The Gordon Commission
on the Future of Assessment in Education

Variety and Drift in the Functions and Purposes of Assessment in K-12 Education

Andrew Ho
Harvard Graduate School of Education

The target of validation—the primary criterion of assessment design and use—is not an assessment, but the uses and interpretations of assessment results. Kane (2006) describes the *interpretive argument* that leads to validation as one that necessarily requires "a clear statement of proposed interpretations and uses" (p. 23). The chapters on validity in the previous editions of the field's handbook, *Educational Measurement*, are consistent in describing a validation approach that must follow from a statement of purpose (Cronbach, 1971; Messick, 1989).

Although the literature on validation is substantial, comprehensive frameworks for the purposes of educational assessments are relatively rare. "Purpose" is thus central to validation, but it more often illustrates validation frameworks than inspires a framework of its own. This is partially explained by the assumed applicability of validation frameworks to any assessment purpose, but it is also consistent with an understanding that the possible purposes of an assessment are numerous, often underspecified, and subject to change over time.

In this paper, I have two goals. The first goal is to review recent frameworks that differentiate among purposes of educational assessments, particularly purposes of large-scale, standardized assessments. The authors of these frameworks agree that interpretive arguments differ across purposes and that assessments are unlikely to serve multiple purposes equally well. The second goal is to reflect on the forces that shape the purposes of any particular assessment over time. I describe the tendency of assessment purpose to evolve, multiply, and, like an adaptive species, become distinct from original purposes and more difficult to isolate and support empirically. The purpose of a modern assessment program is thus better described as an expanding cloud than as a singular, static entity. To use a metaphor, the first goal is to establish a map of assessment purposes. The second goal uses this map to identify migratory patterns for modern assessment programs as they expand across purposes.

To keep this paper manageable in scope, I focus my attention largely on the purposes of assessments for "mainstream" students in K-12 educational institutions. I include only limited discussion of assessments for special populations, including English learners and students with disabilities. I also exclude assessments for licensure and certification in various professions, direct assessments of teachers, and assessments that are given to students by psychologists or school counselors. These assessments serve multiple and important purposes, and indeed I expect their importance and usefulness, as an imaginary proportion of all assessments, to grow

over time. However, although a validation framework may generalize across these applications, a framework of purposes must expand. To keep this expansion modest, the scope is limited.

Recent Descriptions of Variety in the Purposes of K-12 Educational Assessment

In the development of a framework for the purposes of educational assessments, I begin with four contrasting contributions to the recent literature. First, I review the work of Haertel and Herman (2005), where approximately five broad purposes are identified. I extend this discussion to Haertel's recent address at the National Council on Measurement in Education (2012), where he accepted the Council's Career Contributions Award. There, Haertel made an dimension of past validity frameworks explicit by distinguishing between assessment for Measurement and assessment for Influence. I will capitalize the words Measuring and Influencing when I mean to distinguish these as purposive categories. The second part of the paper reflects on the process through which assessment purposes often drift from Measurement to Influence.

I continue with Pellegrino, Chudowski, and Glaser's (2001) NRC report on assessment, *Knowing What Students Know*. The NRC report had a broader focus than large-scale assessment and described three purposes in a parsimonious framework. Finally, I review Kane's (2006) Validation chapter from the perspective of assessment purpose. Kane does not provide a formal framework for assessment purposes but incorporates various purposes to illustrate his process of interpretive argument. Among the purposes Kane reviews in some depth are placement testing, trait identification, theory development, classroom assessment, and accountability programs.

The number of purposes identified by frameworks such as these is the result of a negotiation between parsimony in presentation and accuracy in description. Increasing the number of purposes may decrease the likelihood of inaccurately excluding an existing purpose but leads to weighty frameworks that have limited predictive or conceptual utility. Categories for assessment purposes are neither mutually exclusive nor exhaustive. The purpose of this review is not to "solve" for some optimal number of categories but to understand the advantages and disadvantages of existing frameworks, inform current validation efforts, and predict how validation may need to change as assessment purposes—and the purpose of any single assessment—tend to evolve and multiply.

