Psicothema

Psicothema 2014, Vol. 26, No. 1, 127-135
doi: 10.7334/psicothema2013.258

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Validity evidence based on testing consequences

Suzanne Lane
University of Pittsburgh (USA)

Abstract

Background: Large-scale educational assessments in the U.S. are used as policy tools for improving instruction and evaluating educational programs and the effectiveness of educators. Because of the high-stakes nature of their use, it is imperative that validity evidence based on testing consequences is obtained to support their multiple purposes. **Method:** A comprehensive review of the literature related to validity evidence for test use was conducted. **Results:** A theory of action for a testing program should be delineated. A theory of action reflects an interpretative and use argument and a validity argument for assessment programs and delineates the purposes and uses of the system as well as the outcomes of the system. The articulation of a validity argument allows for the accumulation of evidence not only for, but also against, intended score interpretations and uses. **Conclusions:** As described in the paper, for assessment and accountability systems that are intended to have an effect on both instruction and student learning, the consequences, both positive and negative, of the systems need to be evaluated.

Keywords: Validity, standards, evidence of testing consequences, test use.

Resumen

Evidencia de validez basada en las consecuencias del uso del test. **Antecedentes:** las evaluaciones educativas a gran escala en los Estados Unidos son utilizadas como herramientas políticas para mejorar la instrucción, evaluar los programas educativos y la efectividad de los educadores. Debido al elevado impacto de sus usos, es imperativo obtener evidencias de validez basadas en las consecuencias del uso del test para sus múltiples objetivos. **Método:** se ha llevado a cabo una revisión comprehensiva de la literatura relacionada con la evidencia de validez basada en las consecuencias del uso del test. **Resultados:** debe elaborarse una teoría de la acción para un programa de evaluación mediante tests. Una teoría de la acción refleja un argumento interpretativo, un argumento de uso y un argumento de validez para los programas de evaluación, y define los objetivos y usos del sistema así como los resultados. La articulación de un argumento de validez permite la acumulación de evidencias no solo a favor, sino también en contra, de las interpretaciones y usos previstos de las puntuaciones. **Conclusiones:** como se describe en el artículo, para los sistemas de evaluación y rendición de cuentas que son diseñados para tener un efecto sobre la instrucción y el aprendizaje del estudiante, las consecuencias, tanto positivas como negativas, necesitan ser evaluadas.

Palabras clave: validez, standards, evidencia de las consecuencias del uso del test, uso del test.

In this article I address the need to evaluate the consequences of assessment programs in the context of an argument-based approach to validity which entails both an interpretative and use argument and a validity argument (Kane, 2006, 2013). The central role of consequential evidence in support of the interpretative and use arguments for educational assessments is then discussed. Following this is a review of the results from studies that have examined the consequences of the use of large scale performance assessments from various stakeholders'perspectives and using a variety of methods. The studies reviewed primarily examined the consequences of state-level performance assessment and accountability systems that were implemented in the early 1990's in the United States to serve, in part, as a tool for educational reform. This is followed by a discussion on how a theory of action for educational assessment and accountability programs can guide the delineation of the interpretive and use argument and the validity argument. Within a theory of action for an assessment system, the goals, purposes, and uses of an assessment system; the outcomes of the assessment system (e.g., increased rates of college and career readiness for all students); and the mediating outcomes necessary to achieve the ultimate outcomes (e.g., students will show gains on the assessment, instruction will improve) are articulated (Marion, 2010). Key components of the theory of action are then prioritized and further delineated to support the design of the assessment and the validity argument. Examples of theories of actions are used to demonstrate that consequences are a key component in the validation of educational assessments.

Received: August 27, 2013 • Accepted: October 2, 2013
Corresponding author: Suzanne Lane
Department of Psychology in Education
University of Pittsburgh
15260 Pittsburgh (Estados Unidos)
e-mail: SL@pitt.edu

Argument-based approach to validity

Validation entails constructing and evaluating coherent arguments for and against proposed test interpretations and uses (Cronbach, 1971, 1988; Messick, 1989; Kane, 1992) and therefore validation is an evaluation argument (Cronbach, 1988). A clear statement of the proposed interpretations and uses of test scores

is critical in the validation of assessment and accountability systems. This requires the specification of the inferences included in the interpretations and uses, evaluation of the proposed inferences and their supporting assumptions using evidence, and the consideration of plausible alternative interpretations. The examination of alternative explanations is an important aspect of validity evaluation.

