European Journal of Psychology of Education
1987, Vol. II, n.° 2, 157-165

Large-Scale Educational Assessment As Policy Research: Aspirations and Limitations

Samuel Messick
Educational Testing Service,
Princeton (NJ), U.S.A.

Critical features important for transforming large-scale educational assessment into effective policy research are examined. These include a developmental orientation to time-ordered data, empirically-grounded construct interpretations of measures and relationships, speculative inquiry into the role of diverse contributing factors, and appraisal of alternative perspectives on both the questions and the findings.

In his classic 1951 volume on *The Policy Sciences* edited with Daniel Lerner, Harold Lasswell maintained that in a policy orientation it is «essential to cultivate the practice of thinking of the past and the future as parts of one context, and to make use of 'developmental constructs' as tools for exploring the flow of events in time» (p. 4). Thus, in policy research it is not sufficient simply to document the direction of change, which often may only signal the presence of a problem while offering little guidance for problem solution. One must also conceptualize and empirically evaluate the nature of the change and its contributing factors as a guide for rational decision making. Otherwise we have no research basis for inferring how to protect or enhance positive change, to forestall or reverse negative change, or to initiate appropriate change in the face of shifting circumstances or altered conditions.

From this developmental perspective, it is important to recognize that not only is policy research inherently anticipatory or predictive, but that in many instances its concrete forecasts are contingent upon variable and uncontrolled conditions. This makes the need for general developmental constructs even more important to provide a theoretical underpinning for whatever consistencies, trends, and interactions are observed. These contingent or context-dependent concrete predictions of policy research are in contradistinction to the abstract predictions of much basic science where the caveat of «all other factors being equal» is an indispensable qualification. In policy research, as in most applied science, if implications for action are to be drawn from the findings, one must appraise the likelihood that relevant other factors will indeed remain constant

---

A somewhat abbreviated version of this paper was presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1986. A longer version was presented at the 21st International Congress of Applied Psychology, Jerusalem, July, 1986.

and then empirically assess any changes in these factors and their likely impact on action alternatives.

As a consequence, policy research cannot be constrained to the interplay of a restricted number of focal variables in limited or specialized contexts. According to Merton and Lerner in the same 1951 volume, such research «must include some speculative inquiry into the role of diverse factors which can only be roughly assessed, not meticulously studied» (Merton & Lerner, 1951, p. 304). Furthermore, policy implications are the product of both the focal research and the assessment of contingent conditions or context effects, but these latter estimates are not of the same order of likelihood or precision as estimates of the more fully studied focal relationships.

Thus, there is an inevitable degree of uncertainty in the research implications for action, but nowhere near the uncertainty or risk that obtain if contingent relationships and context effects remain unexamined. This uncertainty creates a gap between policy research and policy formulation — a gap that can only be filled by informed judgment (Merton & Lerner, 1951). Although ultimately this judgment is the prerogative of the policy maker, it should be informed by the empirical findings, contextual qualifications, and evaluated implications or action alternatives of policy research. Indeed, again quoting Lasswell (1951), policy science at its best applies «methods... by which authentic information and responsible interpretations can be integrated with judgment» (p. 4).

In light of this midcentury wisdom, if large-scale educational assessments are to function effectively as policy research — that is, to provide empirically-grounded interpretations or understandings to inform policy judgments — a number of key features must be exhibited. Central among these are, first, the capacity to provide data or measures that are commensurable across time periods and population groups, so that trends and group differences can be meaningfully examined; second, the capacity to provide correlational evidence to sustain construct interpretations; and, third, provision for measuring diverse background and program factors to illuminate context effects and treatment or process differences.

