Language Testing in Asia

RESEARCH Open Access

A validity framework for accountability: educational measurement and language testing

Karen B. Hoeve

Correspondence: kbhoeve@uncg.edu; khoeve@abpeds.org
School of Education, The University of North Carolina at Greensboro, Greensboro, USA

Abstract

High stakes test-based accountability systems primarily rely on aggregates and derivatives of scores from tests that were originally developed to measure individual student proficiency in subject areas such as math, reading/language arts, and now English language proficiency. Current validity models do not explicitly address this use of aggregate scores in accountability. Historically, language testing and educational measurement have been related, yet parallel disciplines. Accountability policies have increasingly forced these disciplines under one umbrella with a common system of rewards and sanctions based on results achieved. Therefore, a validity framework, as suggested in the present paper, is relevant to both.

Keywords: Validity, Validation, Aggregate scores, Accountability, Educational measurement, Language testing

Introduction

Historical and contemporary theories of validity and validation were designed with individual test scores in mind, but in accountability, these scores are aggregated to create a score or index at the school or teacher level. These aggregate scores or indexes are then interpreted in much the same way as an individual score, but at the school level. Sireci and Soto (2016, p. 149) assert that, "Using tests for educational accountability often entails employing the test for purposes beyond which it was originally developed. Like the originally intended purposes, using test scores for accountability purposes also requires evidence and theory to justify their use." Validity is the cornerstone for the use and interpretation of test scores.

Researchers and validity theorists have argued that a comprehensive validity argument needs to include both test development and measurement evidence (e.g., Chalhoub-Deville, 2020; Chalhoub-Deville and O'Sullivan, 2020), documentation of consequences (e.g., Bachman and Palmer, 2010; Chalhoub-Deville, 2016, 2020; Embretson, 2007, 2008, 2017; Kane, 2006, 2013), and consideration for the validation of aggregate-level data in addition to individual student level data (Chalhoub-Deville, 2020, p. 254).

Despite calls for comprehensive validity arguments, neither the measurement nor language testing field has published validity frameworks, models, or theories that fully address the needs of accountability testing. Kane (2006, 2010, 2012, 2013, 2015, 2020) discusses test development, consequences, and accountability in his writing, but his Interpretive Use Argument (IUA) does not explicitly address the needs for validity evidence in these areas. Embretson (2007, 2008, 2017) specifically addresses test development and consequences in her Integrated Framework for Construct Validity but does not discuss aggregate scores and accountability. The predominant models in language testing, such as Bachman and Palmer's (2010) Assessment Use Argument (AUA), while anchored in consequences, do not consider accountability testing and the related validity issues that should be addressed.

The present paper seeks to bring together the publications available in educational measurement and language testing and build on the knowledge they provide. An overview of key validity models and the specific needs of accountability testing, including English language proficiency, is included. An argument is made that validity evidence needs to be gathered during the test development phase, at both the individual student level and the aggregate level, and consideration needs to be given to consequences of accountability policies. Like math and reading/language arts subject area tests, English language proficiency tests are required to be aggregated at the school level for inclusion in the accountability system. Prominent validity theories address consequences but do not consider accountability purposes where the focus is aggregate scores. To address the gaps in prevalent validity models, the IUA is reconceptualized to offer a systematic approach for building a validity argument that begins in the test design and development phase, includes a parallel process for building validity evidence for aggregate scores, and considers the consequences of accountability systems.

English learners

English learners (ELs) figure prominently in accountability. In the United States (US), the Every Student Succeeds Act (ESSA) (2015) requires each state to have a statewide accountability system based on challenging academic standards for reading/language arts and math to improve school success and academic achievement for all students, including ELs. Representing a major shift in the importance of EL test results, the accountability for English language proficiency (ELP) has moved from Title III in the No Child Left Behind (NCLB) Act to Title I in ESSA. Title I is the largest funding allocation in the law and its purpose is "to provide all children significant opportunity to receive a fair, equitable, and high-quality education, and to close educational achievement gaps" (Office of Elementary and Secondary Education, 2021). Prior to ESSA, only those schools receiving Title III funding for EL students and their families were accountable for EL progress.

ESSA requires setting long term goals for the percent of ELs making progress in achieving ELP and requires indicators of this progress in the accountability system. ELs are assessed annually on an ELP test and at the individual student level, these results are used to determine who needs additional help and who may be ready to exit from EL status. These test results are also aggregated for school accountability and therefore validity evidence is needed for the interpretation and use of both the individual and aggregate scores.

Test development argument

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, referred to hereafter as the *Standards*) identify five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and testing consequences. According to the *Standards*, "Content-oriented evidence of validation is at the heart of the process in the educational arena known as alignment, which involves evaluating the correspondence between student learning standards and test content" (p. 15). Sireci and Faulkner-Bond (2014) describe one aspect of content validity as the "appropriateness of the test development process" which "refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material" (p. 101).

