DOCUMENT RESUME

ED 412 766 FL 024 861

AUTHOR	Braun, Henry
TITLE	A Postmodern View of the Problem of Assessment.
PUB DATE	1997-03-00
NOTE	12p.; Revised and expanded version of a paper presented at the Language Research and Testing Colloquium (Orlando, FL, March 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Cognitive Psychology; *English (Second Language); Futures (of Society); *Language Tests; Social Change; Standards; *Test Construction; Test Wiseness
IDENTIFIERS	*Postmodernism

ABSTRACT

A discussion of language testing addresses three questions: why good test construction seems to be increasingly difficult; what forces are shaping the practice of test construction; and what lies ahead in testing. It is proposed that practitioners are constantly redefining what "good" tests are, and those who develop tests are facing greater and more potentially conflicting demands, a common dilemma in the postmodern world. Test design is compared with architectural design in that design is shaped by purpose but must also meet criteria for optimality. In test design, purpose has become more ambitious and multifaceted; cognitive psychology and related disciplines have led to greater understanding of the nature of competence, and more sophisticated models of particular domains. In addition, validity models have become more comprehensive, and standards that testing is held to are becoming more rigorous. It is argued that test designers must learn more about differences in performance among test-takers and understand better the ways in which technology will affect testing. The importance of these factors in the testing of English-as-a-Second-Language competence is emphasized. (Contains 12 references.) (MSE)

A Postmodern View of the Problem of Assessment

Henry Braun
Educational Testing Service

I would first like to thank the LTRC for inviting me to deliver a keynote address at this meeting and, especially, to Professor Antony Kunnan for providing assistance in making the necessary arrangements.

In the time that I have available, I want to address three questions. They are:

> Why does good test construction seem to be an increasingly difficult activity?
>
> What are the forces shaping the practice of test construction?
>
> What lies ahead?

Certainly, I will not be able to fully respond to these questions to anyone's satisfaction, and not only because of time constraints! They are indeed difficult questions and do not admit simple answers.

Let me suggest, though, a short answer to the first question. It is that we are redefining "good" so that there are greater demands on those who must develop tests. Indeed, it is not only that the demands are greater but that they are more likely to come into conflict. This brings to mind a book that I have just read, *In Over Our Heads: The Mental Demands of Modern Life*, by the noted psychologist Robert Kegan. He argues that many of us are living in a post modern psychological state, in which the familiar anchors of family, tradition and religious or civil authority no longer hold sway as they once did. More of us, more of the time, are forced to rely on our own capacities to sort out complicated situations, to make complex judgments and to reach difficult decisions among options that are equally attractive--or equally unattractive.

Kegan makes a strong case that these demands confront us in our roles as spouses, as parents and as workers. So perhaps we who are developing tests are just experiencing the postmodern world firsthand in our own work.

One can think of building a test as a problem that falls under the rubric of "optimal design under constraints." In general, a realized design is a particular combination of design elements or an algorithm for generating such combinations that satisfies certain priori constraints and can be evaluated against one or more orders of merit. Optimality may only mean achieving an acceptable balance among the different orders of merit.

This paper is a revised and slightly expanded version of a presentation made at the Annual Language Research and Testing Colloquium in Orlando, Florida in March, 1997.

From this perspective, test construction may have much in common with other design professions such as architecture. In my view, test designers have been rather insulated from other designers and perhaps we can learn something valuable from the struggles of other design professions to understand what they do and how to do it better. These thoughts have been stimulated by my long-standing involvement in building computer-based simulations of architectural practice as part of a major effort to computerize the entire battery of architectural registration examinations. The research and development during this nine-year period has forced my colleagues and me to grapple with issues in test design, but has also led to a greater appreciation of the practice of architecture itself, and how it has a great deal in common with assessment design.

Some of these similarities are indicated in Table 1 below. In both cases, design is shaped by purpose: What is to be accomplished and for whom. Lack of clarity in purpose or naive overambition often result in poor designs. For both sets of practitioners, critical questions are how to generate candidate designs and how to evaluate them once they are available. The latter question requires explicit criteria for optimality or what I referred to above as orders of merit.