Haertel and Herman (2005)

In a 2005 chapter, Haertel and Herman take a historical perspective on interpretive arguments for accountability testing. From early 20th century "scale books" to modern accountability testing, they review a number of assessment purposes and the evolution of interpretive arguments supporting these purposes. Although their narrative describes eras and their interpretive arguments, it can also be read as the rise and fall of the salience of assessment purposes. They note that two broad purposes that have perennially motivated interpretive arguments: assessment for individual placement and selection and assessment to improve the quality of education. Table 1 lists these purposes and juxtaposes them with purposes listed in frameworks to come.

Haertel and Herman (2005) then describe an increase in emphasis on a separate assessment purpose: evaluation of the effectiveness of educational programs and policies. This was a purpose articulated by Tyler and his colleagues following the Eight-Year Study (Smith & Tyler, 1942) that investigated the application of progressive educational ideals to secondary schools. Evaluation continued to gain in prominence after Sputnik and under Title I of the Elementary and Secondary Act of 1965, where the comparison of the relative effectiveness of curricula was a high priority.

Haertel and Herman (2005) mark the 1983 publication of "A Nation at Risk" (National Commission on Excellence in Education) as an indicator of the rise of two other assessment purposes: shaping public perception and focusing attention on educational reform. The publication used assessment to signal a problem, identify the solution, and demonstrate commitment to that solution. As the upcoming framework makes clear, the Haertel and Herman chronology, from placement testing to shaping public perception, can be read as a transition from Measurement purposes of assessments to Influencing purposes of assessments.

Haertel (2012)

In his 2012 address, Haertel focused more explicitly on purposes than on interpretive arguments as a whole, and he distinguished among seven purposes (Table 1). The previous section listed five of these, student placement and selection, improvement of educational quality,

evaluation of policies and programs, focusing the system, and shaping public perception. To these he added two additional purposes that were less salient in the Haertel and Herman (2005) work. The first is educational management, which encapsulates the measurement of teachers and school administrators and the use of these measurements to make inferences and support decisions about schools. The second is directing student effort, where assessments can motivate students to focus on what is or might be assessed. He clarifies that these are intended purposes and leaves unintended consequences, such as test score inflation and its many degrees and sources, to reviews such as those by Koretz (2008).

Table 1. Rough overlap between recent frameworks for assessment purposes in K-12 education.

Haertel and Herman (2005)	**Haertel (2012)**	**Pellegrino, Chudowsky, and Glaser (2001)**	**Kane (2006)**
Historical perspective on interpretive arguments for large-scale accountability testing	Review of purposes for large-scale assessment	Broad review of assessment purposes	Illustration of types of interpretive arguments
Individual placement/selection	Student placement/selection	Assessment of individual achievement	Placement testing
Improve quality of education	Instructional guidance	Classroom/formative assessment	Classroom assessment
	Directing student effort*		
	Educational management		
Policy/program evaluation	Policy/program evaluation	Policy/program evaluation	Accountability testing
Focusing attention*	Focusing the system*	Signal worthy goals*	
Shaping public perception*	Shaping public perception*		
			Trait identification
			Theory development

* Influencing Purposes

An important contribution of Haertel's 2012 address is his division of his seven purposes into Measuring purposes and Influencing purposes. He describes Measuring purposes as "those that rely directly on informational content of specific test scores" (slide 4, 2012), whereas Influencing purposes are "effects intended to flow from testing per se, independent of specific test results" (slide 4, 2012). For example, a classroom assessment program may direct student effort and increase motivation regardless of test results. Likewise, a policymaker may intend a standards-based assessment to increase attention to desired learning outcomes regardless of scores on assessments. No single assessment program intends solely to influence without measurement, however, the distinction becomes particularly useful in the description of drift in assessment purposes.

The explicit identification of an Influencing purpose is an important contribution because it identifies these purposes as targets for validation. Haertel (2012) argues that it is insufficient to support an interpretive argument for a Measurement purpose when an Influencing purpose may be an ultimate or otherwise essential component. Haertel also observes that attention to the validation of Influencing purposes tends to reveal unintended consequences, including test score inflation and narrowing of educational goals to tested content. Nonetheless, Haertel distinguishes between his Influencing purposes and Messick's (1989) "consequential basis of test interpretation and use," sometimes abbreviated to the ambiguous term, "consequential validity." The latter concerns social consequences and value implications (Messick, 1998), both intentional and unintentional. This is a broader class of concerns that intersects with both Measuring and Influencing purposes. The Influencing purposes in Table 1 are generally more easily identifiable as an explicit purpose of an assessment program than social consequences and value implications. In the second half of this paper, I describe the increasingly frequent and predictable drift of assessment purposes from Measurement to Influence.