A method for test development that facilitates test content evidence for a validity argument is evidence-centered design (ECD). ECD takes interpretation and use claims into account during the test development phase (Mislevy, Steinberg, and Almond, 2003; Plake, Huff, Reshetar, Kaliski and Chajewski, 2015). It is a principled assessment design approach that is engineered towards intended interpretations and uses with explicit design decisions and rationales (Ferrara, Lai, Reilly, and Nichols, 2017). In other words, the building of the validity argument explicitly begins at the design phase of the test development process (Ferrara, Lai, Reilly, and Nichols, 2017; Im, Shin, and Cheng, 2019; Kane, 2015, 2020). "A hallmark of ECD is thus to commence the assessment design process by articulating a chain of reasoning that links evidence to claims about target constructs" (Riconscente, Mislevy and Corrigan 2016 p. 41).

Unit of analysis and consequences in accountability systems

Accountability systems shift responsibility away from students for their performance to holding teachers and schools accountable, and in turn, this shifts the unit of measurement from individual to aggregate scores. With this shift in unit of measure, validity evidence in the traditional individual score unit of measurement now requires consideration of "aggregate and socio-educational consequences" (Chalhoub-Deville, 2020, p. 247). Consequences, also referred to as impact, backwash, and washback, are a subject of discussion and debate among measurement theorists and researchers. The debate is not whether there are consequences in test score interpretation and use, but in whether they fall under the purview of validity, in identifying who is responsible for evaluating them, which ones, and when (Chalhoub-Deville, 2009, 2016). Chalhoub-Deville calls out three levels or units of analysis that are relevant in the interpretation and use of assessments for accountability: individual, aggregate, and educational-social.

Chalhoub-Deville (2016, 2020) observes that traditionally tests have focused on individuals and as such, validity theory has evolved around score use at the individual level. Chalhoub-Deville also observes that accountability testing has moved beyond individual scores to the use of aggregate scores to evaluate teachers and schools. "This aggregated data is a centerpiece of educational reform policies. Aggregated scores are the unit of accountability; this is where validation needs to be anchored" Chalhoub-Deville (2020 p. 253). Chalhoub-Deville (2016, 2020) also argues for the inclusion of consequences in validating accountability testing and further argues that test developers and users have

a shared responsibility in addressing consequences. A validity model that takes accountability systems into account will need to consider the use of aggregate scores during the test development process and consequences of their use for education reform.

Validating the consequences of accountability systems

Messick (1989) proposed that "[v]alidity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 13). Messick included social consequences of test use as part of the definition of the construct validity of score interpretations sparking a controversy that still divides the measurement community (Cizek, 2016; Newton and Shaw, 2014). Researchers such as Lane (1999), Moss (2013), Shepard (2016), Kane (2006, 2013), and Chalhoub-Deville (2009, 2016, 2020) argue that consequences should be considered in validity evidence, while others, like Lissitz and Samuelsen (2007), argue that consequences should not be considered as part of validity evidence. Messick referenced social consequences, but Shepard (2016) further expanded the concept of test consequences to include positive, negative, intended, and unintended consequences as part of score-based inferences.

Consequences of an accountability system are the rewards, sanctions, and interventions imposed on teachers, schools, and districts. Emergent consequences precede rewards and sanctions in anticipation of the possibility that they may be imposed, or they follow the imposed sanctions (Council of Chief State School Officers, referred to hereafter as CCSSO, 2004). Consideration of emergent consequences requires anticipating not only the consequences that may occur after the implementation of the accountability system, but also the consequences that may occur in anticipation of the implementation. Emergent consequences of accountability systems include activities or conditions in the school that may be positive or negative. Examples of positive emergent consequences are improved teaching and learning. Washback, such as narrowing of the curriculum and focusing on test strategies rather than on the knowledge and skills the test intends to measure or decreases in morale because of being identified as a low performing school are examples of negative emergent consequences. Policy interacts directly and indirectly with the consequences of an accountability system. Accountability systems need to be evaluated to ensure that they are achieving intended goals and outcomes, such as closing achievement gaps for ELs, while avoiding potentially negative consequences. The validation plan for an accountability system must analyze both intended and unintended consequences (Kane, 2006; 2013).

Validity frameworks and accountability

This review of validity models focuses on Kane's (2006, 2013) Interpretation and Use Argument (IA/IUA), a Parallel IUA proposed by Acree, Hoeve, and Weir (2016), and Embretson's (2007, 2008, 2017) Integrated Framework for Construct Validity. The extent to which these models consider accountability-testing purposes with regards to aggregate scores, consequences, and consideration of validity in the test development process is analyzed. Finally, a comprehensive model that expands on Acree et al.'s modifications of Kane's IUA is proposed.