Thus, in regard to key characteristics, policy research does not differ in kind from social science research more generally, but rather in the stress applied to conjoint properties exhibited in concert. These key joint characteristics are the *comparability* of measures across time periods and population groups, the *interpretability* of measures in terms of integrative constructs with predictive power, the *generalizability* of measures (or the lack thereof) across diverse contexts and background factors, and the *relevance* of performance measures to manipulable program and process variables amenable to policy influence. This latter property of relevance together with a strong stress on *timeliness* are generally conceded to be the sine qua non of policy research. They are indeed necessary but not sufficient. In the last analysis, the value of large-scale educational assessments for policy making will stand or fall on the basis of more fundamental properties of comparability, interpretability, and either generalizability or its inverse, documented context-dependence.

Let us next consider how these key characteristics were incorporated into a redesigned US National Assessment of Educational Progress (NAEP) and then examine some of the mechanisms by which assessment data can directly and indirectly influence educational policy. But first, a brief word about NAEP. Established in the 1960s to assess the condition and progress of education in the United States, NAEP collected data for the first time in 1969. Since then, over a million 9-, 13-, and 17-year old students, as well as occasional samples of adults and 17-year olds who were not in school, have been assessed in a variety of

subject-matter areas such as reading, writing, mathematics, science, social studies, literature, music, and art. The assessment of reading, writing, and math is legislatively mandated on a five-year cycle. To avoid any implication of national standards, curricula, or tests, the original design for NAEP called for matrix sampling procedures using discrete booklets with reporting at the exercise- or item-level only, so that no student answered more than a small set of items and no test scores as such were reported. National probability samples of schools and of 9-, 13-, and 17-year old students were drawn to permit reporting of results by region of the US and by demographic groups, but *not* by state, city, or school district.

The US national assessment as policy research

To improve the interpretability, comparability, timeliness, and policy relevance of the results and to examine the degree of generalizability of findings across diverse contexts, the new design for NAEP introduces a number of important innovations (Messick, Beaton, & Lord, 1983; Messick, 1985). Field administration moves from an annual schedule assessing one or two subjects to a biennial schedule assessing four subject-matter areas in each wave. Reading is assessed in every wave rather than every four or five years to provide timely reporting of trends in a basic area and to calibrate alternate cohorts; writing is assessed every other wave (i.e., every four years), as are math and science in the alternate waves. In addition to the traditional assessment of 9-, 13-, and 17-year olds, sampling is expanded to permit reporting of the associated modal grades which, with refined definitions of age eligibility, are grades 3, 7, and 11. The newly available grade results can be more readily linked to school practices, state and local assessments, and educational policies, most of which are typically tied to grade level.

There are thus four years intervening between subject assessments, age levels, and grade levels, thereby introducing a systematic cohort matching procedure so that the 17-year olds in a particular assessment are from the same birth cohort as the 13-year olds assessed four years earlier and the 9-year olds assessed eight years earlier. Thus, although NAEP does not provide traditional longitudinal data tracking individual students, it does provide longitudinal data for birth cohorts. Balanced-incomplete block spiralling of exercises, as well as of background and program variables, permits the estimation of intercorrelations among all variables, balanced-incomplete block spiralling, as we shall see, facilitates dimensional analyses within and across subject areas; item-response scaling to improve comparability of scale meaning across ages, population groups, and time periods; and, the examination and structural analysis of various context effects and policy relevant correlates of educational performance.

Given this broad-brush summary, let us now consider the issues of interpretability, comparability, generalizability, and relevance in more detail.

BIB spiralling and the correlational basis for interpretability

Interpretability of findings has been a chronic problem in NAEP as originally implemented because the intended benefits of exercise-level reporting simply were not realized — namely, the unfulfilled hope that the specific learning outcome embodied in a discrete exercise readily conveys its own criterion-referenced standard and that a direct link can be easily perceived between the exercise and the educational objectives it represents. Moreover, the subsequent use of average

percent-correct scoring of composites of exercises apparently reflecting common objectives merely expanded reliance on assumption and judgment. It also creates problems in the assessment of group differences or trends if the composites are not restricted to items common across ages or assessment years. What is needed is a means of moving from single exercises to meaningful, empirically-grounded composites or scales for measuring performance levels. The critical requirement for accomplishing this is to be able to estimate the intercorrelations among the exercises as well as between exercises and other variables. The standard matrix sampling procedure employed in the original NAEP design permitted estimation of correlations only among the exercises appearing in the same administration booklet; in any event, even those correlations were rarely appraised.

The new NAEP design remedies this deficiency by using a powerful variant of matrix sampling called balanced incomplete block (BIB) spiralling. With this procedure, the total assessment is divided not into mutually exclusive booklets each requiring roughly one classroom period of administration time as in the original design, but into blocks of exercises each taking around 15 minutes to complete. Each student is administered a booklet containing three blocks of cognitive exercises as well as a six-minute block of demographic and background questions common to all students. The balanced incomplete part of the method assigns blocks of cognitive exercises to booklets in such a way that each block appears in the same number of booklets and each *pair* of blocks appears in at least one booklet. This generates a large number of different booklets. The spiralling part of the method then cycles the booklets for administration, so typically no two students in any assessment session in a school, and at most only a few students in schools with multiple sessions, receive the same booklet.

With BIB spiralling, correlations may be calculated among all exercises (whether in the same booklet or different booklets) on some subset of students, although different correlations will be based on different random subsamples. This permits estimation of the complete matrix of correlations among exercises within a subject area and the subsequent mapping of the structure of achievement in that domain. Since different exercise blocks may derive from different subject-matter areas, BIB spiralling may yield correlations among exercises not only within subject areas but across subject areas as well. This permits examination of cross-area linkages and the tracing of possible facilitating processes from one area to another. Furthermore, in addition to the common block of background questions taken by all students, two minutes of each cognitive block are currently allocated to student experience and attitude items. These latter questions could instead be variously consolidated and replicated in several booklets to increase the sample size for relating student background to performance measures, thereby increasing the number and precision of subgroup relationships that may be effectivelly explored. Thus, since booklets and blocks are administered to different but random subsamples, BIB spiralling yields correlations between educational performance and a host of background, attitudinal, and program variables.

Item response theory and the quest for comparability

Comparability of findings has also been a chronic problem in NAEP ever since its inception. A key problem is that the relationships between percentage correct and quantitative variables such as those descriptive of background or program characteristics are typically nonlinear. As a consequence, interpretations of the meaning or sources of percentage change, whether at the level of single exercises or composites, are often either misleading or abstruse. This difficulty

may be overcome by employing a statistical scaling model such as item-response theory (IRT) that transforms percentage correct to a logit scale, thus defining latent continua (i.e., ability or performance dimensions) that are typically linearly related to other quantitative variables (Lord, 1980).

An important outcome of this IRT scaling — if the model adequately fits the data — is that item parameters are invariant across groups of examinees, while at the same time estimates of examinee proficiency levels are invariant across sets of items measuring the same ability or skill. Thus, IRT analyses yield a common proficiency scale on which group performance may be estimated and meaningfully compared for any group or subgroup, even though all respondents did not take all the NAEP exercises in a subject area (as is the case with matrix or item sampling in general and BIB spiralling in particular). Furthermore, since many of the same exercises are administered to the different age levels and in different assessment years, a common scale may be established, if the model fits, across age levels as well as across time (Lord, 1980). This enormously simplifies the measurement and interpretation of group differences and trends.

Assessing context and enhancing policy relevance

An important consequence of BIB spiralling is that NAEP exercises and scales may be correlated with any of the diverse background and attitude items that are spiralled into the student booklets (or are taken in common by all students) as well as with teacher, school, and program variables that are tied to the students via teacher and school questionnaires, school records, or other means. These background and program variables may also be used to generate group comparisons, such as students in public versus private schools or language-minority versus nonlanguage-minority students. Given the availability of other background variables characterizing the groups in question, such group comparisons may also be conducted controlling for a variety of demographic, home, and school factors by means of regression analysis or covariance techniques. Although with limited items per student IRT estimates of proficiency on scaled performance dimensions are not reliable enough for characterizing individual students, they are sufficiently reliable for comparisons at the group level as well as for correlational purposes — where in any event unreliability can be taken into account.