The argument-based approach to validity entails both an interpretative argument and a validity argument (Kane, 2006). An interpretative argument "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading to the observed performances to the conclusions and decisions based on the performances" (p. 7). Kane (2013) has shifted from using the term "interpretative argument" to "interpretative/use argument (IU)" to emphasize the need to focus on uses of assessment results. The validity of an assessment and accountability program depends on the synthesis of the evidence for the evaluation of the IU argument (Haertel, 1999). A validity argument provides a structure for evaluating the merits of the IU argument, and it requires the accumulation of theoretical and empirical support for the appropriateness of the claims (Kane, 2006). Each inference in the validity argument is based on a proposition or claim that requires support. The validity argument entails an overall evaluation of the plausibility of the proposed interpretations and uses of test scores by providing a coherent analysis of the evidence for and against the proposed interpretations and uses (AERA, APA, & NCME, 1999; Cronbach, 1988; Kane, 1992; Messick, 1989). Cronbach (1988) argued that the logic of an evaluation argument should provide the foundation for the validation of score interpretations and uses. The specification of a validity argument allows for the accumulation of evidence not only for, but also against, intended score interpretations and uses.

Kane (2006) provided three criteria for the evaluation of IU arguments, including clarity, coherence, and plausibility. A clear IU argument is stated as a framework for validation in that "the inferences to be used in getting from the observed performance to the proposed conclusions and decisions, as well as the warrants and backing supporting these inferences, should be specified in enough detail to make the rationale for the proposed claims apparent" (Kane, 2006, p. 29). A coherent argument logically links the network of inferences from performance to conclusions and decisions, including the actions resulting from the decisions. The plausibility of the argument emphasizes that the assumptions underlying the assessment and score inferences should be credible and judged in terms of supporting and conflicting evidence.

As suggested by Cronbach (1988), researchers have provided advice on prioritizing validity questions (Kane, 2006; Lane & Stone, 2002; Shepard, 1993, 1997). Shepard (1993) proposed three questions to help prioritize validity evaluation:

> "What does the testing practice claim to do?,
> What are the arguments for and against the intended aims of the test?
> What does the test do in the system other than what it claims, for good or bad?" (p. 429).

These questions can be used as a guide in crafting the validity argument for an assessment and accountability system, and unequivocally they indicate that the consequences of the system are integral to the validity argument.

Consequences in the argument-based approach to validity

Considering consequences in the evaluation of validity is not new although it is still debated by scholars (Popham, 1997; Mehrens, 1997; Cizek, 2012). Decades ago Cronbach (1971) considered the evaluation of decisions and actions based on test scores as part of validity evaluation. The soundness of test-based decisions has been an integral aspect of validity, resulting in the need to attend to consequences in the validity framework (AERA, APA, & NCME, 1999; APA, AERA, & NCME, 1985, Cronbach, 1971; Shepard, 1997). The inclusion of the evaluation of the soundness of decisions based on test scores warrants the need to examine consequences which are a "logical part of the evaluation of test use, which has been an accepted focus of validity for several decades" (Shepard, 1997, p. 5).

Claims have been made for examining consequences for both placement tests and tests used as a policy tool for improving instruction and student learning, and evaluating educational programs. When discussing the validity evidence needed for an algebra test used for differential placement, either placement into a calculus course or a remedial algebra course, Kane (1992) indicated that an important claim of the argument is that "the remedial course is effective in teaching the algebraic skills used in the calculus course" (p. 532). In his discussion on how an IU argument framework provides a focus on the intended use of a test, he argued that providing evidence of success in algebra skills and success in the subsequent intended course is a central aspect of the validity argument. Linn (1997) also argued that consequences are integral to validity evaluation for tests that claim differential placement, and stated that their evaluation "demands attention to plausible negative effects as well as the putative benefits of the classification" (p. 15). Such negative effects can include some students not having the opportunity to engage in critical thinking skills. It is evident that a fundamental feature of a validity evaluation is to evaluate test-based decisions in terms of their consequences or outcomes.

Similarly, validation efforts need to attend to consequences when evaluating decision procedures for educational assessment programs that are used as tools for policy. The impetus for state assessment and accountability systems in the United States is to improve the educational opportunities afforded to students so as to improve their learning, and therefore integral to validity evaluation of these systems is the appraisal of test-based decisions in terms of their consequences. As Cronbach (1982) indicated, to evaluate a testing program as an instrument of policy, the evaluation of consequences is essential. The values inherent in the testing program need to be made explicit and consequences of decisions made based on test scores must be evaluated (Kane, 1992). Both positive and negative consequences of test-based accountability programs typically have different impacts on different groups of students and in different schools, and these impacts need to be examined as part of the validity argument (Lane & Stone, 2002). In discussing whether the evaluation of unintended consequences, such as adverse impact, should be integral to a validity argument, Kane (2013) argued that consequences that have "potential for substantial impact in the population of interest, particularly adverse impact and systemic consequences (p. 61)", should be subsumed in the validity research agenda.

For end of course tests, which are intended to increase the rigor of high school courses, consequential evidence that should be

examined include changes in the rigor and depth of the courses and instruction, uniformity of instruction across schools, student course taking patterns, and student dropout rates (Linn, 2009). As indicated by Linn (2009), the impact may be different for tests that are used for graduation requirements than for tests that only contribute partly to course grades. When end of course tests are not required, some students may not take the course and as a result will not be exposed to more rigorous content. Whereas, when end of course tests are required for graduation, they may have an impact on drop out and graduation rates. Another unintended negative consequence for K-12 assessments that has been documented is the narrowing of instruction by some teachers to those topics measured by the assessments (Stecher, 2002). Further, studies examining the effects of state assessment and accountability programs have reported the occurrence of "educational triage," where educators' focus is on students slightly below the cut point for proficient (Booher-Jennings, 2005). As Kane (2013) argued, these types of unintended consequences, both adverse impact and systemic consequences affecting the quality of educational opportunities for students, are integral to the validity argument for assessment and accountability programs.

