in assessment design (Baker, 2001, 2002; Popham, 2001; Ruiz-Primo et al., 2013; Way, 2014).

Here, the literature examining IS with guidance around item development and test design is systematically reviewed. Design principles resulting directly from these connections are organized for the practitioner. This review resulted in a *not-perfectly-consistent* set of guidelines for item writing and test design to maximize IS, to the extent it may be relevant to intended score interpretations and use.

WHY EVEN QUESTION THE INSTRUCTIONAL SENSITIVITY OF TESTS?

By the early 1980s, Airasian and Madaus (1983) noted that standardized tests were beginning to be used for many purposes, moving further away from "a truism" that achievement tests must be linked to instruction, including purposes such as measuring educational equity, evaluating school and program effectiveness, making compensatory funds allocation decisions, measuring teacher effectiveness, accreditation, and placing students in special programs or certifying grade-advancement and high school graduation. Airasian and Madaus took on a couple of specific uses and posed arguments regarding their implications. For example, the use of total achievement scores for evaluating school or program effectiveness is poor practice, as they argued that schools and programs are more likely to differ on measures of specific objectives, rather than total scores that obscure variation in specific objectives. They argued that tests are simply not designed to differentiate school or program-level effectiveness. In addition, they reviewed a number of court cases involving decisions based on test scores, where the courts based part of their decisions on notions related to content validity, curricular validity, and instructional validity – essentially evidence regarding the extent to which the test includes relevant content, content present in the school's/program's curriculum, and content delivered during instruction.

Baker et al. (2001) addressed the question: Why can't state tests be designed to inform instruction to improve student achievement? The report they created from their work on the *Commission on Instructionally Supportive Assessment* contained recommendations for state policymakers to support the design of tests that inform both instruction and accountability. Three relevant requirements include:

- A state must monitor the breadth of the curriculum to ensure that instructional attention is given to all content standards and subject areas, including those that are not assessed by state tests (p. 21).
- A state must provide educators with optional classroom assessment procedures that can measure students' progress in attaining content standards not assessed by state tests (p. 19).
- A state must ensure that educators receive professional development focused on how to optimize children's learning based on the results of instructionally supportive assessments (p. 25).

Under the argument that teaching results in learning (perhaps in many ways), testing is employed in educational accountability systems to "detect differential effectiveness of students' instructional experiences" (Baker, 2008, p. 2). In that light, she took a deep look at the initial development of criterion-referenced tests (CRTs) and the promises of the ongoing development of integrated instructional systems. She recalled the starting point of design for CRTs being the "well-specified domain of content and intellectual (now cognitively-oriented) skills" (p. 2). However her descriptions went far beyond the conceptualization of domains which in the norm-referenced testing (NRT) framework mapped item development to a blueprint that specified a wide-range of topics, often lightly sampled. She argued that CRT items and tasks must more deeply engaging the well-defined content domain and skill boundaries – acknowledging that item uniformity across content and cognitive task (in effect, uniformity of item difficulty) doesn't exist. Difficulty varies across item types and formats, context, conceptual difficulty, and the naturally varying degrees of complexity across the ability levels. Oddly, that simple structure attributed to NRTs now somewhat accurately applies to current so-called standards-based CRTs, where items are lightly sampled across a wide range of topics in the academic standards.

But Baker (2008) also argued that the promise of CRTs was grounded in the simultaneous promise of employing the same domain boundaries to support and directly inform instruction. Just as systematic instructional systems were then expected to directly uncover and address gaps in knowledge and skills among students, CRT performance would directly mirror the gains produced through one or two cycles of revision of instruction. To Baker, the key to CRTs was our ability to define the domain of knowledge, skills, and abilities. And although the results of CRTs were largely reported in terms of percent correct, the process of standard setting soon captured CRT developers, with the designation of performance levels describing what students know and can do. She then detailed how the standards-based accountability testing movement has degraded the promises of CRT into standardized procedures that vastly reduce the test's role in instructional systems, not to mention the direct degradation of instruction through the common finding of narrowing curricula – focusing greater attention to what is tested, rather than testing what should be instructed.

The Board of Testing and Assessment (2009) of the National Academies submitted a letter to the US Department of Education regarding the regulations for the Race to the Top Initiative, to provide guidance regarding the role of measurement and testing. In that letter, they made a number of explicit claims regarding the limits of large-scale testing, specifically addressing state school accountability testing:

> A test score is an estimate rather than an exact measure of what a person knows and can do. The items on any test are a sample from some larger universe of knowledge and skills, and scores for individual students are affected by the particular questions included. A student may have done better or worse on a different sample of questions. In addition, guessing, motivation, momentary distractions, and other factors introduce uncertainty into individual scores. (p. 3)

> The choice of appropriate assessments for use in instructional improvement systems is critical. Because of the extensive focus on large-scale, high-stakes, summative tests, policymakers and educators sometimes mistakenly believe that such tests are appropriate to use to provide rapid feedback to guide instruction. This is not the case.
>
> Tests that mimic the structure of large-scale, high-stakes, summative tests, which lightly sample broad domains of content taught over an extended period of time, are unlikely to provide the kind of fine-grained, diagnostic information that teachers need to guide their day-to-day instructional decisions. (p. 10)

> …BOTA urges the Department to clarify that assessments that simply reproduce the formats of large-scale, highstakes, summative tests are not sufficient for instructional improvement systems. (p. 11)

The critiques leveled against large-scale tests regarding their potential instructional sensitivity are weighty.

DEFINING INSTRUCTIONAL SENSITIVITY (IS)

There are many researchers addressing the many issues and conceptualizations of IS. A sampling of the definitions and conceptualizations of IS include the following (ordered here chronologically).

- IS is defined as the tendency for an item to vary in difficulty as a function of instruction (Haladyna & Roid, 1981, p. 40).
- A primary purpose for instructionally sensitive assessments is to reflect student knowledge/ability as the consequence of instruction (Burstein, 1989, p. 7).
- A test is sensitive to instruction when instruction changes a student's score. If instruction does not change students' score on a test very much, then that test is insensitive to instruction (Wiliam, 2007, p. 5).
- A test's IS represents the degree to which students' performances on that test accurately reflect the quality of the instruction that was provided specifically to promote students' mastery of whatever is being assessed (Popham, 2007, p. 146; 2008).
- Among students with similar ability levels, if students with one set of instructional opportunities perform better on an item than students with a different set of instructional opportunities, the item is considered to be sensitive to instruction (Ing, 2008, p. 25).
- To qualify as instructionally sensitive, assessments should provide evidence that students' scores on the assessments yield valid information about: (1) whether students had the opportunity to learn certain curricular content; (2) the quality of instruction students received; and (3) which aspects of the curricular content may require further attention (Ruiz-Primo et al., 2012, p. 692).
- Assessment tasks should also be instructionally sensitive and educationally useful. That is, they should 1) represent the curriculum content in ways that respond to instruction, and 2) have value for guiding and informing teaching (Darling-Hammond, et al., 2013, p. 11).