Interpretation and use argument (IA/IUA) (Kane 2006, 2013)

An argument-based approach to validity helps to operationalize test validation by providing "a place to start, guidance on how to proceed, [and] criteria for gauging progress and deciding when to stop" (Kane, 2012, p. 8). Kane's IA/IUA offers a roadmap for building a validity argument for trait-based interpretations (Fig. 1). Kane describes "hypothesized empirical relationships," that represent the definition of the trait or construct. His argument-based approach to validity begins with an interpretive argument (IA). The IA specifies the claims or inferences with regards to score use and interpretation. Kane (2006, 2013) identifies four inferences in the IA: scoring, generalization, extrapolation, and implication. The validity argument is an overall evaluation of the claims or inferences being made. Research builds evidence to support the claims or inferences laid out in the IA. Kane concludes that the specified interpretations and uses for test scores are valid if the IA/IUA is complete, coherent, and plausible.

Kane (2006) describes a test development strategy involving three iterative steps: outline an interpretive argument, develop the test, and evaluate the inferences and assumptions in the interpretive argument. While test design considerations are inferred in Kane's "hypothesized empirical relationships," his validity argument does not begin until the scoring inference. As such, his argument-based validity model focuses on score inferences but largely overlooks the test development process in which the trait is defined and the contexts and methods for measurement are considered (Chapelle, 2012; Chalhoub-Deville, 2020). Kane (2006) acknowledged that social consequences of testing were of growing interest and that consequences (positive and negative) play a role in validation. Even though that role is "somewhat contentious" in the field, positive consequences should "outweigh" negative consequences in general (Kane 2006, p. 51). Furthermore, Kane (2006 p. 55) specifically noted that educational reform and accountability call for an evaluation of consequences. "The accountability program is an educational intervention, and a serious evaluation of an accountability program would require an evaluation of both intended and unintended outcomes" (Kane, 2013, p. 54).

Kane's IA/IUA (2006, 2013, 2015, 2020) offers a useful roadmap for the operationalization of validation. However, despite his recognition of the importance of test design, consequences, and score interpretation, Kane's "…model, nevertheless, remains

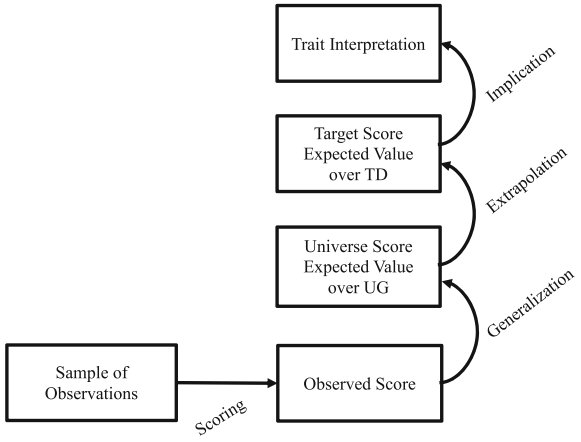

Fig. 1 Kane's IA/IUA

anchored in individual test scores, which does not accommodate accountability testing realities" Chalhoub-Deville (2020).

Parallel IUA (Acree, Hoeve, and Weir, 2016)

While Kane's technique addresses the test score itself, accountability systems present a different, though closely related problem in that the scores in question are aggregate in nature. Building on Kane's validation framework, Acree, Hoeve, and Weir (2016) proposed that Kane's (2006, 2013) IUA for validating uses of individual scores can be extrapolated and expanded for validating uses of aggregate and derivative scores in accountability systems. Individual scores and aggregate scores are similar enough that a common strategy may be used to evaluate the degree to which both are valid. In this framework (Fig. 2. Parallel validation for accountability testing by Acree, Hoeve, and Weir (2016)), individual and aggregate scores are evaluated independently and in parallel. Both branches must be interrogated and interpreted systematically and separately. The validation of accountability systems concerns itself primarily with the group-centered branch.

The parallel, but independent nature of these evaluations maintains that even if strong evidence of validity is established along the individual or student-centered branch, this does not imply the same will hold for the aggregate or group-centered branch. Nor is validity evidence at the individual score level a necessary part of validation for use at an aggregate level. Similarly, this model holds that failure of an inference in the student-centered branch does not necessarily undermine the validity of the group-centered branch. Validity evidence may be found for individual scores, aggregate scores, for both, or for neither.

Integrated framework for construct validity (Embretson 2007, 2008, 2017)

Embretson (2007, 2008, 2017) proposed a validity framework that she described as universal and interactive. According to Embretson (2007), "the system is universal because