The validity framework presented by Messick (1989) explicitly identifies social consequences as an integral aspect of the validity argument. In addressing the intent of assessment and accountability systems, Haertel (1999) remarked that the distinction between intended consequences and social consequences is not clear because their primary purpose is to improve educational outcomes for all students. Cronbach (1988) considered consequences as prominent in the evaluation of validity by suggesting that negative consequences could invalidate test use even if they were not due to test design flaws. When examining the impact of assessment programs designed for school accountability, the consequences should be evaluated with respect to not only the assessment, but also the accountability program in its entirety. The validity of the system as a whole needs to be evaluated in terms of its effects on improving instruction and student learning, and therefore a fundamental aspect of validation is to examine whether these benefits are an outcome of the use of the system. When specific outcomes are central to the rationale for the testing program, the evaluation of the consequences is central to the validity argument.

Linn (1993) argued that the need for consequential evidence in support of the validity argument is "especially compelling for performance-based assessments… because particular intended consequences are an explicit part of the assessment system's rationale" (p. 6). Examples of such intended consequences include improved instruction in student engagement in problem solving and reasoning. In addressing the consequences of performance assessments, Messick (1992) argued that evidence should address both the intended consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of fairness, and that it "should not be taken for granted that richly contextualized assessment tasks are uniformly good for all students… [because] contextual features that engage and motivate one student and facilitate effective task functioning may alienate and confuse another student and bias or distort task functioning" (p. 25).

Consistent with earlier editions, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) reinforces consequences as an integral aspect of validity in that "validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (p. 9). Neither the test nor the scores are validated, but instead the score interpretations and uses are validated. In the validity chapter of the *Standards* (AERA, APA, &NCME, 1999), the standards explicitly address the consequential aspect of validity for educational assessments that are considered to be tools for improving instructional practice. As an example, Standard 1.22 states:

> When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence. (p. 23).

Standard 1.23 calls for the accumulation of validity evidence for the intended consequences of an assessment as well as any potential unintended consequences:

> When a test use or score interpretation is recommended on the grounds that testing or the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Due weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted. (p. 23)

The comment associated with Standard 1.23, further states that "certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices or would clarify students' understanding of the kind or level of achievement they were expected to attain" (p. 23) and therefore, "..the extent that such claims enter into the justification for a testing program, they become part of the validity argument for test use and so should be examined as part of the validation effort" (p. 13).

In the chapter on educational testing and assessment, Standard 13.1 supports the need to consider consequences in the validity argument for educational assessments:

> When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user. (p. 145).

The *Standards* (1999) further reinforces Messick's (1999) claim that evidence about consequences is also relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct irrelevant components. Standard 1.24 in the validity chapter states:

> When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure to represent the intended construct.

The comment for Standard 1.24 indicates that it is important to examine whether any unintended consequences "arise from such sources of invalidity" (AERA, APA, & NCME, 1999, p. 23) such as construct-irrelevant components or construct underrepresentation.

The *Standards* (1999) clearly state the need for validity arguments for assessment and accountability systems that are used as policy instruments to include an evaluation of their effects on improving instruction and student learning. Although not published yet, the next revision of the *Standards* will continue to argue for the need for consequential evidence for educational assessments. The evaluation of the consequences of an assessment and accountability program should also consider relevant contextual variables (Lane, Parke, & Stone, 1998). As an example, school contextual variables, such as teacher and principal stability, access to quality curriculum and resource materials, access to professional development support, and socio-economic status, could enhance the interpretation of the consequential evidence of an assessment and accountability system. As summarized by Linn (1997), "the evaluation of consequences rightly belongs in the domain of validity" (p. 16) and the " best way of encouraging adequate consideration of major intended positive effects and plausible unintended negative effects of test use is to recognize the evaluation of such effects as a central aspect of test validation" (p. 16).