Pellegrino, Chudowski, and Glaser (2001)

In their 2001 NRC Report, Pellegrino, Chudowski, Glaser, and their committee identify three purposes of assessment (Table 1). The first purpose listed in the NRC report is assessment to assist learning. For this purpose, the primary audience for assessment results includes

students and teachers. This purpose is consistent with *classroom assessment* or *formative assessment,* which Black and Wiliam (1998) propose as a necessary feature of effective teaching.

Haertel (2012) is clear that his framework focuses on purposes of large-scale standardized tests, thus his categories do not easily incorporate "assessment to assist learning." Haertel's "instructional guidance" purpose should ultimately assist learning by assisting teachers, but this is indirect, and the gap between instructional guidance and assisting learning is something that a validation endeavor should fill with evidence. "Directing student effort" also relates to assisting learning, but this is one of many ways that learning may be facilitated. This speaks to the multiple mechanisms that might support the assistance of learning and raises the question of how this purpose is to be achieved. As assessment purposes drift, the formative purpose is often taken for granted. Without evidence to address whether assessments do in fact assist learning, this is the fallacy that Kane (2006) has described as "begging the question."

The second purpose listed in the NRC report is assessment of individual achievement. This category includes end-of-course grades assigned by teachers, admissions and selection assessments to college and graduate schools, and individual scores on large-scale assessments used in state accountability systems. These assessments are often called *summative assessments* and contrast with the previous category of formative assessments or assessments to assist learning. This purpose is broad and aligns roughly with Haertel's (2012) "student placement and selection" category, although Haertel's purpose is focused more on large-scale assessment, whereas the NRC category applies also to classroom assessment.

The distinction between these first two NRC purposes should not be viewed as incompatible. Whether distinguished as formative vs. summative assessment, assessment to assist learning vs. assessment for individual achievement, or assessment of learning vs. assessment of learners, summaries of individual achievement are immensely useful for formative feedback. A teacher and a student should both know where the student is in order to understand where the student might be with sufficient instructional support. An assessment that is strictly summative may indeed be useful for formative purposes, although, as noted before, evidence is required to support this purpose. In short, drift from a summative purpose to a formative purpose is not, in and of itself, problematic.

What then justifies statements like those of Black and Wiliam, who caution that "if an optimum balance is not sought, formative work will always be insecure because of the threat of

renewed dominance by the summative" (1998, p. 59). These statements conflate summative assessment with assessment for accountability or assessment that purports to identify a permanent, fixed aptitude. These are indeed less compatible with formative assessment. The latter is at odds with a theory of learning, and the former can subjugate formative feedback to the priorities of external reporting. An assessment for accountability that purports to assist learning creates far more tension with formative goals than an assessment that provides summative information without an accountability framework.

The third purpose listed in the NRC report is assessment for program evaluation. Assessments that meet this purpose include any that support aggregate scores, from assessments designed for small-scale research purposes to large-scale international assessments such as PISA, TIMSS, and NAEP. Program evaluation is often supported by an experimental or quasi-experimental design that supports defensible interpretation of program or other aggregate-level comparisons. A Table 1 shows, this purpose aligns well with previous frameworks.

Late in the NRC report, the authors dedicate a section to the observation that large-sale assessment can signal worthy goals. Here, they establish the most explicit transition between Haertel's (2012) Measurement and Influencing purposes, where assisting learning, individual achievement, and program evaluation are more Measurement-driven, with a focus on measurement of learning, learners, and programs, respectively. The "signaling of worthy goals" category aligns closely to Haertel's purpose of focusing the system, and it also intersects with directing student effort. Both require no test results per se but are instead assumed to follow from the implementation of a testing program. Whether in a classroom or in an accountability testing program, the theory of action relies on clear communication of the assessed domain to teachers and then to students. Whether this results in desired student learning can be addressed empirically.

Kane (2006)

The comprehensive treatment of validation in Kane's 2006 *Educational Measurement* chapter provides a useful practical framework for validation. Kane framed his interpretive argument as having at least four stages, scoring, generalization, extrapolation, and finally an

interpretation, decision, implication, or action. He illustrates these examples with purposes that align roughly with those presented in the previous sections (Table 1).