Table 1

ARCHITECTURE	TESTING
Landscape	Domain
Design Elements	Items/Probes
Engineering Constraints	Modes of delivery Scoring procedures Psychometric tools

Table 2a presents some of the criteria employed by architects while Table 2b presents some of the criteria employed by test designers. Obviously, the purpose of the design effort will influence the salience of the various criteria and the ranges of acceptable or desirable values. Except in the most trivial cases, each feasible design represents a tradeoff among the optimality criteria.

Table 2a

ARCHITECTURAL CRITERIA	
Functionality	Structural integrity Traffic flow Space adjacency
Conformity to Code	Zoning restrictions Safety considerations
Aesthetics	Appropriateness to site Visual attractiveness
Cost	Time-to-build Material cost

Table 2b

TEST DESIGN CRITERIA

Measurement	Distribution of difficulty Reliability Comparability Generalizability
Business	Cost Time Efficiency
Validity	Evidential Consequential

One reason the test developer's job has gotten more difficult is that the design criteria have become more demanding. For example, the modern conception of validity changes the scope of the design world by bringing into consideration a broader set of issues, as the following quote from Sam Messick indicates:

> *Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.* (Messick, 1989)

The above assertion should be compared with the more limited requirements of content and predictive validity. In fact, one can imagine a sequence of increasingly elaborate design worlds induced by increasingly demanding validity models. One could argue that the broadened view of tests embraced by much of the public--in contrast to the more limited view held by the testers--goes to the heart of many criticisms of present day tests. A comment that I vividly recall from a meeting several years ago to the effect that "multiple choice tests are psychometrically immaculate but educationally bankrupt," illustrates the point.

Lest we feel alone in the opprobrium we endure, here is a comment from a critic of another design artifact, a zoning code.

> *America's zoning laws . . . have mutated . . . into a system that corrodes civic life, outlaws the human scale, defeats tradition and authenticity, and confounds our yearning for an everyday environment worthy of our affection.* (Kunstler, 1996)

His point, made throughout the article, is that architects and planners must look beyond building design to consider the functionality of the built environment. The point is the same--the need to take account of a broader set of criteria in evaluating the success (validity) of a design.

Indeed, the practice of test design and construction has become much more difficult. In the first place, purpose has become more ambitious and multifaceted. In school assessments, for example, sponsors seek tests that can both provide useful instructional information for the individual student while also serving accountability roles. Secondly, cognitive psychology and related disciplines have led to a deeper understanding of the nature of competence and more sophisticated models of particular domains. Designers must take account of these new understandings in their work. Advances in technology, particularly the rapid evolution of computers and communication networks, are leading to seismic changes in the infrastructure that supports testing. Finally, as has been mentioned just above, validity models have become more comprehensive and the standards the testing profession is being held to have become more demanding and rigorous.

Test designers must cope with the complex and dynamic interactions among these various aspects of the process, in addition to trying to anticipate future directions. Hampered by reliance on old paradigms and the lack of tools to fully exploit scientific and technical advances, they tend to produce tests that are often very much like the tests of the past.

In the case of "high stakes" assessment for selection, purpose is shifting from providing an assessment of overall proficiency along a unidimensional scale to providing an interpretable score profile that informs educational decision-making. Modern validity requires us to consider what kind of data would support the adequacy and appropriateness of inferences and actions based on test results. For designers, the first question is what types of items or probes, what kind of test structures, and which inferential models would generate the sort of evidence required by the different decision makers.

I believe that we have to understand differences in performance among test-takers in terms of various developmental trajectories and their implications for further learning. Thus, the "static" structural perspective of a domain must be joined with a "dynamic" developmental perspective of performance in the domain.

This will have profound implications for the next generation of psychometric models, an issue that is treated very well by Mislevy (1996).