Validity evaluation for consequences

A framework for the evaluation of intended and unintended consequences associated with assessment and accountability programs was proposed by me and my colleagues at the University of Pittsburgh (Lane, Parke, & Stone, 1998; Lane & Stone, 2002). This framework evolved during our research project that examined the consequences of the Maryland State Performance Assessment Program (MSPAP). MSPAP was comprised entirely of performance tasks that were integrated across disciplines and required reasoning and problem solving skills. It was intended to promote performance-based instruction and classroom assessment, and to provide opportunities for students to be engaged in higher level thinking, resulting in higher levels of student achievement. In our framework the intended outcomes of an assessment and accountability system were organized into a set of IU arguments, from which a set of propositions or claims were generated that could be either supported or refuted by logical and empirical evidence. As indicated by the *Standards* (AERA, APA, & NCME, 1999) a set of claims that support the proposed interpretation of the particular purpose of the assessment needs to be delineated in framing an argument. When an assessment and accountability program is used to improve educational practices and student learning by influencing curriculum and instruction and to hold schools accountable as they were in the U.S. in the early 1990's, five claims were considered to be fundamental to the IU argument (Lane & Stone, 2002, p. 26):

> 1. School administrators and teachers are motivated to adapt the instruction and curriculum to the standards. 2. Professional development support is being provided. 3. Instruction and curriculum will be adapted. 4. Students are motived to learn and put forth their best efforts. 5. Improved performance is related to changes in instruction.

The validation process evolves as these claims are identified and evidence is gathered to evaluate their soundness. Evidence that could be collected to support or refute each of the claims, potential data sources, and potential stake-holders that could provide consequential evidence were outlined. Evidence that could be collected to evaluate the consequences of state assessment and accountability programs included (Lane & Stone, 2002, p. 24).

- Student, teacher and administrator motivation and effort.
- Curriculum and instructional content and strategies.
- Content and format of classroom assessments.
- Improved learning for all students.
- Professional development support.
- Use and nature of test preparation activities.
- Student, teacher, administrator, and public awareness and beliefs about the assessment; and criteria for judging performance, and the use of assessment results.

An evaluation of the intended effects is not sufficient however. The unintended, potentially negative consequences, should also be examined including:

- Narrowing of curriculum and instruction to focus only on the specific standards assessed and ignoring the broader construct reflected in the specified standards.
- The use of test preparation materials that are closely linked to the assessment without making changes to instruction.
- The use of unethical test preparation materials.
- Inappropriate or unfair uses of test scores, such as questionable practices in reassignment of teachers or principals; and
- For some students, decreased confidence and motivation to learn and perform well on the assessment because of past experiences with assessments (Lane & Stone, 2002, p. 24).

Data sources that can be used to obtain consequential evidence include teacher, student and administrator surveys; interviews and focus groups that probe more deeply into survey results; instructional logs; as well as more direct measures such as instructional artifacts, including instructional tasks and classroom assessments, and classroom observations. The latter two sources of evidence, instructional artifacts and classroom observations, complement the other sources, providing richer data with higher fidelity. As an example, the alignment between instruction and the content standards measured by the assessment can be evaluated by both the use of surveys and by the collection of classroom artifacts such as classroom assessment and instruction tasks and test preparation materials. In addition, changes in instruction as evidenced through classroom artifacts and survey data can be related to changes in performance on assessments (Lane & Stone, 2002).

Research on the consequences of large-scale performance assessments in the 1990's

The renewed interest in performance assessments in the United States in the early 1990's was, in part, because performance assessments were considered to be valuable tools for educational reform. Performance assessments help shape sound instructional practice by providing an indicator to teachers of what is important to teach and to students of what is important to learn (Lane & Stone, 2006). Research studies have demonstrated that the implementation of large-scale performance assessments in the

1990's were related to positive changes in instruction and student learning, with a greater emphasis on problem solving and critical thinking skills. In providing validity evidence to support the interpretative argument and the intended outcomes of MSPAP, my colleagues and I demonstrated that there was a positive impact of the assessment and accountability system on both student learning and classroom instruction (Lane, Parke, & Stone, 2002; Parke & Lane, 2008; Parke, Lane, & Stone, 2006; Stone & Lane, 2003), which were the intended outcomes of MSPAP. Our results indicated that most mathematics instructional activities were aligned with MSPAP and the state content standards, however, the classroom assessments had a weaker alignment with MSPAP and the standards (Parke & Lane, 2008). Teacher reported use of performance-based instruction accounted for differences in school performance on MSPAP in reading, writing, math and science. In particular, schools focusing on more performance-based instruction, such as the engagement of students in critical thinking and reasoning skills, had higher MSPAP scores than schools in which their instruction was less performance-based (Lane, Parke, & Stone, 2002; Stone & Lane, 2003). Further, the more impact MSPAP had on instruction, including more of a focus on higher level thinking skills and rigorous content, the greater rates of change in MSPAP school performance in mathematics and science over a five year period (Lane, Parke, & Stone, 2002; Stone & Lane, 2003). The MSPAP results pertaining to mathematics was supported by a study by Linn, Baker, and Betebenner (2002) that demonstrated that trends in mathematics student gains for NAEP and MSPAP math assessments were similar, indicating that increased performance on MSPAP was a result of actual gains in student achievement in mathematics across the school years. Such positive results may have resulted, in part, from schools using MSPAP data along with other information to guide instructional planning (Michaels & Ferrara, 1999).