The Range of Contexts of IS

One of the earliest texts for teachers on educational measurement was Tiegs' (1931) *Tests and Measurements for Teachers*. "The principal function of measurement is to contribute directly or indirectly to the effectiveness of teaching and learning" (p. 3). He continued with a discussion about learning that inextricably ties assessment to teaching:

> Learning does not always parallel teaching; in fact, at many points and in many different ways, there are learning difficulties. Particular measurement devices which will reveal the exact location and the nature of these difficulties will aid the teacher in directing further learning. (p. 11)

Classroom Assessment. IS naturally concerns classroom assessment, although rarely questioned in that context. "Classroom teachers are the ultimate purveyors of applied measurement, and they rely on measurement and assessment-based processes to help them make decision every hour of every school day" (Airasian & Jones, 1993, pp. 241-242). When a teacher can tailor assessment activities to reflect and embody aspects of instruction, this connection between instruction and assessment is inherent and natural. Classroom assessments are not only "one of our indicators of educational outcomes, but these classroom assessments also are part of the very instructional treatments that produce the desired outcomes" (Stiggins & Conklin, 1992, p. 2). From that assertion, it's a wonder why we aren't talking about assessment sensitive instruction.

But once we move beyond a single classroom and the specific approach to instruction, questions regarding IS become more challenging, and these challenges are complicated when embedded in accountability. IS has been a focus of attention in large-scale testing, particularly state-level standards based testing that have been used to make decisions about school and teacher quality. It has also been a long-time concern to professionals addressing students with unique learning needs through special education and intervention models, including the more recent response to intervention and multi-tiered systems of supports.

Response to Intervention. In the context of response to intervention (RTI), the key activities of universal screening, identification or placement in a multi-tiered system of supports (MTSS), monitoring progress, using data to make instructional changes, and meeting individual student needs across the curriculum, depends on instructionally relevant data (Ysseldyke & Burns, 2010). In the context of RTI, the purpose of assessment is multi-phased, including identification of the need for intervention and progress monitoring. The purpose of assessment must be clearly articulated. Assessment must then be instructionally relevant, precise, sensitive to change, and frequent. "Using assessment to guide instructional decision making is the only means to assure individualized instruction to meet the unique learning needs of a child with, or without, a disability" (Ysseldyke & Burns, p. 60).

Statewide School Accountability Testing. School accountability testing presents a topic of national debate, where the intended inferences lead from student test performance to school quality. Inherent in the discussion of school quality is instructional quality via state (or common core) standards. Way (2008) described the many approaches employed in the PARCC assessments that support IS. Primarily, the ECD approach promotes the articulation of student behaviors and the claims about what students know and can do. A great deal of effort was devoted to the design of the content and cognitive complexity frameworks that were employed in the development of items and tasks, including detailed assessment blueprints and evidence statement tables. In addition, items were explicitly written to tap a full range of cognitive skills and represent the kind of instruction that might be employed with low-performing students.

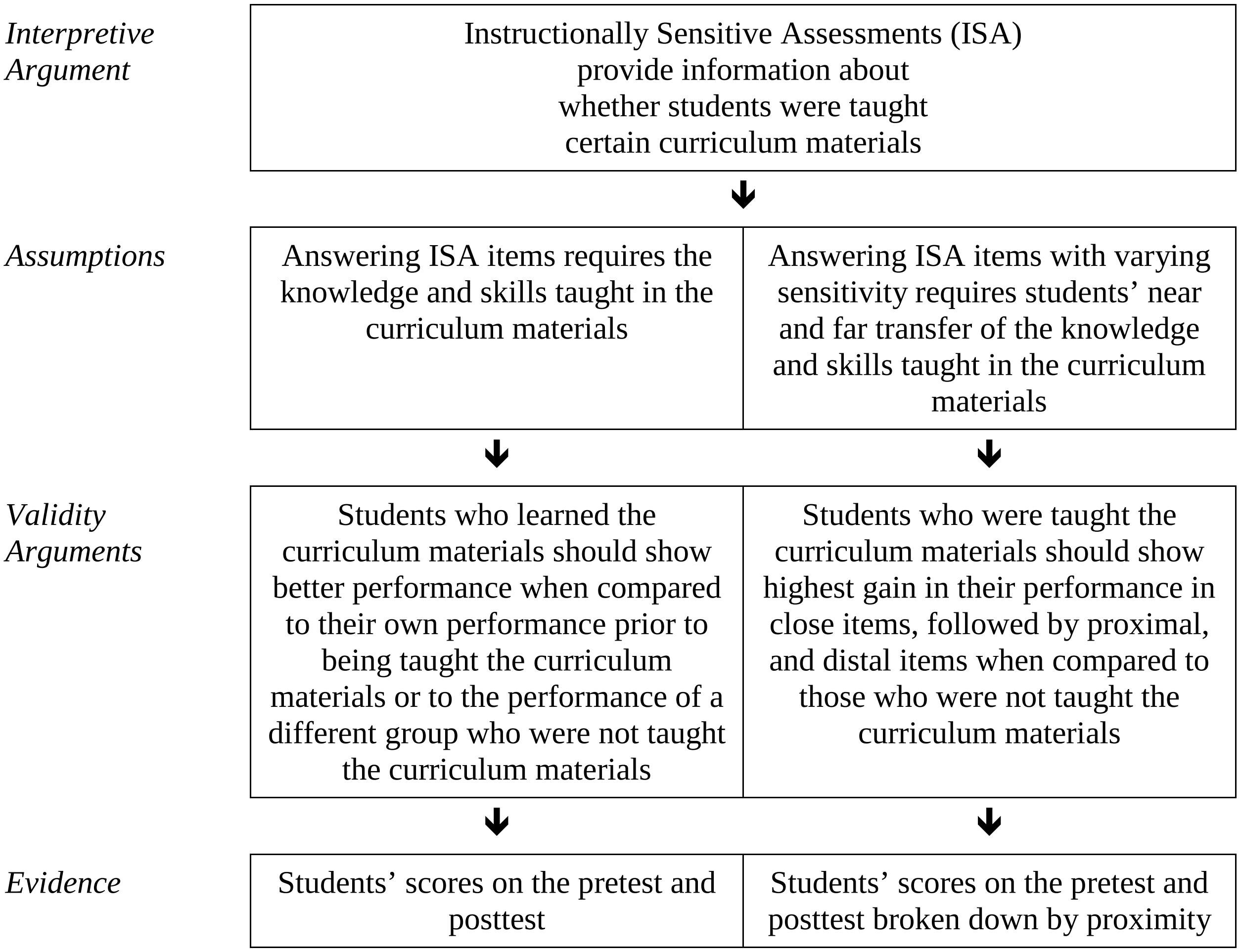
"If we assume that one of the major purposes of assessments is to make inferences about the variables of the learning environment based on the status and progress of student achievement, then an essential component of validating assessments should be instructional sensitivity. Assessments that do not take instructional sensitivity into account cannot adequately monitor the quality of instruction students receive, nor can they adequately evaluate the effectiveness of educational reforms" (Ruiz-Primo et al., 2012, p. 706).