These ideas are by no means new ones, as the following quotations illustrate:

> *. . . modern cognitive psychology conceptualizes the acquisition of cognitive skills in developmental terms. Hence, modern educational and psychological measurement, to enhance its educational usefulness, should be sensitive to developmental differences in subject-matter learning and performance.* (Messick, 1984)
>
> *. . . learning theory is taking on the characteristics of a developmental psychology of performance changes. . . .*
> *. . . measurement must be designed to assess these performance changes. . . .*
> *Coherence of instruction and assessment is the ultimate goal.*
> (Glaser, Lesgold, & Lajoie, 1987)

Until recently, though, these notions have been treated by practitioners as pointing toward idealized goals rather than realistic objectives. However, the development of measures of literacy skills both in large scale assessments and in remedial programs (Kirsch, Jungeblut, & Mosenthal, in press), and the work of Tatsuoka and her associateds on Rule Space Methodology (1997) are important first steps. In the case of adult literacy, a strong theory of competence led to a test design process in which items could be generated to meet specific difficulty targets and different score levels could be given firmly grounded functional interpretations. Rule space methods, when successfully applied, allow cognitively based interpretations of test performance that meaningfully differentiate among individuals at different score levels and even among individuals at similar score levels but with qualitatively different response patterns.

Contemporaneous work by Gitomer and associates (1991) and Mislevy (1996) have shown that we are at the threshold of developing technology-based integrated modular assessment systems that can be tuned to support a range of purposes from instructional assessment to high stakes assessment. These systems are characterized by domain models derived through cognitive task analysis, student models that are informed by the understanding of the nature of expertise and its acquisition, as well as statistical models employing Bayes inference networks that support dynamic assessment and the continuous updating of student models as additional evidence accumulates. These are exciting developments and promise to revolutionize the practice of assessment. They also imply a need for a radical revision in the test design process.

Until this point, I have focused on the impact of validity on test design. In contrast, attention typically tends to be directed toward the impact of technology. Indeed, there is no question that technology advances will influence the design world in many ways, as illustrated in the table below.

Table 3

IMPACT OF TECHNOLOGY
Items/probes utilizing multimedia
Psychometric models relying on rapid real-time computation
Automated scoring of complex constructed responses
Dynamic (adaptive) test designs
Multiple delivery options (test centers, worldwide web)
Cost structures dominated by "seat time"