Kane's (2006) first illustration of his interpretive argument uses the example of placement testing. This aligns closely with the selection purposes of the first two frameworks and represents a specific case of the broader NRC assessment purpose of individual achievement. A second illustration involves classroom assessments, which align closely to Haertel's (2012) purpose of instructional guidance. Kane and Haertel both stop short of following through from instructional guidance to the more general NRC purpose of assisting learning.

A third illustration involves trait identification, where Kane defines a trait as, "a disposition to behave or perform in some way in response to some kinds of stimuli or tasks, under some range of circumstances" (2006, p. 30). Kane notes that trait labels and descriptions imply values and assumptions, and these may make implicit or explicit predictions or justifications of decisions and actions, including, for example, student placement and selection. All of these are subject to an interpretive argument. Trait identification maps to the NRC report's description of individual achievement but hews closer to the psychological literature and terms like "aptitude." Haertel's categories do not incorporate trait identification except to the degree that the interpretive argument for trait identification extends to student placement, selection, and instructional guidance. Such an extension is a vector for purpose drift.

A fourth illustration extends trait identification to theory development, establishing relationships between traits and other observed or latent phenomena. Treated at a distance as the incorporation of trait estimates into various latent or traditional regression models, theory development aligns well with the NRC purpose of program evaluation and with, for example, Haertel's (2012) purpose of informing comparisons among educational approaches. However, Kane's example focused more on mechanisms within and between individuals. This is a psychological framework more than a framework for program or curriculum evaluation.

Nonetheless, the drift of purpose from trait identification and theory development to program evaluation is as easy as following the implications of a regression model that uses a trait as an outcome. In a recent address, John Easton, director of the Institute of Education Sciences, observed that some psychological variables like grit (Duckworth, Peterson, Matthews, & Kelly, 2007) and flow (Csikszentmihalyi, 1990) can predict future outcomes better than test scores (Easton, 2012). This observation is a far cry from recommending the use of these variables as

outcomes in school evaluation. My observation is that it is easy for the purpose of, for example, an assessment of "grit," to drift from trait identification and theory development to program evaluation. The requirements of the validation agenda must follow.

A fifth illustration involves large-scale accountability programs such as those under the No Child Left Behind (NCLB) Act. These programs are increasingly complex, relying on multiple indicators and layered decision rules. The NRC purpose of program evaluation partially subsumes accountability testing, as these tests can support evaluations of states, schools, teachers, and interventions. However, accountability testing programs attach sanctions to evaluations, as well as serving the functions of focusing the system, instructional guidance, and shaping the system. The distribution of purposes that accountability programs explicitly and implicitly serve is a result of purpose drift along vectors that can and should be anticipated.

Additional Framework Dimensions

The frameworks overviewed in Table 1 provide a map on which we can chart the location, expansion, and drift of the purposes of an educational assessment. The vectors along which purposes tend to drift are rarely the same as the directions that validation agendas can easily follow. The difficulty of extending a validation agenda, can be better anticipated by additional dimensions that distinguish purposes from each other. Drift between purposes that differ on multiple dimensions will be far more difficult to defend.

Assessment as Process vs. Product

One way to distinguish between assessment to support learning and assessment of individual achievement is to observe that the former treats assessment more as a process, and the latter treats assessment more as a product. A student may learn *through* an assessment more than *from* an assessment outcome. The difficulty of extending validation from product-oriented purposes to process-oriented purposes is predictable in this light.

Haertel's (2012) distinction between Measurement and Influencing purposes overlaps with but does not completely capture this dimension. The deliberate focus of this framework on large-scale assessments leaves less room for a description of process-oriented purposes. Although Measurement seems product-oriented, and Influence seems process-oriented, the influence that Haertel describes is external and at a distance. Although process-oriented

assessment could be described as intending to influence learning, the Influence descriptor is more aptly reserved for externally mandated tests.

Levels of Aggregation

In addition to the Measurement and Influencing supercategories, Haertel's framework incorporates a number of additional dimensions not described here, including its primary users, the construct being measured, the degree of linkage to the curriculum, the norm- or criterion-referencing of the interpretation, and whether the interpretation is of an individual or group. This latter dimension is one that can be described more generally as the level of aggregation.

