Assessment, Teaching, and Learning

The Gordon Commission on the Future of Assessment in Education

Volume 2, Issue No. 4 ▪ August 2012

(re)Placing Assessments, Remixing an Old Argument

In the work of the Gordon Commission, attention repeatedly has been called to a wide variety of purposes for which assessment can be used in education. In its work, the Commission has tended to emphasize accountability for and the improvement of teaching and learning as privileged purposes of assessment. Educational policy in federal and state governments is dominated by a concern with the use of assessment for purposes of accountability.

Within the Gordon Commission, consensus is building for greater balance in the emphasis given to the multiple purposes of assessment in education, and for special attention to be given to the use of assessment to inform and improve teaching and learning processes and outcomes.

In this issue of *ATL*, we have updated an excerpt from the 1970 Report of the College Board's Commission on Testing, in which the call is made for the qualitative analysis of the SAT® data to generate diagnostic information. In at least four works in progress for the Gordon Commission, we find advocates for the multiple purposes for which assessment can be and is used. In papers prepared for the Commission, one by Chung and one by Resnick, we are introduced to the relational treatment of assessment data from multiple sources and distributed probes to make judgments concerning learning and teaching persons, as well of the institutions and processes to which they are engaged. In Vol. I, No. 2 of *ATL*, we reviewed the National Research Council (NRC) report *Knowing What Students Know*, and called attention to its admonition against the use of assessment data, instruments and procedures for purposes other than the purposes for which these devices were designed.

As policies and practices in education and its assessment change, the pressure to make multiple creative uses of assessments will, no doubt, increase. This trend will be exacerbated as digitalized electronic technologies enable the relational analysis and management of unlimited amounts and varieties of data. In this issue of *ATL*, we have included the abstract from a Commission paper in progress in which you will find Andrew Ho's interesting discussion of the serious problems related to drift

Assessment, Teaching, and Learning is a bi-monthly bulletin that is the primary instrument of communication from the Chairperson of the Gordon Commission on the Future of Assessment in Education to a broad audience of readers who are concerned with the relationships between psychometrics and education. The intent is to use this bulletin to stimulate conversation and debate concerning the multiple purposes of assessment in education; the possibilities for the improvement of teaching and learning processes and outcomes through the more creative use of measurement in education; visions of future change in the nature and practice of education; and the need for change in the capacity of the educational measurement enterprise necessary to the needs implicit in those visions. *Assessment, Teaching, and Learning* is available, without cost to the reader, electronically and in print.

in the use of assessment instruments and data for purposes other than those for which they were designed.

Ho addresses an issue that has been central to the work of the Gordon Commission. That pivotal issue concerns the multiple purposes for which assessment in education is and can be used. A wide variety of purposes have been identified, including:

- ***Measurement of the status of one's developed abilities***
- ***Inventory of one's characteristics and abilities***
- ***Sorting and rank ordering of subjects***
- ***Accountability***
- ***Prediction***
- ***Selection***
- ***Evaluation***
- ***Diagnosis***
- ***Informing and improving teaching and learning***

It is the last purpose in this list, "to inform and improve teaching and learning," to which many members of the Gordon Commission have called attention. The Chairman of the Commission views "to inform and improve teaching and learning" to be inclusive of and superior to all of the other purposes of assessment in education. In his view, the other purposes listed are and should be instrumental to the production of knowledge that can be used to better inform and to improve the processes and outcomes of teaching and learning.

The assertion of this position is not intended to demean other purposes such as sorting, selection, admission, placement and even accountability. But it is to argue that the investment of so much teacher and learner time in the preparation for and the engagement in assessment exercises is justified only by the contribution that such engagement can make to better informing and improving the processes and outcome of teaching and learning themselves.

If students are to spend time in preparation for assessments, the assessments for which they are preparing should demand the competencies that should be the goals of their education.

Increasingly, we argue, those goals are not coterminous with rote memory, recognition of the right answer and regurgitation of factoids.

Now the design of assessment instruments and procedures that both inform and improve teaching and learning may require that the products meet certain criteria. We have not yet turned attention to such specification, but that work should be on the agenda of any continuation of a focus on the future of assessment. The learning sciences and emerging research in curriculum development are beginning to provide some leads. That kind of R&D will not happen by accident. Those of us who use assessment data and technologies — especially those of us who demand and pay for assessment data — as well as teachers and learners who need such data in order to do their jobs well will need to demand that our assessment instruments and programs produce information that informs and improves teaching and learning processes and, eventually, outcomes.