When using test scores to make inferences regarding the quality of education, contextual information is needed to inform inferences and actions (Haertel, 1999). In the MSPAP studies, a school contextual variable, SES measured by percent of students receiving free or reduced lunch, was significantly related to school level performance on MSPAP in math, reading, writing, science, and social studies (Lane, Parke, & Stone, 2002; Stone & Lane, 2003). More importantly, SES was not significantly related to school growth on MSPAP in math, writing, science and social studies. It may be argued that these results indicate that there was no adverse impact of MSPAP for students living in economically disadvantaged areas with respect to school growth on the assessment.

Other state assessment programs that included performance-based tasks have provided evidence of the impact of its assessment on instruction and student learning. Teachers in Vermont reallocated instruction time to reflect the goals of the Vermont Portfolio Assessment in math and writing, such as allocating more time to problem-solving and communication in math and providing students opportunity to engage in extended writing projects (Stecher & Mitchell, 1995; Koretz, Baron, Mitchell, & Stecher, 1996). Teachers in Washington reported that both short-response items and extended-response items on the Washington Assessment of Student Learning were influential in improving instruction and student learning (Stecher, Barron, Chun, & Ross, 2000), and based on observations, interviews and classroom artifacts in a subset of the schools it was found that teachers used math and writing scoring rubrics in instruction in a way that reinforced meaningful learning (Borko, Wolf, Simone, & Uchiyama, 2001). In Kentucky, teachers began using more performance-based activities in math (Borko & Elliott, 1999; Koretz et al., 1996) and writing (Wolf & McLever, 1999) that were aligned to the Kentucky performance-based assessment system. However, based on teacher interviews and classroom artifactsfrom a small number of Kentucky schools it was found that while teachers implemented new instructional strategies, the depth and complexity of content covered did not change in fundamental ways (McDonnell & Choisser, 1997). In examining the relationship between improved instruction and gains in Kentucky school-level scores, Stecher and his colleagues (Stecher, Barron, Kaganoff, & Goodwin, 1998) found inconsistent findings across disciplines and grades. There was a positive relationship however between standards-based instructional practices in writing and the Kentucky direct writing assessment. For example, more 7th grade writing teachers in high-gain schools versus low-gain schools reported integrating writing with other subjects and increasing their emphasis on the process of writing.

Within a mathematics education reform project, the relationship between the presence of reform features of mathematics instruction and student performance on a mathematics performance assessment was examined for schools that served students in economically disadvantaged urban areas (Stein & Lane, 1996). Extensive observations were conducted in the classrooms to examine the quality of mathematics instruction and student engagement. The analyses of instruction focused on the cognitive demands of the instructional tasks as represented in the instructional material, as set up by the teacher in the classroom, and as implemented by students. The greatest student gains on the performance assessment were observed for those classrooms in which the instructional tasks were set up and implemented with high levels of cognitive demands so that students were engaged in using multiple solution strategies and multiple representations, and they were adept at explaining their mathematical thinking.These classroom teachers encouraged non-algorithmic forms of thinking associated with the doing of mathematics. Whereas, the smallest gains were observed in classrooms when instructional tasks were procedurally based and could be solved by a single, easily accessible strategy, and required little or no mathematical communication.

Theory of action and consequential evidence

Underlying program evaluation efforts are theories of action. A theory of action provides a framework for evaluating programs by identifying critical program components and their logical points of impact. A theory of action for program evaluation may include the context in which a program is being implemented; a description of the components of the program; what the program components intend to achieve and how they interact; and short term and long term outcomes. Assumptions or claims that underlie the actions are specified in order to examine the extent to which the proposed activities bring about desired changes (Weiss, 1997).

Given that validation is an evaluation argument (Cronbach, 1988), a theory of action can be used to guide the development of a comprehensive IU and validity argument for an assessment and accountability system by fleshing out the validity evaluation plan. A theory of action for assessment and accountability systems serves as a vehicle to develop a comprehensive validity argument, and should include the rationale for the use of the assessment,

the mechanisms that will lead to the intended outcomes, and the intended outcomes. Studies are needed not only to support the theory of action and the validity argument but also to identify contradictions to the specified claims. Validity evidence to support or refute the claims can be collected based on content, internal structure, response processes, relations with other variables, as well as consequences as outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Consistent with Kane's argument-based approach to the validation of assessments, Bennett (2010) indicated that a theory of action for an assessment and accountability system provides a framework for bringing together score meaning and impact claims, with an emphasis on claims about the intended consequences or outcomes of the assessment on both individuals and schools. The delineation of the claims about the intended consequences within a theory of action is integral to the validity argument. Bennett (2010) proposed that the following features be included in a theory of action for an assessment system:

- The intended effects of the assessment system.
- The components of the assessment system and a logical and coherent rationale for each component, including backing for that rationale in research and theory.
- The interpretative claims that will be made from assessment results.
- The action mechanisms designated to cause the intended effects.
- Potential unintended negative effects and what will be done to mitigate them (Bennett, 2010, p. 71).

An initial theory of action was delineated for the Title 1 law and standards-based reform in the late 1990's (Elmore & Rothman, 1999). The three key components that were included were standards, assessment, and accountability. The ultimate outcome, higher levels of student learning, was mediated by two claims or action mechanisms: 1) clear expectations for students and schools and 2) motivation to work hard. The theory of action as described by Elmore and Rothman (1999) is provided below (see figure 1).

Elmore and Rothman (1999) expanded on this theory of action to reflect features of effective reform. It was evident that the theory of action needed to address both professional development and improved teaching as key mechanisms to enhance student learning. Their revised theory of action reflects the view that standards-based policies affect student learning only if they are linked to efforts that build teacher capacity to improve instruction for students (Elmore & Rothman, 1999) (see figure 2).

The revised theory of action proposed by Elmore and Rothman (1999) explicitly includes action mechanisms that would have an impact on student learning. These action mechanisms are embedded within the validity argument and require evidence to support the claims.

Bennett (2010) proposed a preliminary theory of action for the Cognitively Based Assessment of, for, and as Learning (CBAL) research initiative for developing a model of an innovative K-12 assessment system that "documents what students have achieved (of learning); facilitates instructional planning (for learning); and is considered by students and teachers to be a worthwhile educational experience in and of itself (as learning)" (pp. 71-72). The theory of action included the major components of CBAL, and both intermediate intended effects and ultimate effects. The four main components of CBAL are: Domain-specific competency models and learning progressions, summative assessments distributed across the school year and accumulated for accountability purposes, formative assessment, and professional development support. The intermediate intended effects included:

- A clearer, deeper, and better organized understanding on the part of teachers of the content domain in which they teach.
- An increase in teachers' pedagogical knowledge and assessment skill.
- Greater focus in classroom instruction on integrated performances and other activities intended to promote the development of higher-order skills…
- The routine use of formative assessment practices in the classroom to make appropriate adjustments to instruction; and
- Improved student engagement in learning and assessment. (Bennett, 2010, p. 73).

The ultimate effects included:

- Improved student learning with respect to content standards and
- More meaningful information for policy makers regarding the effectiveness of education at the school, district, and state levels, leading to decisions that facilitate learning. (Bennett, 2010, p. 73).

Provisional interpretative claims, specified at the individual, class, district, and school level, that will require evidence to support them were also specified by Bennett (2010). As an example, three of the five provisional interpretative claims for summative assessments included:

- Aggregated student performance on periodic assessments represents achievement of content standards.
- Students who perform at a designated level of proficiency are ready to proceed to the next grade's work.
- Flagged content areas, class groups, and individual students should be followed-up through classroom assessment because they are likely to need attention. (Bennett, 2010, p. 80).

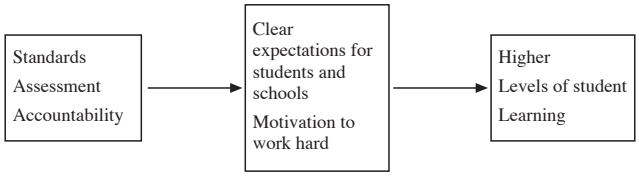

Figure 1. *Initial theory of action for title I*

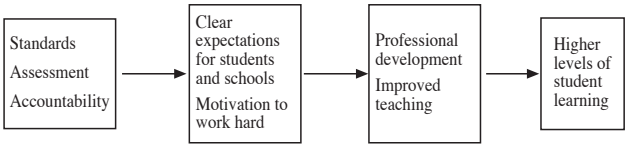

Figure 2. *Revised theory of action for title I*

A set of hypothesized action mechanisms that connect the CBAL components to the intended effects were also identified. Bennett (2010) further stated that "adoption and use of the system components is hypothesized to *cause* those effects" (p. 80). It is evident from this theory of action that when an assessment system has intended outcomes, validation efforts should examine whether these intended outcomes were realized.

The U.S. Department of Education Race to the Top initiative (Department of Education, 2009) calls for state assessment systems to be grounded in academic standards that reflect 21st century skills, including higher level thinking skills that are typically difficult to measure. The Common Core State Standards (CCSS; CSSO, & NGA, 2010) are intended to reflect 21st century skills, representing a set of expectations for knowledge and skills students need so they are prepared for success in college and careers when they graduate from high school. The CCSS emphasize students' ability to reason, synthesize information from various sources, think critically, and solve challenging problems. Under the Race to the Top Assessment Program, the U.S. Department of Education required state consortia to specify a theory of action, including the intended benefits of their assessment and accountability systems (Department of Education, 2010). The call required a "theory of action that describes in detail the causal relationships between specific actions or strategies… and its desired outcomes… including improvements in student achievement and college- and career-readiness" (Department of Education, 2010, p. 18174). The criteria that was used for evaluating the theory of action focused on "the extent to which the eligible applicant's theory of action is logical, coherent, and credible, and will result in improved academic outcomes" (Partnership for Assessment of Readiness for College and Careers, 2010, p. 34).