Selection tests support inferences about individuals. Educational management requires inferences at the classroom (teacher) and school (administrator) levels. Comparisons among educational approaches require inferences about variables and relationships between variables, and these can also be considered as an aggregate inference above and beyond individual scores. Assessments like NAEP, TIMSS, and PISA report scores at the level of states, nations, and demographic subgroups. As the purposes of an assessment drift across levels of aggregation, the defensibility of the inferences they support will not follow as easily.

Length of Feedback Loops

A related dimension is identifiable when considering, as an ultimate purpose of an assessment, the provision of useful feedback to inform teaching and support learning. This is clearly not the aim of all assessments and purposes, however, it is a particularly desirable purpose to claim and a destination for many vectors of purpose drift. To evaluate the difficulty of satisfying this claim, a useful dimension is the length of the feedback loop in time and space from the assessment event. In other words, once an assessment is given to a student, how long does it take for useful feedback from that assessment to return to that student, and how far must the information travel?

Formative assessments generally have the shortest distance to travel in both time and space. Indeed, the metaphor of a feedback loop is poorly suited for a formative assessment model, where the assessment is a more of an ongoing process rather than an outcome that travels back to a student. Assessments of individual achievement, from trait identification to those that might be used for placement, have feedback loops dependent on the timeliness of scoring,

reporting, and the decision based on the report. For many large-scale commercial assessments that require secure scoring, the wait and the distance traveled are both lengthy. For assessments that build theory or support program evaluation, the feedback loops are even longer and more abstract, as a research consensus is developed, and new curricula, programs, or interventions are eventually implemented. Large-scale national assessments never provide feedback to individual students. Assessments for educational management may only make a difference to learners if they directly result in retraining or refocusing of the teacher, or in cases where they influence a personnel decision.

The length and probability of the successful feedback loop are useful validation checks for assessment programs whose purposes drift towards informing teaching and supporting learning. When loops are lengthy or improbable, this raises the question of whether non-formative assessments that serve Measuring purposes can defensibly claim the formative mantle. The more realistic theory of action for these assessments improving student learning lies in the Influencing purposes, where students, teachers, and administrators are arguably more directed in their efforts in response to incentive structures.

Stakes

A fourth dimension is stakes. The stakes on a particular assessment-based decision or interpretation will certainly vary within assessment purposes, but it is also possible to describe how stakes tend to vary across purposes. Formative assessments generally have low stakes, as assessments are frequent (if not ongoing) and any single incorrect decision is unlikely to do harm. Large-scale assessments like NAEP are also low-stakes, although they increasingly support comparisons that motivate policy and change public perception. Theory-building and program evaluation may have low stakes at the level of academic research, but stakes rise as consensus forms and motivates implementation of consequential programs.

Stakes are generally higher for trait identification of individual students, and higher still for tests supporting selection and placement. Stakes on assessments supporting educational management decisions are variable but have increased rapidly in recent years. It is common to find test-based metrics influencing teacher and administrator salary and promotion. As assessment purposes drift to those with higher stakes, particular evidence is required, often

involving at a minimum, sufficient reliability evidence, protection against score inflation, and monitoring to ensure that stakes motivate desired responses.

This is not to suggest that purpose drift toward lower-stakes purposes requires less evidence to justify. It is both cavalier and incomplete to assume that, for example, test score gains are evidence that large-scale accountability assessments support instruction. The risk here is less of the consequences of an incorrect decision than one of false promises: the formative mantle, undeserved.

Completeness of Interpretive Arguments

Finally, a fifth dimension is the completeness of the interpretive arguments for assessments that serve these purposes. Like stakes, the completeness of interpretive arguments varies across assessments within categories, but there are clear cross-category contrasts in the sophistication of interpretive arguments as well. Black and Wiliam (1998) were critical of many classroom assessment uses, and their research was an effort at providing evidence supporting interpretive arguments for formative assessment use. As Haertel (2012) notes, Influencing purposes tend to be the most poorly articulated with the greatest potential for unintended consequences.

As a recent example, a class of so-called "value-added models" has become widely used policy tools for educational management and focusing the system. Haertel (2012) is one of many who shares concerns about the lack of validation for this particular assessment use. The ascription of "value added" to a teacher is an example of what Kane (2006) describes as a reification fallacy, where a distance between observed and expected average student scores is assumed to be the value of a teacher. Reardon and Raudenbush (2009) describe some of the conditions required to support a "value added" interpretation. However, even if a "teacher effect" were a defensible interpretation, theories about how the teaching corps or an individual teacher should improve are often left unspecified. Even rarer is a description of an explicit mechanism through which incentivizing value-added scores improves generalizable student learning outcomes.