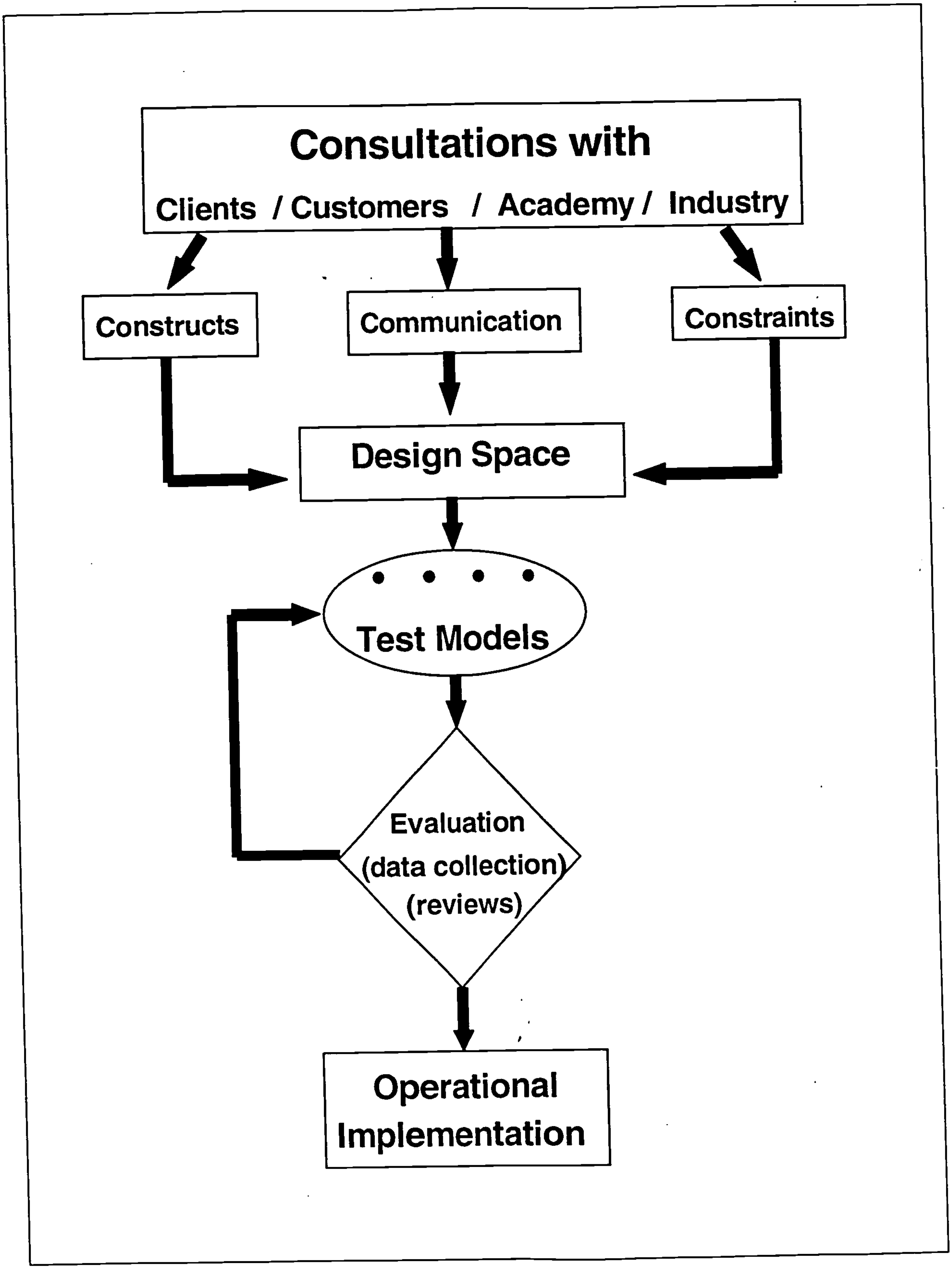
It is also important to recognize areas that technology may influence only indirectly. For example, the demand for authentic performance assessment coupled with multimedia capabilities will lead to the need for automated scoring of complex student-produced responses. In another forum (Braun, 1994), I have argued that the development and implementation of these expert systems will lead to more rigorously defined tests with improved measurement properties. In particular, in order for an automated scoring system to operate accurately for a wide variety of instances of a particular problem type, developers are forced both to craft tighter problem specifications and to clarify the rules of evidence for scoring. This leads to greater comparability over time which is particularly important in an "on-demand" testing environment with the concomitant requirement for large item pools to maintain test security. This has certainly been the case in the architectural licensing effort. See also Bejar (1995).

As the design process becomes more clearly delineated, technology will also facilitate a more experimental approach to the practice of test construction; that is, it will be possible to take a more generative approach, in which multiple candidate designs can be produced and then examined, leading to new cycles of generation and evaluation until a satisfactory design is found. This technique of automated design generation is being practiced in such disparate areas as architecture and biology with interesting results.

In fact, it is already serving us well at ETS in various investigations. We are employing Automated Item Selection (AIS), a tool developed originally by Swanson and Stocking (1993) to provide near final form linear tests; and now, also, to produce computer adaptive tests operationally in real time. At the heart of the system is a clever dynamic optimization algorithm that sequentially selects items from a pool so that the final result is a test that meets the varied constraints and requirements that embody the target construct. It is now used to generate multiple instances of a test under a particular set of conditions, permitting developers to experimentally determine the effects of different combinations of constraints or different item pool compositions on the properties of the resulting tests. Such a program of research would never have been feasible in the past when the assembly of a test could require as much as four days and not four minutes!

One model of a revamped test design process is presented in Figure 1 below.

Figure 1



In this scenario, consultations with various constituencies provide test developers with three essential building blocks: 1) the **constructs** or underlying targets of the measurement process, 2) the **communication** goals or the kinds of information that is to be conveyed on the basis of the test results, and 3) the **constraints** or the relatively unchanging features of the setting in which the test will be designed, developed and delivered.

Together, the three "C's" determine the design space, the universe of feasible test designs that conform to the three C's. Various candidate designs can then be generated by different means, with the goal of exploring different regions of the design space. These designs are evaluated using appropriate criteria. On the basis of these evaluations, one or more of the designs can be modified or entirely different designs can be generated. After some number of cycles, a satisfactory design is attained and operational implementation commences.