The only limitation on the number and nature of potential context effects and of educational and policy questions that can be addressed in this fashion is set by the scope of the relevant background and program variables that are included in the student, teacher, and school questionnaires or are derivable from other sources. Fortunately, the extensiveness of treatment of such policy-related variables is markedly expanded in the new design. Specifically, the opportunity to elicit student information bearing on policy issues is greatly amplified by means of BIB spiralling — as an example, 351 background and attitude items were administered to the 13-year olds in the 1983-84 assessment. These student questions covered demographic characteristics and home environments; educational background and current practices; exposure to courses and computers; use of time both in and out of school; and, orientation toward school, studying, and subject matters.

The teacher data derive from a random sample of teachers of the assessed students. The selected teachers are administered a questionnaire covering background, education, and training; characteristics of the instructional program; and, teacher perceptions of the school and its curricula These teacher charac-

teristics are associated with each assessed student of a given teacher, thereby generating a teacher-linked probability subsample of the total student sample selected for assessment.

In addition to the teacher questionnaire, extensive contextual data also derive from a school questionnaire covering characteristics of the principal, staff, and student body; of standards, programs, and computers; and, of school climate, finances, and resources. Thus, the new NAEP design affords ample opportunity to examine the background and program correlates of student educational performance in assorted educational contexts in relation to a variety of policy issues. Examination of context effects is especially important since context may influence not only the accuracy and meaning of the measures in particular educational settings, but also the nature and appropriateness of the implications for action that may be drawn from them (Messick, 1984b).

Furthermore, the new NAEP design also includes provision for special probes to gather timely information in depth about particular policy issues or particular subject areas of immediate interest. Relevance to educational practice is also enhanced in the new NAEP plan by affording interested states the opportunity to conduct state assessments concurrently with the national assessment, administering some or all of the NAEP exercises in the particular wave to state samples selected according to NAEP specifications. The states in turn may encourage school districts to conduct district assessments in the same manner.

Mechanisms for influencing educational policy

With these powerful new capabilities incorporated into the NAEP design, a diverse array of analyses become possible. For instance, with commensurable and interpretable measures of achievement status and trends, one can document relative achievement and change for different demographic groups and identify student subpopulations in likely need of additional services. With data from the teacher and school questionnaires, one can quantify prevalent instructional and school practices and relate them to student performance. By contrasting student outcomes in relation to program and background factors, one can develop a heightened understanding of potential contributors to differential performance and of the varied effects of different learning contexts. Examples of such analyses appear in the NAEP reading trend report (*Reading Report Card*, 1985), in an inquiry into the determinants of student computer use (Lockheed, 1986), in a report of Catholic school reading proficiency in relation to national averages (Lee, 1986), and in a descriptive and path analysis of the reading performance of language minority students in relation to home, school, and process variables (Baratz & Duran, 1987).

The value of such assessment results for educational planning and policiy making depends on the effectiveness with which they are reported and disseminated. Especially important are the appropriateness and comprehensiveness of the targets of dissemination as well as the nature and amount of interpretation provided to clarify the meaning of the results and of their implications for action.