Purpose Drift

The second goal of this paper has been to describe the tendency of modern assessments towards a kind of *purpose drift*. The metaphor of *drifting* is not perfect, as it implies a passive, glacial process, and the actual adoption of new purposes can be strategic, opportunistic, and relatively sudden. Eva Baker has called this *purpose creep* (personal communication, April 14, 2012), a catchy and accurate alternative. I keep the drift metaphor as a conservative alternative that assigns little value or blame but helps to describe the tendency toward proliferation of purposes in the lifespan of an assessment program.

Much of the struggle with purpose is captured by the rhetorical difficulty of communicating that validity is not a property of an assessment but a use or interpretation. It is far easier to defend "validity" once, usually in test development, than to exhaustively defend "the validity of uses and interpretations" that are plentiful and require imagination and careful consideration. The *Standards of Educational and Psychological Testing* concedes that, for practical purposes, "appropriate test use and sound interpretation of test scores are likely to remain primarily the responsibility of the test user" (AERA/APA/NCME, 1999, p. 111). Shepard (2005) clarifies this in the context of drift, noting that, "when users appropriate tests for purposes not sanctioned and studied by test developers, users become responsible for conducting the needed validity investigation" (p. 8).

It is worth stepping back and taking a broad view on the process of quantification and the appealing features of numbers. This notion that an assessment, once validated, can be used for anything, is consistent with the idea that numbers "travel." As Porter (1996) has noted, users are weak to numbers and tend to ascribe them with various meanings as purposes dictate. In addition, numbers travel easily across levels of aggregation, from students to classrooms to schools, states, and countries. Aggregation lends itself to program evaluation and narratives about "added value" that can be uncritical. These common practices are examples of appealing fallacies, including reification fallacies, naming fallacies, and ecological fallacies. Most fundamentally, it is the belief that a number can be imbued with an unerring meaning that remains fixed even as purposes, aggregation levels, and stakes will drift.

In the next subsections, I present illustrative examples of purpose drift in educational assessments. As this paper makes clear, I do not hope to prevent drift as much as explain it and anticipate it. I do not see a danger in drift itself. It can in many cases be entrepreneurial,

creative, and a result of or a force for advances in assessment science. However, drift without validation of newly claimed purposes risks unintended consequences. The following examples of drift have all inspired thoughtful validation agendas, but in many cases the validation work has lagged too far behind the drift itself. It is this that I hope we can remedy in the future.

From Selection to Evaluation: The Maine SAT Initiative

Although tests like the SAT and ACT are primarily tools for college admissions, they have recently been incorporated as components of high-stakes school evaluation programs. The ACT is currently required of all public high school juniors in Colorado, Illinois, Kentucky, Michigan, North Dakota, Tennessee, and Wyoming (ACT, 2012), and Louisiana will join this group in the spring (Louisiana Department of Education, 2012). The ACT has distinguished itself from the SAT through its explicit use of high school curricula in test development, a possible explanation for the fact that Maine is the only state that requires the SAT. I focus on Maine because the considerations are particularly well documented.

Hupp, Morgan, and Davey (2008) review the evolution of what became known as the Maine SAT Initiative. The Maine Commissioner of Education officially began consideration of the initiative in the spring of 2005, and mandatory SAT testing for public school juniors began in the spring of 2006. Hupp, Morgan, and Davey describe the motivation in terms of marginal improvements over the existing high school assessment that suffered from low student engagement and perceived irrelevance of results. The SATs were already being taken by two-thirds of Maine students and seemed to align well with the state department's goal of college and career readiness for their students.

From the perspective of replacing an existing but undesirable alternative, the expansion of SAT purpose flowed downhill, as if to fill a vacuum. The demand predated the drift. Test-driven accountability policies like NCLB thus facilitate purpose drift as well as, in the rhetoric of the policy itself, represent purpose drift. Once demand is taken for granted, principles of economy and convenience also motivate drift. Development of a new assessment program may be more expensive than adaptation of an existing assessment program. In the case of the Maine SAT Initiative, where the SATs were already taken by a majority of students, the question becomes, why have two assessments when one may suffice?