General Guidance on High Quality Assessment Pertaining to IS. Darling-Hammond, Herman, Pellegrino, et al. (2013), helped define a small number of characteristics of high quality assessments. In those, they identified the following: (a) measures higher-order cognitive skills,

(b) represents high-fidelity assessment of critical abilities, (c) employs internationally-benchmarked standards, (d) includes instructionally sensitive items that are educationally valuable, and (e) results in evidence regarding validity, reliability, and fairness. Their focus on instructionally sensitive and educationally valuable items and tasks includes the representation of the curriculum in instructionally responsive ways that can guide and inform teaching. They argued that instructionally sensitive items should be designed in way that allows the underlying concepts to be taught and learned, rather than reflect test-taking skills or socioeconomic or cultural characteristics of students. This guidance was high-level and did not include explicit instruction on item development or design.

The CCSSO (2014) criteria for procuring and evaluating high quality assessments includes a criteria for “ensuring that assessments are valid for required and intended purposes” (p. 4). Among the supporting evidence, the guidance recommends including “evidence that the items are ‘instructionally sensitive,’ that is, that item performance is more related to the quality of instruction than to out-of-school factors such as demographic variables” (p. 4).

Ruiz-Primo and colleagues (2012) articulated an interpretive and validity argument around IS assessments (Figure 1). Their work is reviewed in more depth below, but the assumptions and arguments are informative at this point. A more typical line of argument is presented on the left; the unique contribution of Ruiz-Primo and her colleagues is presented on the right, including a discussion of the distance between test items and instruction. Interestingly, this is part of their validity assumptions and argument, where, as we will see shortly, this is also a source of critique of large-scale assessments (that they are too distant and different from instruction to be IS).



Interpretive Argument

Instructionally Sensitive Assessments (ISA) provide information about whether students were taught certain curriculum materials

Assumptions

Answering ISA items requires the knowledge and skills taught in the curriculum materials

Answering ISA items with varying sensitivity requires students' near and far transfer of the knowledge and skills taught in the curriculum materials

Validity Arguments

Students who learned the curriculum materials should show better performance when compared to their own performance prior to being taught the curriculum materials or to the performance of a different group who were not taught the curriculum materials

Students who were taught the curriculum materials should show highest gain in their performance in close items, followed by proximal, and distal items when compared to those who were not taught the curriculum materials

Evidence

Students' scores on the pretest and posttest

Students' scores on the pretest and posttest broken down by proximity

Figure 1. Validation model for instructionally sensitive assessments.
Source: Ruiz-Primo et al. (2012).

THE INTEGRATION OF INSTRUCTION, LEARNING, & ASSESSMENT

From a theory-design perspective, there are multiple layers to the argument regarding the importance or relevance of IS of assessments. We need to believe that teaching does impact learning; that standards are represented on the test and the standards are present in the curriculum and instruction. We also need to believe that the test scores contain instructionally relevant information and that information can be used to identify student performance on the standards.

Rodriguez and Albano (2017) summarized several "models" of thinking about the role of assessment (Figure 2). This also reflects the sentiment of Tiegs (1931), discussed above.

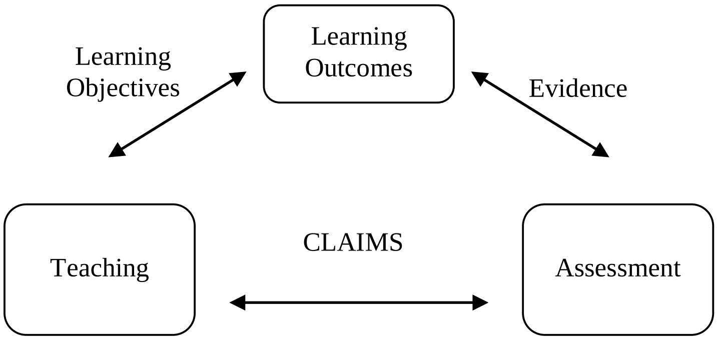


Figure 2. A common model of instruction, learning, and assessment.

The principles behind this simple framework include the notion of validity as argument (evidence in support of intended interpretations and uses, or claims), evidence centered design (explicitly connecting expecting learning objectives with task and evidence models), constructing measures (ala Wilson, 2005), and others. For Pellegrino and colleagues (2001), assessment is one of three components, also represented by a triangle including the vertices of curriculum, instruction, and assessment, with a "theory of learning and knowing" at the core center. They also promoted an assessment triangle, with vertices including a model of student ***cognition*** and learning, assumptions regarding the kinds of ***observations*** providing evidence of student competencies, and a sense-making ***interpretation*** process. The assessment triangle is fruitless without intentional connections among curriculum, instruction, and assessment.

With respect to the practice of item development and test design, Airasian and Madaus (1983) argued that test-item formats also limited the connection between tests and instruction, as the typical MC format does not reflect the nature of instruction (which is typically not a selected response interaction between teacher and student, but more commonly a constructed-response interaction). This incongruence between test format and instruction format leads to weaker connections between tests and instruction.

Implication: Reflect the formats and modes of instruction in the test format/modes.

Burstein (1989) argued that existing psychometric models and the standard statistical analysis routines applied to test score data serve the purposes of monitoring individual differences and prediction, but are woefully inadequate for identifying instructional factors that influence performance, item content and process features that are sensitive to instructional practices, or ways to support the instructional utility of test results. Burstein suggested that the selection of analytical method for test score analysis should be driven by its sensitivity to the substantive processes underlying performance and the educational structures where instruction occurs. As examples, a number of relevant contexts with implications for analytical models were offered, including mathematics taught through the presentation and use of formulas, or symbolic manipulations of equations, or the use of formulas within real-world problems. These examples not only suggest different test designs, but also analyses intended to detect such context differences with the hope of informing instructional decisions. IS as the impact of heterogeneous learning environments on performance must be accommodated in the psychometric modeling for the evaluation of IS (Burstein, 1989).

A Closer Look at Burstein's Work

To shift direction regarding test design and analysis, Burstein (1989) articulated the goals set out by the Center for Research on Evaluation, Standards, and Student Testing (CRESST), to address "how ability/achievement measures reflect educational and other experiences, or, in other words, on tests/assessments as measures of instruction/schooling/education" (p. 7). They addressed how differences in knowledge, skills, and abilities result from educational experiences. Whereas standard psychometric methods approach testing from the knowledge (ability) side, they argued that we should approach testing from the instruction side. The proposed psychometric model is one with test item characteristics, ability and experience characteristics, and their interactions. Burstein articulated a number of propositions:

1. The purpose of instructionally sensitive testing is to reveal student ability as a consequence of instruction.
2. Tests must accurately reflect the relevant multiple dimensions of knowledge and cognitive processes in a subject area.
3. The analysis model should be able to separate ability-relevant from construct-irrelevant dimensions from the performance under review.
4. Regarding item development, the item format and response mode should not interfere with the measurement of underlying ability.
5. Psychometric models that are instructionally sensitive are able to differentiate subject-matter teaching from teaching to the test (item-specific knowledge versus domain knowledge).

Implication:	Employ formats and response modes that support the measurement of the underlying abilities.

Burstein (1989) provided an extended example from Muthén, Kao, Burstein (1991). The instructionally sensitive psychometric approach to analyzing a math test employed a structural equation model (SEM) with item response theory (IRT) parameterization. In their analyses, they identified two levels of the effects of OTL. On one level, OTL influences the general latent achievement trait being measured. On another level, OTL influences performance on specific items. What they initially thought was evidence of instructionally sensitive items, was the direct effect of OTL – which they label as a form of item bias. The bias is due to the teacher specifically teaching the concept contained in an item because it is represented in the test blueprint, rather than instruction of the concept as it is represented in the broader domain, which should be comprehensively covered in instruction. This modeling incorporates the estimation of general and specific abilities as represented in instruction and the test.

Implication:	Reflect concepts addressed in instruction, but not so that they are tied too closely to the concept as presented in instruction, rather to address the broader concept intended by the standards.

Consistent with Burstein's framing, D'Agostino, Welsh, and Corson (2007) found significant association between how achievement standards were taught and tested on the 2003 fifth-grade mathematics Arizona state test. They found that teachers who emphasized the relevant standards more and whose teaching was consistent with the test had students with higher performance, conditioned on prior achievement: "students had to learn the standards similar to the way they were tested" (p. 19).

Implication:	Create tests that emulate the way students learn the standards.

Testing for Teacher Quality and IS

In a partnership between WestEd and CRESST, the *Assessment and Accountability Comprehensive Center* (Herman, Heritage, & Goldschmidt, 2011) identified a number of design characteristics in their guidance for developing (or selecting) measures of student growth for use in teacher evaluation. To support the use of growth measures for teacher evaluation, a validity framework was articulated, based on a number of propositions. This included the proposition that standards clearly define what students are expected to learn, that tests are designed to accurately and fairly measure what students are expected to learn, and that student growth scores can be accurately and fairly ascribed to teaching practices (p. 3). If these propositions can be supported with evidence, then test scores can be used to make judgements about teacher effectiveness.

Among the Herman and colleagues (2011) propositions was a set of design practices to establish an evidence basis for the validity framework. To support the proposition regarding the ability of tests to accurately and fairly measure what students are expected to learn and their growth, the following design claims were made:

1. Test specifications reflect the breadth and depth of learning expectations;
2. Test items comprehensively reflect learning expectations;
3. Test design, administration, and scoring produce reliable scores;
4. Test accessibility and fairness is secured for all students;
5. Tests accurately measure growth over the course of the year;
6. Cut scores are justifiable; and
7. Tests are designed to be instructionally sensitive.

Unfortunately, the evidence sources for these propositions did not extend far beyond the typical approaches, including expert review of alignment, administration, scoring, and sensitivity. An additional source of evidence included unspecified "research studies."

Regarding the proposition the scores reflect the efforts of teachers, the following design claims were made:

1. Scores are instructionally sensitive
2. Scores representing teacher contributions are sufficiently reliable and relatively free of bias.

The evidence sources included research studies on IS and statistical analyses of reliability and bias.

In the deeper descriptions of these claims and evidence sources, the authors present sound arguments regarding the need for test specifications to include the full range of milestones of knowledge and skill development relevant over the course of a year, to capture the full range of potential progress.

Implication:	Item development should be mapped across the range of content as it is relevant to instructional content. Items shouldn't simply address the "exit" criteria, but represent the entire learning progression.

Regarding the claims of IS, the authors argue that this is an essential design characteristic. When tests are used to support teacher evaluation, inferences about IS are inherent. They argued that items must reflect core goals represented in the standards and learning progressions (assuming that standards reflect learning progressions). Unfortunately, the evidence sources were not well articulated: "Independent review of assessment items and tasks by subject matter and teacher experts can provide evidence of sensitivity of the test design" (p. 10). There are a number of assumptions in that statement that need to be clarified for a productive review.

IS ITEM & TEST DESIGN GUIDANCE

IS Design Guidance á la Popham

To turn the discussion upside down, Popham (2008) asked: Why would test items be instructionally *insensitive*? He proposed several reasons, including, weak alignment, items that are too easy or too difficult, poor item writing, strong connections to socioeconomic status (or other background characteristics), and a measure that is too dependent on academic aptitude. He argued that once we (in the measurement community) became aware of the detrimental effects of bias (e.g., DIF), we addressed it intensely; we should be able to do the same thing with instructional insensitivity. Popham suggested two strategies, including a judgement-based strategy and an empirical approach by contrasting (effectively) taught v. untaught (ineffectively taught) students. Popham and Ryan (2012) endorsed three types of evidence to support the IS evaluation of tests, including curricular clarity, judgmental evidence of item sensitivity, and empirical evidence of item sensitivity. Regarding the judgmental evidence, Popham provided examples of the characteristics of items that might cause judges to rate it as instructionally insensitive, which are reminiscent of his 2008 insensitivity critique of tests:

1. Not aligned with the intended curricular aim
2. Too easy or too difficult
3. Poorly written or includes an item-writing flaw

4. Correct responses are associated with student characteristics (e.g., SES) that are construct irrelevant.
5. Measures general aptitude or intelligence.

He argued that item-review committees that identify under-developed items, such as item bias review committees, can be used to identify instructionally *insensitive* items. This allows the test developer to modify or remove the item from the item pool.

Implications: Design items that are connected with curriculum, at the appropriate level of difficulty, well written and designed consistently with high-quality guidelines, bias free, and not heavily loaded on general aptitude/intelligence.

Popham (2001) recommended four rules to support the instructionally-relevant information from standards-based assessment. The four rules are (a) test the highest priority outcomes; (b) include tasks that require students to use key knowledge and skills; (c) clearly describe knowledge and skills tested so educators can address the required cognitive demands; and (d) review test items and descriptions with the rigor appropriate with the stakes and intended use of the test. Note that these are reflective of the item-writing guidelines presented below (from Haladyna & Rodriguez, 2013). His definition of instructionally sensitive tests includes explicit assumptions about the quality of instruction. Common state test score interpretations and uses include inferences about the effectiveness of instructional programs, curriculum coverage, and associated school quality factors; if the test is instructionally insensitive, such inferences are indefensible (Popham, 2014).

Implications: Write high-quality items. Provide high quality instruction to all students.

Popham (2011) provided guidance regarding the expert review of item IS. Among these, he included the following sources of influence:

1. SES. The experts should rate the extent to which the likelihood of a correct response is affected by family socioeconomic factors.
2. Aptitude. Similarly, experts should rate the extent to which the likelihood of a correct response is affected by verbal, quantitative, or spatial aptitudes. This also addresses the extent to which the test is a measuring grade-level content standards, rather than general ability or intelligence, the long-term accumulation of abilities, which is much less affected by recent learning.
3. Instruction. Experts should rate the extent to which quality instruction on the topic/skill being addressed in the item should affect the likelihood of a correct response.

Instructionally sensitive items should receive responses of No to the first two question and Yes to the third.

Implications: As items undergo content and sensitivity review, add a cycle of review that examines items for the potential influence of SES, aptitude, and instruction.