Dissemination for direct and indirect impact

In regard to the targets of dissemination, NAEP has systematically highlighted findings in appropriately tailored form to the general public through print and video media, to school boards, to superintendents, principals, and teachers, to local, state, and federal agencies as well as to legislators, to educational associations, and to the educational research community. In so doing,

it was important to recognize that the impact of such comprehensive dissemination might be more indirect than direct. To be sure, it was hoped that NAEP results and their implications would be taken into account by legislators and education decision makers in shaping policy. But it was deemed at least as important to raise the consciousness of the public, of the education community, and of policy makers about the current status of problems and issues on the educational scene. For example, the *NAEP Reading Report Card* (1985) simultaneously displayed the striking progress in reading performance made by minority groups in recent years along with the finding that Black 17-year olds on the average currently read at about the same level as do White 13-year olds. This juxtaposition underscores not only the severity of a national educational problem, but the enormous efforts likely needed to accelerate even further the current pace toward rectifying the situation.

Interpretation as the evaluation of alternative perspectives

In regard to the nature and amount of interpretation to be provided in reporting assessment results, there are widely divergent points of view. One stance is that the facts should speak for themselves. To be sure, compendia of facts often serve useful purposes, but they are primarily useful in the course of making or defending interpretations. Another viewpoint as expressed by Kaplan (1964) is that «data are the product of a process of interpretation, and though there is some sense in which the materials for this process are 'given' it is only the product which has scientific status and function» (p. 385). Perverse as it sounds, this latter viewpoint — that data are not input to the interpretative process but its product — should prevail because it recognizes the interdependence of scientific inference on facts, theories, and values. That is, the nature of the questions asked, the kind and amount of data collected, the type of analysis chosen to yield results, the form in which the findings are cast are all guided at least tacitly by theories and ideologies. Indeed, it is this theory- and value-dependence of scientific research, and all the more so of policy research, that leads a prudent researcher to evaluate systematically a range of plausible alternative perspectives on the questions, the analyses, and the findings (Churchman, 1971; Messick, 1980). And this inevitably leads to more interpretation, not less.

The reason that these ideological issues appear even more salient for policy research than for social science research generally is that policy research is not only theory- and value-dependent but policy-dependent.[1] That is, in the very formulation of the problem, the policy-maker typically embraces a set of values, either tacitly or explicitly, that places limits on the scope and nature of the applied research deemed relevant, Merton and Lerner (1951) call these constraints «value constants» that circumscribe the alternative lines of action to be investigated. These constraints often take the form of assumptions that certain features of the problem situation are constant or given and under no circumstances to be modified.

Two common types of implicit constraints or unwitting misformulations of a problem are overspecification and overgeneralization. When a policy maker overspecifies the practical problem, researchers must clarify the situation by searching out the prime or fundamental objective, thereby often redefining the problem. For example, the US Office for Civil Rights, charged with ensuring local school districts' compliance with equal protection under the law for

[1] My thanks go to Hiroshi Azuma for suggesting this point as a discussant of the AERA symposium.

minority students, observed disproportionate placement of Black children in special education classes, especially classes for the educable mentally handicapped. Consequently, it asked a National Research Council panel to determine the factors that account for this disproportionate minority representation and to identify placement criteria or practices that do not affect minority students disproportionately. As reformulated by the panel, however, the key problem became one of determining the conditions under which inequality of placement constitutes inequity of treatment (Heller, Holtzman, & Messick, 1982; Messick, 1984a). In contrast, when a policy maker overgeneralizes the practical problem, researchers must clarify the situation by searching out a variety of problem perspectives and action alternatives and by determining the potential consequences of each (Merton & Lerner, 1951).

In most policy research, it is important to examine plausible alternative perspectives concerning two major issues — namely, the meaning of obtained measurements and relationships and the import of implications for action derived therefrom. A given finding or relationship may have plausible alternative meanings, and each of those meanings may imply plausible alternatives for action depending on the circumstances. An appeal to other findings in the data or the conducting of additional analyses may render some of these alternatives less plausible. But in any event, this view of interpretation as the evaluation of alternative perspectives yields not specific policy recommendations, except in clear-cut cases, but rather illuminated alternatives to challenge policy makers to informed choice. It should be noted that the counterstrategy of sticking to the so-called facts, of settling for much less than this level of interpretation in policy research brings with it the prospect of incompletely or inadequately analyzed data.