It is expected that the implementation of the Common Core State Standards and the next generation of state assessments based on the CCSS will lead to improved educational opportunities for students and as a result, improved student learning. In order to evaluate these intentions it is necessary to specify how they will occur, and the mechanisms that will bring about these outcomes. In crafting a theory of action, the intended purposes of the assessments need to be made explicit. As an example, the fundamental goal of the Partnership for Assessment of Readiness for College and Careers (PARCC, 2010) is to increase the rates at which students graduate from high school prepared for success in college and the workplace. The intended purposes of the PARCC assessment that will guide their theory of action are:

1. Determine whether students are college- and career ready or on track.
2. Assess the full range of the Common Core State Standards, including standards that are difficult to measure.
3. Measure the full range of student performance, including the performance of high and low performing students.
4. Provide data during the academic year to inform instruction, interventions and professional development.
5. Provide data for accountability, including measures of growth.
6. Incorporate innovative approaches throughout the system (PARCC, 2012).

Inclusion of students with severe cognitive disabilities in alternate assessments intended to improve learning for those students requires evidence of the impact on instruction for these students and effects on their learning. Marion and Perie (2009) argued that consequential evidence is of particular importance for evaluating the validity for alternate state assessments because these assessments provide a mechanism to promote grade-level academic instruction for students who are typically underserved. The theory of action for the National Center and State Collaborative (NCSC) Alternate Assessments treats the summative assessment as a component of the overall system, and must be considered in light of the other system components (e.g., professional development for teachers, appropriate communication methods for the student and teacher, instruction aligned to grade-level content standards) when evaluating the system goals (Quenemoen, Flowers, & Forte, 2013).

Incorporating the claims and outcomes in the next generation of assessment programs, including the alternate assessments, into a comprehensive validity argument will facilitate evaluating the consequences of the assessment systems. Some of the intended consequences and unintended, negative consequences that will need to be examined based on their theories of actions include:

- Teacher engagement in professional development, instruction and student learning.
- Student engagement in learning and engagement on the assessment.
- Teacher and administrator use of assessment results to inform instruction.
- Changes in curriculum, instruction, and classroom assessment (innovative instructional techniques, alignment with CCSS, focus on problem solving and reasoning, etc.).
- The relationship between changes in instruction and changes in student performance on the assessments.
- Impact on subgroups (students with disabilities, ELLs, and minority subgroups) with respect to improving instruction and narrowing achievement gaps.
- Use of assessment results for evaluating the system (e.g., changes in teachers' course patterns).
- Use of assessment results for educator accountability.
- Impact on college readiness – overall and for subgroups:
  - HS graduation rates.
  - College admission patterns.
  - Remedial course patterns.

Haertel's (2013) framework for classifying mechanisms of intended testing effects as direct effects and indirect effects can help clarify our thinking on the validity evidence needed for assessment and accountability systems. The direct effects of educational assessments, instructional guidance for students, student placement and selection, informing comparisons among educational approaches, and educational management (e.g., the use of assessments to help evaluate the effectiveness of educators or schools), involve interpretations or uses that rely directly on the information scores provide about the assessed constructs. Whereas, indirect effects, including directing student effort, focusing the system (i.e., focusing curriculum and instruction), and shaping public perceptions that have an impact on actions, have no direct dependence on the information provided by test scores, but are linked closely to the purposes or claims of assessment (Haertel, 2013). These indirect mechanisms of action, which are key components of the interpretative and use argument, are critical

in the evaluation of consequences of educational assessments and accountability programs. Further, the potentially negative unintended consequences tend to be embedded within these indirect effects (Haertel, 2013). The use of Haertel's framework provides a foundation for a comprehensive, principled approach to studying both the intended and unintended consequences of an assessment system.

Concluding thoughts

A research agenda for validity evidence based on testing consequences should be a priority for educational assessments, in particular, assessment and accountability programs that are used as policy tools for improving instruction and evaluating educational programs. A theory of action that encompasses the IU argument and a validity argument for educational assessment and accountability programs can explicitly address both the direct and indirect mechanisms of action. Within a theory of action the purposes and uses of the system and the outcomes of the system, including the mediating outcomes necessary to achieve the ultimate outcomes are delineated. The key components of the theory of action can then be prioritized and further delineated so as to support assessment design and the validity argument. Because assessment and accountability systems are intended to have an impact on both instruction and student learning it follows that the consequences, both positive and potentially negative, of the systems are integral to the theory of action and validity argument.

Author note

This article draws heavily on two of my papers, *The Interplay Among Validity, Consequences, and a Theory of Action*, presented at the 2012 annual meeting of the National Council on Measurement in Education, Vancouver, Canada, and *Impacts on Classroom Instruction and Achievement in the 1990's and Implications for the Next Generation of Assessments*, presented at the 2013 annual meeting of the National Council on Measurement in Education. I would like to thank Steve Sireci for his thoughtful comments on an earlier version of this paper and José-Luis Padilla for his help with the Spanish translation.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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.

Bennett, R. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70-91.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Borko, H., & Elliott, R. (1999). Hands-on pedagogy versus hands-off accountability: Tensions between competing commitments for exemplary math teachers in Kentucky. *Phi Delta Kappan, 80,* 394-400.

Borko, H., Wolf, S.A., Simone, G., & Uchiyama, K. (2001). *Schools in transition: Reform efforts in exemplary schools of Washington*. Paper presented at the annual meeting of the American educational Research Association, Seattle, WA.

Cizek, G.J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

Council of Chief State School Officers and National Governors Association (2010). *Common Core Standards for English Language Arts*. Retrieved on June 25, 2010 from www.corestandards.org.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17), Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), *Intelligence: Measurement theory and public policy,* (pp. 147-171). Urbana, IL: University of Illinois Press.

Department of Education (April, 2010). Overview information: Race to the Top fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register, 75*(68), 18171-18185.

Elmore, R.F., & Rothman, R. (1999). *Testing, teaching, and learning: A guide for states and school districts*. National Academies Press: Washington, DC.

Haertel, E. (1999). Performance assessment and education reform. *Phi Delta Kappan*, *80*(9), 662-667.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives, 11*(1-2), 1-18.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th edition, pp. 17-64). Washington, DC: American Council on Education/Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(*1), 1-73.

Koretz, D., Barron, S., Mitchell, M., &Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS*). Rand Corporation.

Lane, S., Parke, C.S., & Stone, C.A. (1998). A framework for evaluating the consequences of assessment programs, *Educational Measurement: Issues and Practice, 17*(2), 24-28.

Lane, S., Parke, C.S., & Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on Mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*(4), 279-315.

Lane, S., & Stone, C.A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational Measurement*. New York: American Council on Education &Praeger.

Lane, S., & Stone, C.A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and* Practice, *21*(1), 23-30.

Lane, S., Parke, C.S., & Stone, C.A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24-28.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-6.

Linn, R.L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 21*(1), 14-16).

Linn, R.L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity* (pp. 195-212).Charlotte, NC: Information Age Publishers.

Linn, R.L, Baker, E.L., & Betebenner, D.W. (2002). Accountability systems: Implications of the requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 20*(8), 15-21.

Marion, S. (April 16, 2010). *Developing a Theory of Action: A foundation of the NIA Response*. Retrieved on August 2, 2010 from http://www.nciea.org/papers-TheoryofAction_041610.pdf.

Marion, S.F., &Perie, M. (2009). An introduction to validity arguments for alternate assessments. In W.D. Schafer & R.W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 321-334). Baltimore, MD: Brookes.

McDonnell, L.M., &Choisser, C. (1997). *Testing and teaching: Local imp mentation of new state assessments*. CSE Technical Report 442. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mehrens, W.A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 21*(1), 16-18.

Messick, S. (1989).Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Michaels, H., & Ferrara, S. (1999). Evolution of educational reform in Maryland: Using data to drive state and policy local reform. In G.J. Cizek (Ed.), *Handbook of Educational Policy.* San Diego, CA: Academic Press.

Parke, C.S., & Lane, S. (2008). Examining alignment between state performance assessment and mathematics classroom activities. *Journal of Educational Research*, *101*(3), 132-147.

Parke, C.S., Lane, S., & Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12*(3), 239-269.

Partnership for Assessment of Readiness for College and Careers (June 23, 2010). *Application for the Race to the Top Comprehensive Assessment Systems, Competition*.Retrieved on September 18, 2010 from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf.

Partnership for Assessment of Readiness for College and Careers (2012). *PARCC Assessment Design*.Retrieved on March 1, 2012 from http://www.parcconline.org/parcc-assessment-design.

Popham, W.J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 21*(1), 9-13.

Quenemoen, R., Flowers, C., & Forte, E. (2013, April). *Theory of action for the National Center and State Collaborative Alternate Assessments and its role in the overall assessment design*. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA.

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education*, *19*(1), 405-450.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 21*(1), **5-8,13.????**

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. In L.S. Hamilton, B.M. Stecher & S.P. Klein (Eds.), *Making sense of test-based accountability*. Santa Monica, CA: RAND.

Stecher, B., & Mitchell, K.J. (1995). *Portfolio driven reform: Vermont teachers' understanding of mathematical problem solving*. CSE Technical Report 400. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000, August). *The effects of the Washington state education reform in schools and classrooms* (CSE Tech. Rep. NO. 525). Los Angeles: Center for Research on Evaluation, Standards and Student Testing.

Stecher, B., Barron, S., Kaganoff, T., & Goodwin, J. (1998, June). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing* (CSE Tech. Rep. No. 482). Los Angeles: University of California: National Center for Research on Evaluation, Standards, and Student Testing.

Stein, M.K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50-80.

Stone, C.A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1-26.

U.S. Department of Education (2009). *Race to the Top Program Executive Summary* Retrieved on April 15, 2010 from http://www.ed.gov/programs/racetothetop/resources.html.

Weiss, C.H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review, 21*(4), 501-524.

Wolf, S.A., & McIver, M. (1999). When process becomes policy: The paradox of Kentucky state reform for exemplary teachers of writing. *Phi Delta Kappan*, *80*, 401-406.