Hupp, Morgan, and Davey (2008) review answers to this question, primarily focusing on the concerns about the lack of content alignment of the SAT to state standards and research about the coachability of SAT scores. In order to comply with federal regulations, the SAT had to be augmented by a short Mathematics assessment as well as a Science assessment. This reflects an intuition about assessment that is an essential mechanism of purpose drift: the distinguishing characteristic of an assessment is what it measures. This focus on content and measurement over interpretation and use facilitates the intuition that a test may be used for any purpose as long as what it measures is well understood.

From Low-Stakes to High-Stakes Evaluation: The National Assessment of Educational Progress

Another common vector for purpose drift is from low-stakes to high-stakes evaluation. The National Assessment of Educational Progress has long been intended as a low-stakes audit (Haertel, Beauregard, Confrey, Gomez, Gong, & Ho, et al., 2012), even as the results have become increasingly consequential. In 1996, Musick's use of NAEP to compare state proficiency standards marked the beginning of now formal ongoing effort to map state standards onto the NAEP scale (Bandeira de Mello, 2011). This has led to upward pressure on state performance standards which have been, with few exceptions, lower as indicated by the percentage of students exceeding "proficiency." As Hill anticipated in 1998, the stakes on NAEP have risen along with the visibility of its audit role, and, following Campbell's Law (Campbell, 1976) the integrity of the audit diminishes as NAEP becomes an explicit target.

An additional indicator of the expansion of NAEP purpose can be seen in the increasing number of comparisons that it supports and the levels of aggregation at which those comparisons occur. The first assessments were at the national level and were paired with the establishment of the long-term trend in 1969 and 1970. Trial state-level results were added in 1990, and trial district-level results followed in 2002. Beginning with the 2003 state NAEP administration, all 50 states were required to participate in the Reading and Mathematics assessments in order to receive federal funding, and these state results have been reported biennially since. Most recently, the 2011 NAEP-TIMSS Linking Study is embedding item booklets across conditions to explore whether an internationally referenced NAEP and state benchmark can be defensible.

Additional comparisons do not guarantee higher stakes or threaten NAEP's audit role in and of themselves. However, if relative comparisons rise in prominence and become targets of rhetoric or incentive structures, as they do, for example, in college rankings, audit interpretations would be weakened considerably, as Hill (1998) had anticipated. NAEP and its board have had a history of cautioning against these uses (National Assessment Governing Board, 2002). Recent and dramatic reports of gaming and cheating in college and law school rankings (Perez-Pena & Slotnik, 2012) are a reminder that these cautions will likely continue to be necessary to preserve NAEP's audit role.

Across Levels of Aggregation and Abstraction: Secondary Data Analysis

Due to NCLB assessment requirements and additional initiatives like the Department of Education Statewide Longitudinal Data Systems (SLDS) grant program, an increasing number of rich data structures are available to researchers. A creative secondary data analyst is able to generate a practically limitless number of research questions, data displays, and summary reports that effectively repurpose these data. The earlier example of value-added modeling is but one example of an analytical framework that can add multiple purposes to an assessment program long after the design and implementation of that assessment is complete. In this data-rich age, it is not possible to exhaustively list the purposes that an assessment (or in this case, its data) can serve. This makes anticipation of likely vectors of purpose drift all the more important.

Let me be clear that I believe open and flexible data access policies for researchers are a good thing. However, secondary analysis raises easily anticipated issues. Increased distance from the original assessment design principles facilitates reification fallacies. Distance in both time and space from the actual act of measurement precludes direct formative purposes and leads primarily to evaluative analyses. Validation work tends to follow, focusing more on the appropriateness of evaluative inferences than on formative impact. As an example, although validation of the Measurement purposes of value-added models is proceeding rapidly (e.g., Chetty, Friedman, & Rockoff, 2011; Kane & Staiger, 2012; Rothstein, 2010), validation of Influencing purposes has not proceeded apace.

Anticipation of and Response to Purpose Drift

If known forces cause the purposes of an assessment program to deviate from the purposes originally validated, then conventional validation approaches proposed in the assessment literature are incomplete. Validation is generally framed as reactive to purpose, not proactive in anticipation of known vectors of purpose drift. Calls for end users to validate any extended uses of tests are helpful, but underestimate the problem of purpose drift as an exception when it is increasingly a rule. These users are also increasingly outside of the measurement community and have less access to the validation frameworks described in this paper. If incentives exist for publishers and policymakers to stay silent on the matter of purpose, then validation will become an increasingly toothless endeavor, an eventual scolding of end users long after consequential and potentially indefensible decisions are made.