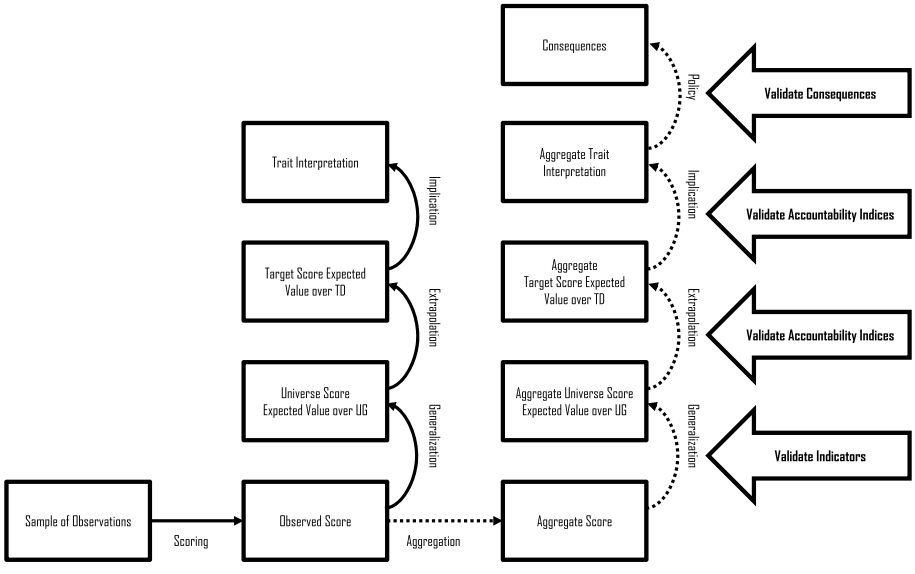

Fig. 2 Parallel validation for accountability testing by Acree, Hoeve, and Weir (2016)

all sources of evidence are included and may be appropriate for both educational and psychological tests" and "interactive because the adequacy of evidence in one category is influenced or informed by adequacy in the other categories" (p. 452). Embretson's validity framework is divided into internal and external aspects of construct validity. Test development processes are characterized as internal aspects of validity. Practical constraints, latent process studies, a conceptual framework, and psychometric analysis directly inform item design and test specifications; and in turn, test specifications inform scoring models. Embretson includes the five sources of evidence defined in the *Standards* (2014) in her framework, i.e., test content, response processes, internal structure, relationships to other variables, and consequences with the latter two being external aspects of construct validity.

To summarize, Embretson (2017) emphasizes test development processes as essential internal aspects of her validity framework and identifies the need for a conceptual framework (like that offered by evidence-centered design). Embretson (2007, 2008, 2017) advocates for including categories of evidence that would be evaluated during the test development cycle as part of her validity system. She includes categories of evidence for practical constraints (e.g., test administration methods and scoring mechanisms); item design principles (e.g., formats, context, complexity, and specific content); domain structure (specification of content areas and levels); and test specifications (e.g., blueprints).

Impact or consequences are also explicitly included as an aspect of Embretson's (2007, 2008, 2017) validity framework. Embretson describes concern for differential item functioning among groups and the potential impact on selection or placement at the individual level. She also recognizes a role for test developers in consequences saying, "there may be aspects of test specifications and item design that could be changed to reduce impact" (2017, p. 108).

Of the validity models reviewed, Embretson's (2017) framework is the most comprehensive. However, aggregate scores and the potential consequences of their use in accountability systems are not specifically addressed in Embretson's validity framework (2007, 2008, 2017).

Proposed validation framework: accountability IUA

Historically, validation approaches have been proposed for study at the individual score level where the test user wants to draw an inference about an individual test taker (e.g., placement testing, achievement testing). Accountability systems use group-level aggregate test scores and derivatives of test scores to draw conclusions about schools and teachers. Neither historical nor current theories of validity and validation explicitly address the use of group-level test scores, as used in accountability systems.

To address the omission of validation during the test development phase of an accountability system, the Parallel IUA framework (Fig. 2) proposed by Acree et al. (2016) for validating accountability systems has been expanded into an accountability IUA framework (Fig. 3) for accountability systems that includes a test-centered branch. The proposed Accountability IUA validation framework (Fig. 3) offers a roadmap for building a validity argument that promotes consideration of validity in the test development phase, outlines a parallel process for building validity evidence for aggregate

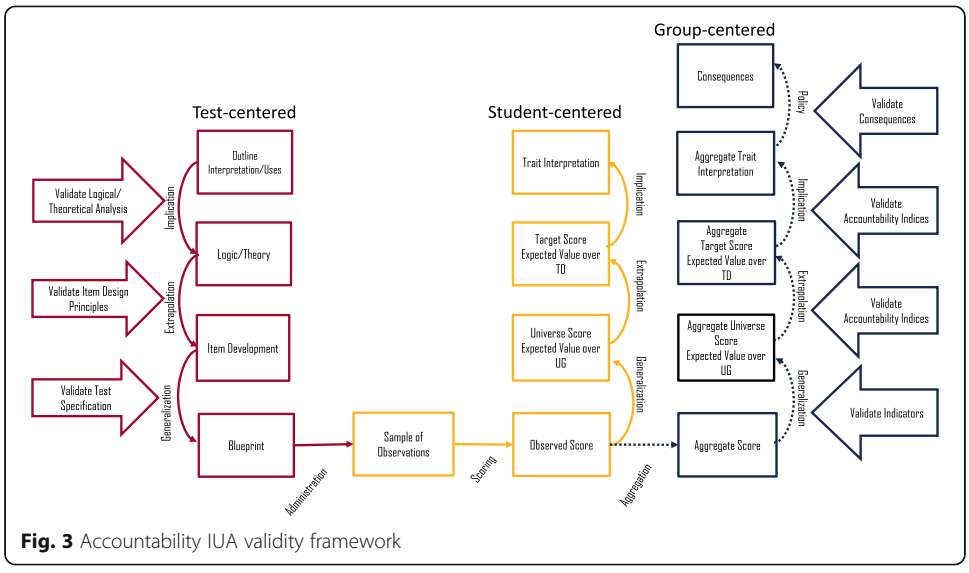

Fig. 3 Accountability IUA validity framework

scores in addition to individual scores, and considers the consequences of aggregate score use in test-based accountability systems.