Alas, this seems to be the fate of large-scale educational assessments. Under pressures of timeliness and limited funding, the evaluation of alternative perspectives is usually relegated to secondary analyses. Ironically, the price that is paid for this delayed enlightenment is timeliness.

But irony aside, we should not lose sight of the enormous potential of large-scale educational assessment as effective policy research — provided that key indispensable conditions are incorporated into the enterprise. Especially important features are the capacities to provide data or measures that are commensurable across time periods and demographic groups, correlational evidence to support construct interpretations, and multiple measures of diverse background and program factors to illuminate context effects and treatment or process differences. Combining these features with the power of multivariate analytic techniques and with an interpretation strategy making explicit provision for the exploration of multiple perspectives can yield cogent clarifications both of the nature of the problem and of the relevant policy alternatives for action.

References

Baratz, J. C., & Duran, R. (1987). *The educational progress of language minority students: Findings from the 1983-84 NAEP reading survey.* Princeton, NJ: National Assessment of Educational Progress.

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organization.* New York: Basic Books.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity.* Washington, DC: National Academy Press.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Lerner, D., & Lasswell, H. D. (1951). *The policy sciences: Recent developments in scope and method.* Stanford, CA: Stanford University Press.

Lee, V. (1986). *1983-84 NAEP reading proficiency: Catholic school results and national averages.* Report to the National Catholic Education Association.

Lockheed, M. E. (1986). *Determinants of student computer use: An analysis of data from the 1984 National Assessment of Educational Progress.* Princeton, NJ: National Assessment of Educational Progress.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J: Erlbaum.

Merton, R. K., & Lerner, D. (1951). Social scientists and research policy. In D. Lerner & H. D. Lasswell (Eds.), *The policy sciences: Recent developments in scope and method.* Stanford, CA: Stanford University Press.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1984a). Assessment in context: Appraising student performance in relation to instructional quality. *Educational Researcher, 13* (3), 3-8.

Messick, S. (1984b). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215-237.

Messick, S. (1985). Response to changing assessment needs: Redesign of the National Assessment of Educational Progress. *American Journal of Education, 94,* 90-105.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report no. 83-1). Princeton, NJ: National Assessment of Educational Progress.

The reading report card: Progress toward excellence in our schools (NAEP Report No. 15-R-01). Princeton, NJ: National Assessment of Educational Progress.

Evaluation pédagogique à grande échelle et politique d'éducation: aspirations et limites

L'auteur examine les aspects les plus importants à prendre en considération pour que l'évaluation pédagogique à grande échelle conduise à des résultats utilisables dans la détermination d'une politique d'éducation. Les conditions à réaliser seraient les suivantes: examiner d'un point de vue développemental des données recueillies séquentiellement, adopter à l'égard des mesures retenues et de leurs relations des corps d'interprétations fondées empiriquement, s'interroger sur le rôle des différents facteurs à considérer, et enfin, évaluer l'intérêt d'autres perspectives possibles, relatives aussi bien aux questions posées qu'aux résultats obtenus.

Key words: Assessment, Interpretability, Comparability, Policy judgment.

Received: November 1986
Revision received: December 1986

Samuel Messick. Educational Testing Service, Princeton, New Jersey 08541.

Current theme of research:

The integration of two programmatic strands of inquiry: The interrelatedness of personality and cognition in human behavior, especially as revealed in creativity and cognitive style, and the interrelatedness of validity and values in measurement.

Most relevant publications in the field of Educational Psychology:

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing policy. *Educational Psychologist, 17,* 67-91.

Messick, S. (1982). The values of ability testing: Implications of multiple perspectives about criteria and standards. *Educational Measurement: Issues and Practice, 3,* 9-12, 20, 26.

Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist, 19,* 59-74.

Messick, S. (1985). Progress towards standards as standards for progress: A potential role for the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 4,* 16-19.