IS Design Guidance á la Ruiz Primo

A concerted effort was undertaken to systematically investigate the potential of instructionally sensitive items and tests, from conceptualization, to design, to administration and use, to identify the effects of education reform on student achievement. A core contributor to these efforts is Ruiz-Primo. In some ways, tests that are instructionally sensitive are similar to tests that indicate the effects of educational reform – to the extent that reform is instructionally relevant. To support this work, Ruiz-Primo and colleagues (2002) undertook a study to respond to what appeared to be weak effects of the National Science Foundation's (NSF) State Systemic Initiatives in mathematics and science education. Employing the notion of proximity of indicators to the underlying processes being observed, as described by Richard Snow (1968), they employed a multilevel framework to guide judgements about the distance between the curriculum efforts and student performance (test) results. They included five levels of assessment distances. Each was specified in terms of the intended goals, the content specifications, and the nature of the assessment tasks themselves, as a function of the distance to instruction.

1. Remote. Typically includes very general measures of achievement, for example, the National Assessment of Educational Progress (NAEP). The content is in practice different than the relevant curriculum for a particular student or school and the nature of the assessment tasks is completely different than what students experienced during the curriculum-based instruction.
2. Distal. Assessments that are based on broader standards across a given domain or subject matter; for example, state tests or CCSS consortium tests. The content domain is expected to be similar, although with the use of topics that were not directly included in instruction, with assessment tasks that are more advanced than those used during instruction.
3. Proximal. Assessments of knowledge and skills that are relevant to the curriculum, but based on topics that are related to but not studied in relevant units; for example, assessments that cover the same content covered in curricular-based lessons, but not based on classroom activities. The content covers the same topics addressed in the curriculum, but through the use of different concepts or procedures, where the assessment tasks are similar to those employed in instruction.
4. Close. Assessments that are close to the content and activities of the curriculum; for example, embedded assessments that are periodic, covering more general knowledge and skills associated with instructional lessons; possibly includes common-course assessments today. These are more typically at the level of an instructional unit or chapter. The topics, concepts, and procedures are the same as those employed in instruction, as are the assessment tasks and activities.
5. Immediate. Direct artifacts of the enactment of the curriculum; for example, a science class lab-book, class journal, or classroom assessments. These are typically tied to specific instructional activities, within a unit of instruction.

In their study of science classrooms, teachers, and students, Ruiz-Primo and colleagues (2002) found that close assessments were more sensitive to changes in student performance compared to more proximal assessments, where the characteristics of the assessments had more influence on evaluating student improvement than expected.

Implication:	IS tests are closer in proximity to instruction. It is also possible that large-scale tests that cover more general content with items/tasks functionally different than those experienced in any given instructional setting are not IS by design.

Example assessments provide keys into the design characteristics. This work was extended and further supported by Ruiz-Primo (2012), through NSF support to develop and evaluate IS assessments. From a comprehensive review of the literature on the topic of IS (I highly recommend reading that review), Ruiz-Primo and colleagues found that most prior work focused on the evaluation of IS, primarily through judgmental or empirical methods. Prior researchers focused their efforts on existing tests and assessment, and had not attempted to construct IS items or tests. IS assessments have three characteristics (Ruiz-Primo & Li, 2008; Ruiz-Primo et al., 2012): they (a) represent the curriculum content in instruction and (b) reflect the quality of instruction in a way that (c) supports formative purposes. Their approach to IS was dependent on the proximity of the assessment to instruction (the curriculum as delivered). Their item design process focused on promoting mental models the explore "why" questions or big ideas. Given the definition of learning goals and curriculum characteristics as realized in instructional activities, and the identification of big ideas, item development then followed a relatively systematic process to design tests of varying levels of IS:

1. Begin with items that have the characteristics of items *close* in proximity to instruction.
2. Extend the development of more distant items by manipulating the item characteristics that distinguish items at different levels of proximity.
3. Gather validity-related evidence to support the IS of items as intended based on their proximity to instruction.

Without making the connections explicit, this approach has strong connections to evidence-centered design (Hendrickson, Huff, Luecht, 2010; Huff, Steinberg, Matts, 2010).

To support the design of varying levels of IS, the research team clarified that distance is not the same as cognitive complexity. Whether we call it cognitive taxonomy, cognitive demand, or depth of knowledge, this is not determined or restricted based on distance or proximity to instruction. Distance, as conceptualized by Ruiz-Primo and colleagues is about "whether students have the opportunity to learn the content tapped by the item at hand, and whether what they learned can be transferred within the same knowledge domain but in different contexts" (p. 694).

Implication: Cognitive complexity is not manipulated by designing items to be more or less proximal to instruction.

The items developed by Ruiz-Primo and colleagues (2012) were in MC format, and developed in triads, including one close item, one *near* proximal and one *far* proximal. The close items were designed to evaluate the extent to which learning occurred following instruction, in ways that were consistent with instructional activities. The proximal items were explicitly designed to evaluate the extent to which students were able to transfer learning. The authors argued that content expertise was required to allow for systematic and consistent modification of items relative to IS (the use of external item writers not as familiar with the content or instructional activities was not as successful). To develop these triads, five dimensions of items were manipulated:

1. Characteristics of the question regarding the familiarity of the question and its similarity to the kinds of questions asked during the instructional activities.
2. Exposure students received to the big ideas; essentially, the opportunity to learn given a basic frequency metric of none, a few items, multiple times.

3. The cognitive demands that are consistent between the cognitive strategies addressed during instructional activities and those present in the test items.
4. The contexts and settings addressed during instructional activities and those present in the test items. In their study of science assessments, they addressed four specific aspects of setting, including type of organism or target object, type of process, the scenario or set-up of the item (activity), and graphical representation.
5. Regarding specific items that focus on experiments in science, the setting of the experiments with respect to the independent and dependent variables and their similarity between instructional activities and those present in the test items.

In each case, items were manipulated by introducing no change, small changes, and big changes from those that were presented during instruction. They tracked which component of the MC item was manipulated, in terms of the stem, options, or both.

Implications: Possible IS features of MC items that are malleable, include similarity to specific instructional activities in terms of specific content, settings or contexts, and cognitive tasks; and the extent to which the content of the item is frequently presented in instruction (dosage).

Ruiz-Primo and colleagues (2012) argued that the design of proximal items requires “intimate knowledge of the specific curriculum” (p. 707). They further recognized that there can be significant differences between the intended and enacted curriculum – that using the intended curriculum can result in a moving target – “what is close for some teachers may be proximal for others” (p. 707). And although the authors promoted the use of this approach for classroom and district level application, as well as curriculum development, they didn’t see promise for the approach to result in instructionally sensitive tests without intimate knowledge of the curriculum.

Implication: To the extent that instruction varies across teachers, it is unlikely that any given item will be equally IS across teachers.