A deeper understanding of the structural incentives and historical precedent for purpose drift allows for improved anticipation of consequences. This foresight calls for the raising of standards of validation to proactive efforts from developers, policymakers, and analysts. This need not require explication of how an assessment cannot be used, although this may be an effective deterrent in some cases. Instead, this may take the form of the validation agenda that we know will soon be necessary. It is increasingly common that scores will be aggregated, stakes will rise, Influencing purposes will be assumed, and formative claims will be asserted. These are easily anticipated vectors of purpose drift. These define a validation agenda, in short, a research agenda, that can be described in technical manuals as necessary work for extending assessment purposes. The goal need not be the prevention of purpose drift but the facilitation of interpretive arguments and supporting evidence to travel along the same vectors.

References

ACT (2012). 2012 ACT National and State Scores. Retrieved from http://www.act.org/newsroom/data/2012/states.html#.UGz2U67CTIZ

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association

Bandeira de Mello, V. (2011), *Mapping state proficiency standards onto the NAEP scales: Variation and change in state standards for Reading and Mathematics, 2005–2009* (NCES 2011-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC: Government Printing Office.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Campbell, D. T. (1976). *Assessing the impact of planned social change.* The Public Affairs Center, Dartmouth College, Hanover, New Hampshire.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011, December). *The Long-term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. Working Paper #17699. National Bureau of Economic Research.
Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement, 2nd ed*. (pp. 443-507). Washington, DC: American Council on Education.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.

Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology. 92*, 1087-1101.

Easton, J. (2012, March 3). National Arts Education Association Talk. Retrieved from http://ies.ed.gov/director/pdf/Easton030312.pdf

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Haertel, E. H. (2012, April). *How is testing supposed to improve schooling?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.

Haertel, E. H., Beauregard, R., Confrey, J., Gomez, L., Gong, B., Ho, A. D., et al. (2012). *NAEP: Looking ahead. Leading assessment into the future*. National Center for Education Statistics. Initiative on the Future of NAEP. Washington, DC.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement*. *The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 1-34). Malden, MA: Blackwell.

Hill, R. (1998). *Using NAEP to compare state data—While it's still possible*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA. Dover, NH: Advanced Systems, Inc.

Hupp, D., Morgan, D. L., & Davey, T. (2008). The SAT as a state's high school NCLB assessment: Rationale and issues confronted. Paper presented at the National Conference on Student Assessment, Orlando, FL. Retrieved from http://research.collegeboard.org/publications/sat-states-high-school-nclb-assessment-rationale-and-issues-confronted-2008-ccsso

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement, 4th ed*. (pp. 17-64). Westport, CT: Praeger.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Measures of Effective Teaching Project. Bill and Melinda Gates Foundation.

Koretz, D. (2008). *Measuring up: What educational testing really tells us?* Cambridge, MA: Harvard University Press.

Louisiana Department of Education. (2012, April 17). Education leaders advance initiatives to expand college and career ready opportunities for high school students. Press release. Retrieved from http://www.louisianaschools.net/offices/publicaffairs/press_release.aspx?PR=1618

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, 3rd ed*. (pp. 13-103). New York: Macmillan.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35-44.

Musick, M. (1996). *Setting education standards high enough*. Atlanta: Southern Regional Education Board.

National Assessment Governing Board. (2002). Using the National Assessment of Educational Progress to confirm state test results. Retrieved from http://www.nagb.org/content/nagb/assets/documents/publications/color_document.pdf

National Commission on Excellence in Education (NCEE). (1983). *A Nation At Risk: The Imperative Educational Reform* (Report No. 065-000-00177-2.) Washington, DC.: Government Printing.

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

Perez-Pena, R., & Slotnik, D. (2012, January 31). Gaming the college rankings. *The New York Times.* Retrieved from http://www.nytimes.com/2012/02/01/education/gaming-the-college-rankings.html?pagewanted=all

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton, NJ: Princeton University Press.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *American Education Finance Association, 4*, 492 – 519.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 1*, 175-214.

Shepard, L. A. (2005). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-13.

Smith , E.R., and Tyler, R.W. (1942) *Appraising and recording student progress*. New York: Harper and Row.