In the proposed Accountability IUA framework, Kane's (2006, 2013) generalization, extrapolation, and implication inferences occur in the reverse order during the test-centered (test development) branch. In a sense, test development involves beginning with the end in mind, so the logical progression of building validity arguments is reversed. The intended interpretations and uses are defined and evidence gathered for the logical and theoretical analysis of the implications (the decision or action being taken from the scores). Item design principles, such as ECD provide evidence for extrapolation (using the scores as a reflection of real-world performance).

Validation of the test specifications supports generalization (using the scores as a reflection of performance in the test environment). Once the test administration is operationalized, building evidence for the student-centered and group centered IUA branches begins. Evidence must be compiled to support the validity of inferences based on the accountability system indicators. Compiling evidence based on content, internal structure, and generalizability is especially pertinent to the validation process of accountability systems.

Evidence that is based on content links the features of a test to the construct of interest (*Standards*, 2014). Evidence based on internal structure (Wilson, 2008) is produced by comparing the results from statistical analyses of the relationships between items of a test (through factor analysis, structural equation modeling, etc.) to theoretical characterizations of the construct. Evidence supporting the generalizability of an indicator links the indicator to situations beyond the immediate interpretation of that indicator (*Standards*, 2014).

Extrapolation, implication, and accountability indices

Indicators such as end-of-grade test scores, value-added model or growth scores, graduation rates, and now ELP scores are combined in accountability systems to make judgments about teachers, schools, and school districts (*Standards*, 2014). This combining of scores serves a similar function to extrapolation and implication inferences,

which extend interpretations of test scores to a target domain and trait interpretation in Kane's (2006, 2013) IUA framework.

Extrapolation and implication are distinct, but for the purpose of this discussion, both are defined as inferences intended to broaden test score interpretations to include "real-world" performances (Kane, 2013, p. 28). In accountability systems, decision rules and accountability indices are the mechanisms by which extrapolation and implication occur (CCSSO, 2004; Kane, 2013). Indices synthesize data based on decision rules to provide a single score that is used to make judgments about educational quality and student success (*Standards*, 2014). It is the interpretation and use of these indices that must be validated as an argument that is built for the system as a whole (Kane, 2013). The decision rules and indicators used to construct indices are to be evaluated in relation to the operational definitions of educational quality and school success, set by state and federal mandates such as ESSA (2015). These operational definitions are, in effect, the target domain and trait for accountability systems.

Analytic and empirical evidence are gathered to support extrapolation and implication inferences (Kane, 2006). For individual test interpretations, analytic evidence is found in development, when the content, tasks, and processes included in the test are compared with those encompassed by the target domain (Kane, 2013). The closer the test mirrors the target domain, the easier it is to support the extrapolation inference (Kane, 2013). Empirical evidence is derived from external criteria. Test score interpretations are compared against other measures of the same target domain (Kane, 2006). This can include performance-based assessments that may only be practical for smaller sample sizes (Kane, 2006).

In aggregate score interpretations, analytic evidence should be gathered during the development of accountability indices (CCSSO, 2004). The weight assigned to the ELP indicator determines how important EL progress is in evaluating school success. Because of the elusiveness of terms such as educational quality and school success that define the target domain, a pragmatic approach to validation is most appropriate (Moss, 2013; Kane, 2013). The weights given to indicators of quality and success in accountability indices should be scrutinized using stakeholder definitions and interpretations. Data triangulation should be used to inform judgments based on aggregate scores, giving voice to policy makers, school leaders, and teachers alongside strict, quantitative decision rules (Moss, 2013; Kane, 2013). The extent to which that triangulation occurs is evidence for the validity of aggregate score interpretations. Along the same line, extrapolation and implication arguments must include evidence of fairness and transparency (Kane, 2010, 2013). Empirical evidence stems from a critical perspective in both accountability and test-based validity (CCSSO, 2004; Kane, 2006). Once tests and accountability indices are operational, evidence is gathered and compared with other criteria linked to the same target domain or trait.

For accountability systems, longitudinal studies may be used to support score interpretations for groups of students by comparing them to long-term student outcomes (CCSSO, 2004). Other indices (e.g., Adequate Yearly Progress, Education Value-Added System (EVAAS), stakeholder surveys, and document analysis could also be used as criteria for comparison (Lane and Stone, 2002). Empirical evidence must also refute threats of trait under-representation and irrelevant variance (Kane, 2006). To do so, it is important to consider the effects of external factors, such as educational opportunity,

English learner status, race, and socioeconomic status on aggregate score interpretations.