To achieve this end, we will need to bring greater balance to our national and state education policies that currently place disproportionate emphasis on accountability as the primary purpose of assessment. Our nation is making a sizeable investment of money and human capital in the pursuit of improved instruments and programs of educational assessment. Any investment in education is welcome, but it is our belief that such capital should be invested in assessment instruments and procedures that better inform teachers and learners with respect to how we teach better and how we learn better. Accounting for what we have done may contribute less to that end than the production of knowledge that tells us how to do it. Ho correctly cautions against the tendency to drift from the purpose for which an instrument or procedure was designed. There are serious

transgressions in contemporary policy and practice — for example, student achievement tests were not designed to be solitary indicators of teacher quality, nor are these tests the best indicators of the ways in which a specific student learns or of the developable potential of specific students.

Our instruments and procedures should be used only for the purposes for which they were designed, until they have been subjected to relational analysis and R&D investigations to determine their suitability for use in combinations for other purposes. As we move forward, assessment in education will need to be directed at:

1. ***the recognition of the multiple purposes for which assessment data, instruments and procedures can be used;***
2. ***ensuring the empirically supported use of these devices in combinations for purposes other than the purposes for which each device or procedure was designed; and***
3. ***the achievement of greater balance between our attention to these multiple purposes, so as to avoid the disproportionate and possibly distorting influence that the emphasis on a single device or purpose (such as accountability) may have on education and on the character of the assessment enterprise itself.***

The enterprise could place so much emphasis on holding us accountable for meeting the common core standards, for example, that we will lose the opportunity to develop the capacity to produce the ends that we desire, as well as to produce the actual instruments and procedures that can help us teach and learn what such standards are or should be intended to achieve. This concern with purposes of assessment and appropriate ways to address these multiple purposes is an example of the several issues with which the Gordon Commission on the Future of Assessment in Education is concerned.

[Still] Toward More Whole Assessments: Reiterating Recommendations on How We Understand and Implement Measurement

Edmund W. Gordon

Much of the impetus for the development of a technology of assessment related to intellective function and achievement resulted from and has been maintained by a supply-and-demand approach to access to education and distribution of educational opportunities. Access to a limited supply of educational opportunities has been guarded by selection procedures that prior to the 20th century were based on the prospective student's social status. In the pre-Reformation period, access to education was limited to the political and religious nobility and later to other privileged classes, while the 20th- and 21st-century selection procedures have come to be dominated by the student's demonstrated or predicted intellectual status. Where the supply of opportunities has been limited, great emphasis has been placed on the selection of students and the prediction of their performance when exposed to those opportunities. Binet's work in intelligence-test development was directed toward the creation of an instrument that could be used to identify those pupils who were likely to benefit from schooling.

His admonitions that education also turn to treatment of those exposed as not likely to succeed were generally ignored. In a period of scarce educational opportunities, Binet's concern for the educability of intelligence did not gain favor. Society found greater utility in the promise of the predictive and selective validity of his new test.

This emphasis on selection and prediction has continued even though the social conditions that gave rise to it have changed. In recent years, we have seen in America a growing concern with universal access to secondary and higher education. The educational requirements of the nation are increasingly defined as post-high school educational opportunities for almost all youth and continued learning for most people. If this trend continues, selection and prediction can no longer be allowed to dominate in the technology of psychoeducational appraisal.

Rather, the stage must be shared with an emphasis on description and prescription — that is, the qualitative description of intellective function leading not to the selection of those most likely to succeed but to the prescription of the learning experiences required to more adequately ensure that academic success is possible.

Psychological testing obviously can be used to measure achieved development. From those achievement patterns, subsequent achievement in the same dimensions of behavior under similar learning-experience conditions can be predicted with reasonable validity. Thus, people who have learned an average amount during one learning period (high school) may be expected to learn an average amount in the next learning period (college).

However, adequate attention has not been given to the facts that psychological testing can be used to describe and qualitatively analyze behavioral function to better understand the processes by which achievement is developed, to describe non-standard achievements that may be equally functional in subsequent situations requiring adaptation, or to specify those conditions in the interaction between learner and learning experience that may be necessary to change the quality of future achievements.

In the present situation confronting those concerned with access to higher education for larger numbers of young people and for youth from more diverse backgrounds than those from which college students previously were chosen, it is not enough to simply identify the high-risk students.

The tasks of assessment and appraisal in this situation are to identify atypical patterns of talent and to describe patterns of function in terms that lead to the planning of appropriate learning experiences.

Accordingly, it is recommended that we:

1. Explore possibilities for adding to its quantitative reports on the performance of students, reports descriptive of the patterns of achievement and function derived from the qualitative analysis of existing tests.

2. Explore the development of test items and procedures that lend themselves to descriptive and qualitative analyses of cognitive and affective adaptive functions, in addition to wider specific achievements.