To investigate differences in effects of IS that appeared in prior studies, which may be the result of the specific instructional lessons or the format of the assessments (SR v. CR), Ruiz-Primo et al. (2013) examined the item format directly. Eight MC items were developed to be close, near proximal, and far proximal in bundles of threes; distal items were selected from released items from state and international (e.g., TIMSS) assessments. To develop the CR items, the options were removed from the eight MC items and the stems were slightly modified to ensure the prompt was in question form. Students only took items in one format (although there were only eight MC or CR items across the three instructional lessons). They examined pre-post instructional differences by item and test scores. Although the two formats differed in their IS, the number of items was too small to support strong conclusions about which format was more likely to result in higher IS; there was a lot of variation in the format effects across the three instructional lessons.

Implication: Test item format might be less relevant to creating IS tests than other item characteristics.

Similar to the work of Ruiz-Primo (2002), Ing (2008) suggested that tests of specific lessons that have close proximity to instruction should be more instructionally sensitive than those tests of more general achievement or are more distant from instruction. Some tests are designed to be more instructionally sensitive to others. Ing argued that the focus of the work on IS is on strengthening the validity of inferences regarding instructional quality. Furthermore, any study of this link between the test and instruction requires (a) detailed information about instruction, (b) multiple measures of student performance, and (c) direct analyses of the link between instruction and student test scores.

IS Design Guidance á la PARCC

In 2008, Denny Way produced a memo for the Partnership for Assessment of Readiness for College and Careers (PARCC) on IS. The memo reviewed the PARCC purposes and positioned IS at the heart of the validity argument inherent in PARCC. In the memo, he included a section on relevant test design characteristics. That section is reproduced here (pp. 8-9):

> Test Design and Instructional Sensitivity
>
> The PARCC assessments have been developed using an Evidence Centered Design (ECD) approach. In the ECD framework, an argument is made from observed student behaviors in response to particular tasks and items to support claims about what students know and can do. The ECD process thus facilitates the existing rationales and evidence supporting intended score interpretation and use.
>
> In this context, the first chain in a validity argument addressing instructional sensitivity is found in the fidelity of the PARCC test design with the CCSS. This relates to curricular validity as shown Instructional Sensitivity August 15, 2014 Page 9 in Figure 1: the degree to which test items represent the objectives of the curriculum. PARCC has released a set of test specification documents, including assessment blueprints and evidence statement tables, to document the design of the PARCC assessments. Evidence statement tables and evidence statements describe the knowledge and skills that an assessment item or a task elicits from students, and are aligned directly to the CCSS.
>
> For both ELA/literacy and mathematics, PARCC has developed Cognitive Complexity Frameworks that explicitly describe how rigor (from a cognitive standpoint) is woven into the item and task development process. These frameworks help to serve one of the stated purposes of the PARCC assessments: to assess the full range of student performance, including high- and low-performing students. At least conceptually, items and tasks written to a lower level of text complexity are more likely to accurately measure low-performing students, and may also be more sensitive to instruction provided to low-performing students. Still another feature of the PARCC test design that may relate to instructional sensitivity is the use of partial credit scoring for many of the item types on the test. That is, compared to traditional test items with binary (right/wrong) scoring, lower performing students may be more likely to achieve partial credit score on items after instruction even if they are not able to obtain full credit.
>
> Thus, several aspects of the PARCC test design and development process are relevant to instructional sensitivity: the ECD process; explicitly defined task models

and associated documentation; developing items to a range of cognitive complexity, and developing items with partial credit scoring. All of these elements can be used in supporting claims about the instructional sensitivity of PARCC items and assessments, although empirical data to be collected will be critical to supporting (or possibly refuting) these claims.

Implications:	To extend the capacity of large-scale statewide tests to provide IS information is to build test blueprints with additional layers of supports for item writers and instructional leaders. This includes tools used by PARCC, such as evidence statement tables, cognitive complexity frameworks, and consider partial-credit scoring.

IS Design Guidance á la Traditional Item Development

In considering item development and test design from a traditional perspective, Haladyna and Rodriguez (2013) addressed the role of IS. In this context, IS was introduced as the ability of items to discriminate among students due to the effectiveness of instruction. These authors, as item developers, acknowledged that IS can have an impact on both item difficulty and item discrimination. As item statistics change from pre-instruction to post-instruction, or differ markedly between students who received relevant instruction versus those who have not, expert reviewers can make informed judgements regarding IS potential of the item. However, judgements about the effectiveness of instruction are best made across items – that is, with many items or total test scores, as any one item is an unreliable index of the larger evaluative inferences.

Whatever the intended inference from a specific test, high-quality item development must be achieved. It is the subject-matter expert that then translates the specific purpose of the test, guided by the test blueprint and design specifications, by writing items under that guidance.

A line of research syntheses and reviews of expert guidance led to the development of item-writing guidelines (Haladyna & Downing, 1989a, 1989b; Haladyna, Downing, & Rodriguez, 2003; Haladyna & Rodriguez, 2013). Many of these guidelines are “item-writing niceties” (Mehrens, personal communication, 1996). Item-writing researchers have long lamented the underdeveloped science around item development (beginning with Ebel, 1951). Only a handful of the 27 guidelines in Table 1 have empirical evidence supporting them (see Haladyna & Rodriguez, 2013, for a summary of that evidence). Nevertheless, these guidelines represent the best of what the field has to offer regarding the design of sound items. As items are the building blocks of any test, care must be given to their design for any purpose. And high-quality item writing is a requirement for IS based on the Popham approach.

Implication:	High-quality item writing is necessary for IS.

Table 1.
Guidelines for Writing Multiple-Choice Items (Source: Haladyna & Rodriguez, 2013).

CONTENT CONCERNS

1. Base each item on one type of content and cognitive demand.
2. Use new material to elicit higher-level thinking.
3. Keep the content of items independent of one another.
4. Test important content. Avoid overly specific and overly general content.
5. Avoid opinions unless qualified.
6. Avoid trick items.

FORMAT CONCERNS

7. Format each item vertically instead of horizontally.

STYLE CONCERNS

8. Edit and proof items.
9. Keep linguistic complexity appropriate to the group being tested.
10. Minimize the amount of reading in each item. Avoid window dressing.

WRITING THE STEM

11. State the central idea clearly and concisely in the stem and not in the options.
12. Word the stem positively, avoid negative phrasing.

WRITING THE OPTIONS

13. Use only options that are plausible and discriminating. Three options are usually sufficient.
14. Make sure that only one of these options is the right answer.
15. Vary the location of the right answer according to the number of options
16. Place options in logical or numerical order.
17. Keep options independent; options should not be overlapping.
18. Avoid using the options *none-of-the-above, all-of-the-above,* and *I don't know*.
19. Word the options positively; avoid negative words such as NOT.
20. Avoid giving clues to the right answer:
    a. Keep the length of options about equal.
    b. Avoid specific determiners including always, never, completely, and absolutely.
    c. Avoid clang associations, options identical to or resembling words in the stem.
    d. Avoid pairs or triplets of options that clue the test taker to the correct choice.
    e. Avoid blatantly absurd, ridiculous options.
    f. Keep options homogeneous in content and grammatical structure.
21. Make all distractors plausible. Use typical errors of test takers to write distractors.
22. Avoid the use of humor.