Discussion

A comprehensive validity argument for accountability systems needs to be initiated during the test development process, policy consequences need to be evaluated, and validity evidence for the use of aggregate scores needs to be gathered. Validity frameworks need to be reconceptualized in consideration of test-based accountability systems where scores are aggregated to measure the performance of teachers, administrators, and schools (Chalhoub-Deville, 2020). Reform-driven policies such as NCLB (2002) and ESSA (2015) mandate and attach consequences to schools and teachers based on aggregates and derivatives of student test scores. Therefore, consequences are inextricably tied to the validation of test use and interpretation in accountability.

None of the three validity models reviewed, Kane's IA/IUA (2006, 2013), Acree et al.'s Parallel IUA (2016), nor the Integrated Framework for Construct Validity by Embretson (2007, 2008) address all these demands of accountability testing. To address the gaps in these validity models, the proposed accountability IUA (Fig. 3) offers a systematic approach for building a validity argument beginning in the test design and development phase, includes a parallel process for building validity evidence for aggregate scores, and considers the consequences of accountability systems.

Implications

The accountability IUA offers a reconceptualization of validity frameworks to account for the demands of accountability systems where aggregate scores are used as measures of the success of teachers and schools in educating all students, including ELs.

Test development

Operationalizing the proposed accountability IUA requires test developers to consider the potential for aggregation of test scores and lay a foundation for an accountability validity argument in the test development phase. Potential interpretations and uses at the student or aggregate level should be anticipated. Principled assessment design approaches like ECD guide test developers in articulating the chain of reasoning that links evidence to claims about target constructs. Logic and theory guides test specification and item development.

Consequences

Bachman and Palmer's (2010) AUA, the preeminent validity model in language testing, merits highlighting for its unique focus on consequences as the basis for the validation argument in the test development and design phase. In accountability test contexts, however, Chalhoub-Deville (2020) advocates for investigating consequences at the policy-making stage before test development begins. She distinguishes this approach "from what is proposed by Bachman and Palmer (2010) where consequences are considered at the beginning of test development, after a policy has been finalized and rolled out" (p. 257). In the Accountability IUA, implications and consequences are evaluated as part of the use of aggregate scores when educational reform policies are being

defined. Consideration should be given to maximizing the opportunity to meet the objectives of the accountability system while anticipating and minimizing unintended negative consequences. Theory of action (TOA) offers a framework within which to consider the impact of accountability systems on students, teachers, and schools. The effectiveness of the accountability system in achieving the intended goals can be evaluated through a TOA (Chalhoub-Deville, 2016). A TOA explicitly states the intended outcomes as well as the action mechanisms through which they will occur conceptually and operationally. Furthermore, potential implementation problems and negative consequences are identified. A clearly defined TOA allows for a meaningful evaluation of the accountability system (Bennett, 2015). As part of the validation plan, it is necessary to identify and map key intended or imposed consequences including rewards, sanctions, and interventions (CCSSO, 2004). According to Chalhoub-Deville (2020, p. 259),

> *The use of frameworks such as TOA invites systematic and anticipatory research that can help us move beyond traditional, individual-focused test scores and related technical quality documentation. Such frameworks can help us attend to actual desired socio-educational goals embedded in a policy (or a client's request) and address research into unintended outcomes.*

Future editions of the *Standards* (2014) need to hold test developers accountable for anticipating consequences of the interpretation and use of their tests at the individual and aggregate level in accordance with the zone of negotiated responsibility (ZNR) described by Chalhoub-Deville (2016).

Roles and responsibilities of test developers, test users, and policy makers

Often the use of aggregate scores and the design of accountability systems is a policy decision. This means test developers must engage with policymakers regarding their testing needs and their goals for interpretation and use of tests. Test developers cannot monitor, or control all uses of their tests, but as Chalhoub-Deville (2020) describes, "[t]est providers create tests with *some* understanding of the consequences entailed by the testing program, but they are reluctant to engage in validation to uphold those consequences" (pp. 255–256, emphasis in original). Historically, test developers have hidden behind test specifications and the interpretations and uses laid out therein to absolve themselves of responsibility for unintended consequences of the use of their tests beyond the originally defined scope. Knowing that test scores may be aggregated and used in accountability systems and that accountability systems impose consequences such as sanctions and rewards, obligates education test developers to consider the consequences of such use.

Sireci and Forte (2012) point out that tests are at the center of accountability and that "the use of tests is initiated and mandated by policy makers" (p. 27). Elected officials such as general assemblies, governors, and politically motivated and directed chiefs of state education have become the decision makers in the design of many state accountability systems. Decisions regarding what indicators are included in an accountability system and how they are calculated is often a political conversation as evidenced by the fact that state statute often regulates the definition of accountability systems (Education

Commission of the States, 2018). Sireci and Forte (2012) argue that "it is an ethical imperative for the measurement community to do all we can to inform policy makers of the strengths, benefits, and limitations of educational tests" (p. 27). Testing programs and accountability systems necessitate "[c]ommunication and engagement with policy makers, education professionals, and other key stakeholder groups beyond the measurement community" (Chalhoub-Deville, 2020, p. 245).