3. Explore the development of report procedures that convey the qualitative richness of these new tests and procedures to students and institutions in ways that encourage individualized prescriptive educational planning.

4. Explore the development of research that will add to understanding of the ways in which more traditional patterns of instruction will need to be modified to make appropriate use of wider ranges and varieties of human talent and adaptation in continuing education.

A. In the development of new tests, attention should be given to the appraisal of: (1) Adaptation in new learning situations; (2) Problem solving in situations that require varied cognitive skills and styles; (3) Analysis, search, and synthesis behaviors; (4) Information management, processing and utilization skills; (5) Nonstandard information pools.

B. In the development of new procedures, attention should be given to the appraisal of: (1) Comprehension through experiencing, listening and looking, as well as reading; (2) Expression through artistic, oral, nonverbal and graphic as well as written symbolization; (3) Characteristics of temperament; (4) Sources and status of motivation; (5) Habits of work and task involvement under varying conditions of demand.

C. In the development of tests and procedures designed to get at specific achievements, attention should be given to: (1) Broadening the varieties of subject matter, competencies and skills assessed; (2) Examining these achievements in a variety of contexts; (3) Open-ended and unstructured probes of achievement to allow for atypical patterns and varieties of achievement; (4) Assessing nonacademic achievements such as social competence, coping skills, avocational skills, and artistic, athletic, political or mechanical skills.

Outlines of a Commission Paper

Outlines of a Commission Paper provides a glimpse into Gordon Commission work in real time with themes that are being developed across a collection of more than two dozen Gordon Commission Papers in progress. The following summary of *Variety and Drift in the Functions and Purposes of Assessment in Education* is by Andrew Ho and, as was expressed in the Chairman's introduction, addresses a "pivotal issue concern[ing] the multiple purposes for which assessment in education is and can be used."

Introduction

***Validity* is a quality of the interpretation and use of an assessment, rather than the assessment itself. It is based on an interpretive argument grounded in a clear statement of purpose. However, outside of the literature, it is rare to see an assessment framework built with "purpose" as a central concept. Instead, assessments are used for multiple, underspecified purposes over time.**

Andrew Ho

Assistant Professor of Education,

Harvard University

Andrew Ho is a psychometrician working at the intersection of educational statistics and educational policies. His research informs and improves the development, use and interpretation of large-scale educational accountability metrics. He has studied the consequences of "proficiency-based" accountability metrics, the validation of high-stakes test score trends with low-stakes comparisons, and the potential for alternative accountability structures — such as "growth models" and "index systems" — to improve school- and classroom-level incentives.

Here, different frameworks for classifying the purposes of assessments — in particular, large-scale, standardized K–12 general education assessments — are addressed. There also is reflection on the forces that shape the uses an assessment is put to and the expansion of those purposes over time.

First, Haertel's distinction between assessment for measurement and assessment for influence is explored by mapping the NRC report *Knowing What Students Know* onto this framework and overviewing Kane's (2006) chapter on validity with a focus on assessment purpose. Also, the tension between a focus on presentation (limiting the number of discussed purposes at the cost of accuracy) and a focus on accuracy (increasing the specificity of discussed purposes at the cost of conceptual utility) is discussed.

Purposes of Assessment

Haertel identifies seven purposes of assessment. Two broad purposes are consistent throughout time: assessment for individual placement and selection, and assessment to improve the quality of instruction. The Elementary and Secondary Education Act of 1965's emphasis on comparing the relative effectiveness of curricula gave rise to the use of assessments in educational program evaluation.

The National Commission on Excellence in Education's *A Nation at Risk* report marked the rise of two other purposes: using assessments to shape public perception and to focus the attention of the education system on reform. Assessments are used both to identify underperforming students and schools and as a barometer of their success in and commitment to addressing the problem. International comparisons arise from this assessment purpose.

In a 2012 address, Haertel added two additional purposes. The first is education management via the measurement of teacher and school effectiveness in a way that supports making inferences and decisions about teachers and schools. The second is directing student effort, in which assessments inform the areas on which students should focus their efforts. Haertel leaves the unintended consequences of those intended purposes for other authors to untangle.

Knowing What Students Know

The 2001 NRC report *Knowing What Students Know* identifies three purposes of assessment: assisting learning, assessment of individual achievement and program evaluation. Assisting learning can be understood as related to the concept of the use of formative assessment to inform instruction. In Haertel's categories, this is understood as two separate purposes: on the Measurement side, "instructional guidance" covers the use of specific test results to assist teachers in improving instruction; and on the Influence side, "directing student effort" covers the indirect impact ongoing formative assessment has on facilitating student engagement.