From the list of 27 guidelines, there are several that address issues more directly related to IS. But, some of these might work against IS given the arguments and evidence presented above. Below, a few guidelines are discussed briefly, informed by research and practice concerning IS.

1. Base each item on one type of content and cognitive demand.
   This is a key guideline in all cases, but particularly acute in the IS arena. It likely requires an additional level of specificity, regarding the close connection to instructional activities for items that are intended to be IS at a close level.

2. Use new material to elicit higher-level thinking.
   This guideline is somewhat challenging. The extent to which the item or test materials are "new" will likely determine the nature of the IS of the items and test. If the materials are quite different, this will be helpful in testing transfer of KSAs, but likely lead to items/tests that are less IS, at least less proximal.

3. Keep the content of items independent of one another.
   This guideline is helpful for some purposes, particularly with IRT scaling (regarding assumptions of local independence). However, richer item sets (even sets of MC items) can be integrated in ways to express more complex processes, but such item sets might need to be scored as testlets or with partial credit. This is consistent with the idea of being more sensitive to a broader range of cognitive abilities and likely will enhance IS.

4. Test important content. Avoid overly specific and overly general content.
   Especially regarding IS, to the extent that instructional activities are focused on important content, close alignment to instruction will ensure that trivial content is not present in the test (simply recognizing the expense of real-estate on any large-scale test). However, as more distal tests adopt more general overviews of broad standards, this guideline is likely to be violated, at the cost of IS.

9. Keep linguistic complexity appropriate to the group being tested.
   This goes a long way in terms of maintaining accessibility of the test to the potentially very diverse intended test taking population. It also helps achieve a focus on the content and skills as they were addressed in instructional activities. Item writers should be wary of trying to manipulate item difficulty by employing complex language structures or infrequent vocabulary, as this is likely to reduce IS, making the items more distal to instruction.

10. Minimize the amount of reading in each item. Avoid window dressing.
    As with the guideline on linguistic complexity and focusing on one type of content and cognitive demand, this reminds us to purify the item of extraneous materials and skills. Unless the task is to differentiate between what is relevant and what is not, we've learned from a fair amount of research in test accessibility that focus and getting to point in the item improves the item functioning for all test takers. It also helps us focus on the design specifications, potentially improving the IS characteristics of the items – as we stay closer to the elements of instruction and not introduce irrelevant content (that is likely more distal to instruction).

11. State the central idea clearly and concisely in the stem and not in the options.
    Again, here is another guideline that helps the item writer focus on the intended content and cognitive skill – by addressing the big idea or core problem immediately, clearly, and unambiguously.

13. Use only options that are plausible and discriminating. Three options are usually sufficient.
    Although this recommendation is grounded in a unanimously consistent body of empirical research (Rodriguez, 2005), the goal here is to know that the reason to use a specific number of options is because that is what fits the item. Typically, three options are sufficient; but sometimes, we need balance in the options (two positive/two negative). But item writers should have a reason for writing the number of distractors they have written, and not to follow a "blind" unjustified rule such that all items must have four options or five options.

20. Avoid giving clues to the right answer:
    a. Keep the length of options about equal.
    This is such a common error, particularly among teacher-made tests, that I can't leave it out of this discussion. It's a simple thing to avoid, but most of us inadvertently tend to write the correct option in a little more detail, to defend its correctness. Easy to fix.

    b. Avoid specific determiners including always, never, completely, and absolutely.
    c. Avoid clang associations, options identical to or resembling words in the stem.
    d. Avoid pairs or triplets of options that clue the test taker to the correct choice.
    e. Avoid blatantly absurd, ridiculous options.
    f. Keep options homogeneous in content and grammatical structure.
    Issues a through f are all common item-writing error that are easy to fix. These errors provide clues, which become construct-irrelevant advantages to test-wise students. This absolutely interferes with our assessment of IS, as the test then measures test-wiseness instead of the impact of instruction.

21. Make all distractors plausible. Use typical errors of test takers to write distractors.
    Finally, this augments the guideline (13) regarding the number of options, and is among the most useful to support IS goals. The distractors can provide diagnostic information about the misconceptions or typical errors made by students, if the distractors represent the likely errors. Errors and misconceptions that students share during instruction make excellent distractors. But this requires expertise and knowledge of the relevant instructional activities in close proximity.

CONCLUDING THOUGHTS

There is a great deal of consistency in the propositions regarding the meaning of IS, ways of evaluating IS, and its importance. There is less consistency regarding the mechanisms that enable

IS and how to secure them, in the service of enhancing IS. There is also significant doubt as to the ability of large-scale tests, particularly at the state or national level, can embody IS.

Burstein (1989) presented one of the deepest challenges to the identification and role of IS, in the context of OTL. On one hand, we want the test to be sensitive to instruction and reflect the status of student knowledge relative to the standards. But if a teacher teaches a specific concept because it is in the test blueprint, potentially at the expense of the broader conceptualization of the content standards, this may reflect a kind of item bias, a direct effect of highly specific OTL.

Yet, there are multiple authors who have advocated creating tests that emulate the way students learn (D'Agostino, Welsh, & Corson, 2007; Ruiz-Primo et al., 2012) as a way to achieve IS. Popham (2008, 2011) promoted good item development, good test design, and the use of an item-review panel to evaluate IS, noting that it must begin with an assumption of quality instruction.

Ruiz-Primo and her many colleagues took on the challenge directly and attempted to create items that varied in IS. Unfortunately, their hierarchical system of distance between instruction and the test renders IS to a simple function of proximity. One might read this body of work and conclude that only classroom tests can truly embody IS; that tests that are designed further from the classroom (district tests, state tests, national and international tests) will naturally have less IS. Because of this, state and national tests (and perhaps district tests) will unlikely result in instructionally relevant information (similarly argued by Ing, 2008). Through efforts to develop tests with different degrees of proximity to specific instructional lessons and activities, Ruiz-Primo and her colleagues were able to design items that functioned as expected given the extent to which they represented, closely or distally, the instructional activity versus the broader standards. In numerous ways, they declared that securing items with strong IS required "intimate knowledge" of the curriculum and how it was enacted through instruction, going as far as to say that proximity is teacher-specific.

PARCC (as well as Smarter Balanced and other efforts, including redesign of AP science exams and others) have developed powerful approaches to item development that promotes IS. These approaches employ strong item-writing guidelines, elements of ECD and universal design, and attend closely to content, curricular assumptions, and goals of improving instruction. But they all depend on high-quality instruction and uniform OTL.

Popham spoke most directly to the role of the item developer, to ensure high-quality items and to engage in expert review for features specifically enhancing or hindering IS. The traditional item-writing guidance also embody many of the characteristics of direct and indirect item development guidance from the IS researchers reviewed here (highlighted in the implications called out throughout the paper). In all cases, high-quality test items are the foundation of each effort to provide instructionally relevant information.

To provide easy access to those guidelines, through the noted implications of this deep body of work, they are summarized here, without repeating the item-writing guidelines presented in Table 1.