Smith and Benavot (2019) have argued that discussions of accountability exclude the "voices of stakeholders who work, learn, and teach in schools and other educational institutions" (p. 193). They advocate for the inclusion of these stakeholders, particularly in discussions of planning and evaluation through what they have labeled "structured democratic voice." Their "collaboration for structured democratic voice" diagram (Smith and Benavot, 2019, p. 202) offers a useful vision for the engagement of stakeholders; however, they have not included the measurement community in that collaboration. A modification to their depiction of a "collaboration for structured democratic voice" to include the measurement community is shown in Fig. 4.

Writing to the measurement community, Sireci (2019) said, "[e]ducation policy makers, state and local department of education staff, superintendents, principals, and teachers are all involved in educational testing. It is time for us to get involved with them." Researchers and experts in measurement have an obligation to step outside of theory, and perhaps outside of their comfort zone, to engage with and inform test users and policy makers. Future editions of the *Standards* need to hold test developers accountable for engaging in this collaboration. Chalhoub-Deville's ZNR offers guidance for test developers, test users, and policy-makers to engage in meaningful discussion of the shared responsibility for outcomes and consequences (2016, 2020). Further guidance is given by Sireci and Forte (2012) who "discuss the types of information that are important to communicate to policy-makers, how to best convey this information in a manner in which it can be understood, and how to be seen as a valuable source of information to education policy makers" (p. 27).

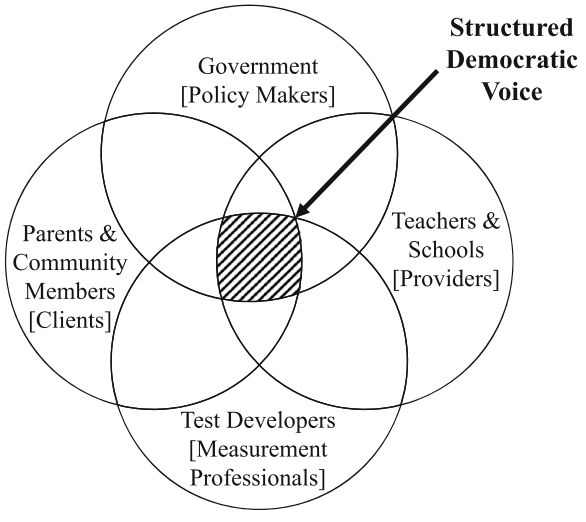

Fig. 4 Adapted from "Improving accountability in education: The importance of structured democratic voice," W. C. Smith and A. Benavot, 2019, *Asia Pacific Education Review, 20*, p. 202. CC BY 4.0.

Conclusion

Accountability testing in the USA has increasingly combined subject area and ELP testing and holds teachers and schools accountable for student performance. Under ESSA, schools are accountable for the progress of ELs in English language proficiency, math, reading/language arts, and in other accountability indicators such as graduation rates. The proposed IUA framework for accountability (Fig. 3) maintains not only that a test may be valid at the individual level and not at the aggregate level, but also that a test may be valid at the aggregate level and not at the individual level, depending on the use and interpretation. If tests are going to be used at both the individual and the aggregate level with the same interpretation, then the individual and aggregate scores must both be validated for that interpretation and use. A strength of the proposed IUA framework for accountability systems is that it allows for separate and parallel validation of individual and aggregate scores. This paper provides a validity framework for the validation of accountability systems and suggests a methodical approach for building validity evidence for the use of aggregate scores that is relevant not only to accountability EL testing, but also to the wider measurement community.

Abbreviations

AUA: Assessment use argument; CCSSO: Council of Chief State School Officers; EL: English learners; ELP: English language proficiency; EVAAS: Education value-added system; ESSA: Every student succeeds act; ECD: Evidence-centered design; IA: Interpretive argument; IA/IUA: Interpretive argument/interpretive use argument; IUA: Interpretive use argument; TOA: Theory of action; VA: Validity argument; ZNR: Zone of negotiated responsibility

Acknowledgements

No acknowledgements to declare.

Author's contributions

Sole author. The author read and approved the final manuscript.

Author's information

Author is affiliated with the University of North Carolina at Greensboro and is also a psychometrician at the American Board of Pediatrics.

Funding

No sources of funding to declare.

Availability of data and materials

Not applicable.

Declarations

Competing interests

No financial or non-financial competing interests to declare.