Individual achievement includes various kinds of summative assessments including end-of-course grades, admission and selection assessments to postsecondary institutions, and individual scores on state accountability assessments. Many aspects of this purpose align with Haertel's more specific "student placement and selection category." The specificity of Haertel's category allows us to avoid conflating summative assessments that inform instruction with those used for individual or school accountability, or conflating formative assessments summarized for a summative accountability purpose with those used to make judgments about learning trajectories.

Program evaluation assessments include those that support aggregate scores, from small-scale research to large-scale international assessments such as the Programme for International Student Assessment (PISA) or the National Assessment of Educational Progress (NAEP) in the United States. This purpose aligns with Haertel's categories of "informing comparisons

among educational approaches" and "educational management." Making inferences about teachers and principals instead of programs involves making a distinction between school personnel and their actions, which can be difficult. The NRC report explicitly aligns assisting learning, individual achievement and program evaluation to Measurement goals around learning, learners and programs, respectively, aligning with Haertel's "focusing the system" category. The signaling of those goals requires no test results per se, but is assumed to focus the implementation of the testing regimen.

Validity and Educational Measurement

The comprehensive treatment of validation in Kane's 2008 *Educational Measurement* chapter provides a useful practical framework for validation. Kane illustrates *trait identification*, defining 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." Trait labels and descriptions imply values and assumptions and make predictions and justifications that require interpretative arguments. Haertel's categories do not incorporate trait identification except to the degree that the interpretive argument for trait identification extends to student placement and selection and instructional guidance.

Kane extends trait identification to *theory development*, in which relationships between traits and other phenomena are established. The incorporation of traits into regression models aligns with Haertel's "informing comparisons" purpose. Kane goes further to provide a framework for the validation of large-scale accountability programs, the purposes of which are explicitly laid out in Haertel's framework.

The Union of Frameworks

Haertel's framework is intended primarily for standardized, large-scale achievement testing and has incomplete applicability to classroom and formative assessment and trait and theory development. Understanding formative assessment in this framework requires a conception of formative assessment as a process rather than a product, one that incorporates teacher and student training in assessment and feedback. From this perspective, although the measurement of learning is crucial to fulfilling the Measurement goal of "instructional guidance," high-quality formative assessment practices change classroom practice regardless of the results of the assessments themselves, fulfilling the Influence purpose of "directing student effort."

Incorporating trait estimation and theory development requires representing Measurement endeavors of "student placement" and "informing comparisons." The level of the theorized impact of the traits and models distinguishes these. Both trait estimates and student placement require inferences about individuals, while theory building and informed comparisons require inferences about relationships at the aggregate level. Large-scale national assessments such as NAEP are examples of measurement at a still higher level of aggregation. Their Measurement purpose of informing comparisons is clear, but their Influence purpose is unclear. While they are influential in shaping public perception and focusing the system, the methods by which they do so are deeply dependent on results, making those impacts more of a Measurement purpose. Fitting them to Haertel's dimensions may require the incorporation of an additional Measurement purpose for large-scale demographic and national comparisons.

Anticipation of and Response to Purpose Drift

There exists the tendency of modern assessments towards *purpose drift* or *purpose creep* — the strategic, opportunistic and relative adoption of new purposes for existing assessments. Much of the struggle with

the purposes of assessment springs from the difficulty of explaining to non-academics that validity is not a property of an assessment but rather of its use and interpretation. Validity as it is defined and defended during test development has little bearing on the responsibility of the test user to appropriately utilize an assessment. The notion that an assessment, once validated, can be used for anything is consistent with the common idea that numbers "travel" — an idea combining a host of appealing fallacies of reasoning that allows test users to ascribe various meanings to numbers as their shifting purposes dictate. Selection tests like the SAT are not designed to be used as components of state accountability testing, and existing tests are not designed to be fit into a statistical model for making "value-added" judgments about teachers.

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 inadequate. Validation needs to be framed proactively in anticipation of purposes to come.

If publishers and policymakers don't change their practices, validation will be reduced to toothlessly scolding end users long after high-stakes, indefensible decisions have been made about students, teachers and schools. A deeper understanding of purpose drift calls for raising the standard of validation to proactively stem the anticipated drift of assessment score purpose. While purpose drift may be impossible to prevent, we know that it occurs and changes the consequences of the implementation of an assessment.

In line with Chairman Gordon's thinking, *ATL* is committed to pushing forward innovative and practical considerations from scholars who take seriously the advancement of human capital through the development of strong minds. Perspectives will be anchored in the desire and need to do better in the utilization of assessment, and will be supplemented in future issues with readings, resources, and lists that help to frame the future of assessment in a way that is responsive to 21st-century learners. We look forward to public discourse and trust our readers also will make their perspectives known through contacting us.