Table 2

Summary of Item and Test Design Implications from Research on IS

General Implications

1. Create tests that emulate the way students learn the standards.
2. Write high-quality items. Provide high quality instruction to all students.
3. As items undergo content and sensitivity review, add a cycle of review that examines items for the potential influence of SES, aptitude, and instruction.
4. To the extent that instruction varies across teachers, it is unlikely that any given item will be equally IS across teachers.
5. To extend the capacity of large-scale statewide tests to provide IS information is to build test blueprints with additional layers of supports for item writers and instructional leaders. This includes tools used by PARCC, such as evidence statement tables, cognitive complexity frameworks, and consider partial-credit scoring.
6. High-quality item writing is necessary for IS.

Content Implications

7. Reflect concepts addressed in instruction, but not so that they are tied too closely to the concept as presented in instruction, rather to address the broader concept intended by the standards.
8. Item development should be mapped across the range of content as it is relevant to instructional content. Items shouldn't simply address the "exit" criteria or what students know by the end of a year, but represent the entire learning progression.
9. Design items that are connected with curriculum, at the appropriate level of difficulty, well written and designed consistently with high-quality guidelines, bias free, and not heavily loaded on general aptitude/intelligence.
10. IS tests are closer in proximity to instruction. It is also possible that large-scale tests that cover more general content with items/tasks functionally different than those experienced in any given instructional setting are not IS by design.
11. Cognitive complexity is not manipulated by designing items to be more or less proximal to instruction.
12. Possible IS features of MC items that are malleable, include similarity to specific instructional activities in terms of specific content, settings or contexts, and cognitive tasks; and the extent to which the content of the item is frequently presented in instruction (dosage).

Format Implications

13. Reflect the formats and modes of instruction in the test format/modes.
14. Employ formats and response modes that support the measurement of the underlying abilities.
15. Test item format might be less relevant to creating IS tests than other item characteristics.

References

Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103-118.

Baker, E.L. (2008). *Empirically determining the instructional sensitivity of an accountability test*. Paper presented at the annual meeting of the American Educational Research Association. Retrieved at http://www.cse.ucla.edu/products/overheads/AERA2008/baker_sensitivity.pdf

Baker, E.L., Berliner, D.C., Camp Yeakey, C., Pellegrino, J.W., Popham, W.J., Quenemoen, R.F., Rodriguez-Brown, F.V., Sandifer, P.D., Sireci, S.G., & Thurlow, M.L. (2001). *Building tests to support instruction and accountability: A guide for policymakers*. The Commission on Instructionally Supportive Assessment. American Association of School Administrators. Retrieved at https://www.aasa.org/uploadedFiles/Policy_and_Advocacy/files/BuildingTests.pdf

Board of Testing and Assessment. (2009). *Comments on the Department of Education's proposal on the Race to the Top Fund*. Washington DC: National Academy of Sciences. Retrieved at https://www.nap.edu/catalog/12780/

Burstein, L. (1989). *Conceptual considerations in instructionally sensitive assessment* (CSE Tech Report 333). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, University of California. Retrieved at http://cresst.org/wp-content/uploads/R333.pdf

Chen, J., & Kingston, N. (2012, April). *Detecting item sensitivity to instruction: A comparison between Mantel-Haenszel and logistic regression procedures*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, CA. Retrieved at https://aai.ku.edu/presentations

D'Agostino, J.V., Welsh, M.E., & Corson, N.M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1-22.

Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Ebel, R.L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational Measurement* (1st ed., pp. 185-249). Washington DC: American Council on Education.

Haladyna, T.M. (1976, April). *The quality of domain-referenced test items.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1,* 37–50.

Haladyna, T.M., & Downing, S.M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51–78.

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*(3), 309–334.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*(1), 39-53.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education*, *23*(4), 358-377.

Herman, J.L., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version*). Lost Angeles, CA: University of California, national Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design large scale assessment. *Applied Measurement in Education, 23*(4), 310-324.

Ing, M. (2008). Using instructional sensitivity and instructional opportunities to interpret students' mathematics performance. *Journal of Educational Research & Policy Studies, 8*(1), 23-43.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current application*. San Francisco, CA: McCutchan, 1974.

Monroe, W.S., & Clark, J.A. (1924). Measuring teaching efficiency. *University of Illinois Bulletin, 21*(22), 3-26. Retrieved at https://www.ideals.illinois.edu/bitstream/handle /2142/32447/measuringteachin25monr.pdf

Muthén, B.O. (1988). *Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study*. CSE Technical Report 286. Los Angeles, CA: UCLA Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B.O., Kau, C-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1-22.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Polikoff, M.S. Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3-14.

Popham, W.J. (2008, December). *Can instructionally insensitive accountability tests ever evaluate educators fairly?* Presentation at the Winter Conference of the Washington Educational Research Association, Seattle, WA.

Popham, W.J. (2011). Instructional insensitivity of tests: Accountability's dire drawback. In A.C. Ornstein, E.F. Pajak, S.B. Ornstein (eds.), *Contemporary issues in curriculum* (5th ed., pp. 295-302). Upper Saddle River, NJ: Pearson Education, Inc.

Popham, W.J., Berliner, D.C., Kingston, N.M., Fuhrman, S.H., Ladd, S.M., Charbonneau, J., & Chatterji, M. (2014). Can today's standardized achievement tests yield instructionally useful data? *Quality Assurance in Education, 22*(4), 300-316.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

Popham, W.J., & Ryan, J.M. (2012, April). *Determining a high-stakes test's instructional sensitivity*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, CA.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Rodriguez, M.C., & Albano, A.D. (2017). *The college instructor's guide to writing test items. Measuring student learning*. New York, NY: Routledge.

Ruiz-Primo, M.A., Li, M., Wang, T., Giamellaro, M., Wills, K., Orgeron, M., & Zhao, D.Y. (2013). *Comparing item formats of instructionally sensitive assessments*. DEISA Paper 1. Denver, CO: School of Education and Human Development, University of Colorado. Retrieved at http://source.ucdsehd.net/deisa/

Ruiz-Primo, M.A., Li, M., Wills, K., Giamellaro, M., Lan, M-C., Mason, H., Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research on Science Teaching, 49*(6), 691-712.

Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L.S., & Klein, S. (2002). On the evaluation of systematic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Scheerens, J. (Ed.) (2017). *Opportunity to learn, curriculum alignment and test preparation: A research review*. New York, NY: Springer.

Stiggins, R.J., & Conklin, N.F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany, NY: State University of New York Press.

Tiegs, E.W. (1931). *Tests and measurements for teachers*. Cambridge, MA: The Riverside Press.

Way, W.D. (2014). *Memorandum on instructional sensitivity considerations for the PARCC assessments*. Partnership for Assessment of Readiness for College and Careers. Retrieved at www.parcconline.org/files/65/.../Instructional_sensitivity_memo_final08_15_14.pdf

Wiliam, D. (2007, April). *Three practical, policy-focused procedures for determining an accountability test's instructional sensitivity: III: An index of sensitivity to instruction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Earlbaum.

Yoon, B., & Resnick, L.B. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance* (CSE Tech Report 484). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California. Retrieved at https://www.cse.ucla.edu/products/reports/TECH484.pdf