Received: 7 December 2021 Accepted: 27 December 2021
Published online: 01 February 2022

References

Acree, J., Hoeve, K.B., Weir, J.B. (2016). Approaching the validation of accountability systems. Unpublished paper and presentation. ERM 600: Validity and Validation, University of North Carolina at Greensboro.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Bennett, R. (2015). *Validity considerations for next-generation assessment: a "theory of action" perspective. Paper presented at National Conference on Student Assessment*. San Diego: ETS (Educational Testing Service

Chalhoub-Deville, M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics*, *29*, 118–131. https://doi.org/10.1017/S0267190509090102.

Chalhoub-Deville, M. (2016). Validity theory: reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453–472. https://doi.org/10.1177/0265532215593312.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: theoretical development and integrated arguments*. Equinox Publishing Limited.

Chalhoub-Deville, M. B. (2020). Toward a model of validity in accountability testing. In *Assessing English language proficiency in US K–12 schools*. New York: Routledge. https://doi.org/10.4324/9780429491689-13.

Chapelle, C. A. (2012). Validity argument for language assessment: the framework is simple …. *Language Testing*, *29*, 19–27.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212–225. https://doi.org/10.1080/0969594X.2015.1063479.
Council of Chief State School Officers (CCSSO) (2004). *A framework for examining validity in state accountability systems*. Washington, DC: Council of Chief State School Officers.
Education Commission of the States. (2018). 50-State Comparison: States' School Accountability Systems. Retrieved 15 August 2021 from https://www.ecs.org/50-state-comparison-states-school-accountability-systems/.
Embretson, S. (2008). Construct validity: a universal validity system [PowerPoint Slides]. Retrieved 2 July 2019 from https://marces.org/conference/validity/8Susan%20Embretson.ppt
Embretson, S. (2017). An integrative framework for construct validity. In *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications*, (pp. 102–123).
Embretson, S. E. (2007). Construct validity: a universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455. https://doi.org/10.3102/0013189X07311600.
Every Student Succeeds Act (ESSA), 20 U.S.C. § 6301 (2015). Retrieved from https://www.congress.gov/bill/114th-congress/senate-bill/1177.
Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled approaches to assessment design, development, and implementation. In *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications*, (pp. 41–74).
Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia*, *9*(14) Retrieved from. https://doi.org/10.1186/s40468-019-0089-4.
Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement*, (4th ed., pp. 17–64). Westport: Greenwood Publishing.
Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467.
Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210.
Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000.
Kane, M. (2015). Validation strategies: delineating and validating proposed interpretations and uses of test scores. In *Handbook of test development*, (pp. 80–96). Routledge.
Kane, M. (2020). Validity studies commentary, educational assessment, *25*(1), 83–89. https://doi.org/10.1080/10627197.2019.1702465.
Lane, S. (1999). *Validity evidence for assessments. Reidy interactive lecture series*. Pittsburgh: University Pittsburgh.
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. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448. https://doi.org/10.3102/0013189X07311286.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*, (3rd ed., pp. 13–103). New York: Macmillan.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–7.
Moss, P. (2013). Validity in action: lessons from studies of data use. *Journal of Educational Measurement*, *501*(1), 91–98. https://doi.org/10.1111/jedm.12003.
Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage. https://doi.org/10.4135/9781446288856.
No Child Left Behind Act of 2001 (NCLB), Pub. L. No. 107–110, 115 Stat. 1425 (2002). Retrieved from https://www2.ed.gov/policy/elsec/leg/esea02/index.html.
Office of Elementary & Secondary Education (2021). Retrieved 18 November 2021, from https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/title-i-part-a-program/.
Plake, B. S., Huff, K., Reshetar, R. R., Kaliski, P., & Chajewski, M. (2015). Validity in the making: From evidenced-centered design to the validations of the interpretations of test performance. In M. Faulkner-Bond, & C. Wells (Eds.), *Educational measurement: Foundations to future*, (pp. 62–73).
Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2016). Evidence-centered design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development*, (pp. 40–63). Routledge/Taylor & Francis Group.
Shepard, L. A. (2016). Evaluating test validity: reprise and progress. In *Assessment in Education, 23, 2, 268-280.* Amherst: Center for Educational Assessment, University of Massachusetts.
Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256.
Sireci, S. G. (2019). From the president: you, me, and NCME! Retrieved from https://www.ncme.org/blogs/megan-welsh1/2019/06/30/you-me-and-ncme.
Sireci, S. G., & Forte, E. (2012). Informing in the information age: how to communicate measurement concepts to education policy makers. *Educational Measurement, Issues and Practice*, *31*(2), 27–32. https://doi.org/10.1111/j.1745-3992.2012.00232.x.
Sireci, S. G., & Soto, A. (2016). Test validation for 21st-century educational assessments. In *Meeting the challenges to measurement in an era of accountability*. Routledge.
Smith, W. C., & Benavot, A. (2019). Improving accountability in education: the importance of structured democratic voice. *Asia Pacific Education Review*, *20*(2), 193–205. https://doi.org/10.1007/s12564-019-09599-9.
Wilson, M. (2008). *Constructing measures: an item response modeling approach*. Mahwah: Lawrence Erlbaum Associates.

Publisher's Note