Of course, this is a highly simplified view of the test development process. Nonetheless, there is a key notion of a generative phase in which an explicit effort is made to examine the attractiveness of a variety of very different designs. This is not standard practice and the usual result is a lack of innovation in the design process.

With all the excitement attendant on the role of technology, it is important to note that technology changes neither the purpose of measurement nor the criteria by which we judge the adequacy of an instrument with respect to the demands of contemporary psychometric practice and test validity theory. In my view, if the design profession takes the modern conception of validity seriously, the consequences for assessment will be as great as the more visible effects of technology.

Validity theory compels us to adopt a more ecological approach to test construction by fundamentally broadening the scope of the design world. Indeed, elaborating the theoretical and practical implications of validity theory is essential to forestalling the ascendancy of an impoverished techno-centric approach to test design. It is only by respecting the emerging validity standards and employing technology thoughtfully that we will, over time, produce better tests--tests that are generated through a craft of test design that is at once more principled, more disciplined and more innovative.

These ideas are particularly germane to the area of language testing. For millions around the world, English language competence is the key to information, educational opportunity and employment. In ESL testing our purpose should be to help people realize their educational and career goals, while assisting institutions in making the resource allocation decisions they must. A successful and valid assessment will have to take into account such factors as: the multiplicity of purposes, the heterogeneity of language backgrounds, differential instructional strategies, as well as the role of psychological and social psychological factors in performance.

This is a complex and challenging undertaking that will, I am convinced, defeat ordinary test development practice. Indeed, I believe that serious consideration of the ecological approach to test design in this area will lead us to the construction of assessment systems that will support both extended instruction and relatively short certification episodes. This will lead to fundamental changes in the practice of assessment and promises an exciting future for all of us.

References

Bejar, I. I. (1995). From adaptive testing to automated scoring of architectural simulations. In E. L. Mancall & P. G. Bashook, Assessing Clinical Reasoning: The Oral Examination and Alternative Simulations. Evanston, IL: American Board of Medical Specialties.

Braun, H. I. (1994). Assessing technology in assessment. In E. L. Baker & H. F. O'Neill, Technology Assessment in Education and Training. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Gitomer, D. H., Cohen, W., Gallagher, A., Kaplan, R., Steinberg, L. S., Swinton, S. & Trenholm, H. (1991). Design rationale and data analysis for Hydrive content and structure. Princeton, NJ: Educational Testing Service.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive psychology. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Kegan, R. (1994). In over our heads: The mental demands of modern life. Cambridge, MA: Harvard University Press.

Kirsch, I., Jungeblut, A., & Mostenthal, P. B. (In press). Interpreting the adult literacy skills and literacy levels. Technical Report for the National Adult Literacy Survey. Washington, D.C.: National Center for Education Statistics, U.S. Government Printing Office.

Kunstler, J. H. (1996, September). Home from nowhere. The Atlantic Monthly, pp. 43-66.

Messick, S. (1989) Validity. In R. L. Linn (Ed.), Educational Measurement (3rd. ed., pp. 13-104). New York: American Council on Education/MacMillan Series on Higher Education.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21(3), 215-237. (Invited Address, National Council on Measurement in Education and American Educational Research Association.)

Mislevy, R. J. (1996). Test theory reconceived. Journal of Educational Measurement, 33(4), 379-416.

Swanson, L. C., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement,17, 151-166.

Tatsuoka, K. (1997). Computerized cognitive diagnostic adaptive testing: Effect on remedial instruction as empirical validation. Journal of Educational Measurement. 34(1), 1-20.

FL 024861

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A Postmodern View of the Problem of Assessment

Author(s): Henry I. Braun, Educational Testing Service

LTRC presentation? X yes — no If no, was this presented elsewhere? — yes — no Specify:

Publication Date: March 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

□

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title:
Henry I. Braun
Vice President for Research Management

Organization/Address:
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Telephone: (609) 734-1239

FAX: (609) 734-5010

E-Mail Address: HBraun@ets.org

Date: 11/5/97

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on
Languages & Linguistics
1118 22nd Street NW
Washington, D.C. 20037

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to: