IMPROVING

LARGE-SCALE ASSESSMENT

Resource Paper No. 9

Pamela E. Aschbacher
Eva L. Baker

Editors

National Center for Research on
Evaluation, Standards, and Student Testing (CRESST)

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, CA 90024-1521

1991

The project presented or reported herein was performed pursuant to a grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of OERI/ED, and no official endorsement by OERI/ED should be inferred.

TABLE OF CONTENTS

PREFACE

This volume, *Improving Large-Scale Assessment*, was initiated as part of the Monitoring and Improving Testing and Evaluation Innovations (MITEI) Project conducted at the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA from 1986-1990. The goal of this project was to support the quality of large-scale student assessment programs by encouraging dialogue among testing directors, measurement experts, commercial publishers and policymakers, and by disseminating informative papers on key issues in the field. This volume was originally created as a looseleaf notebook with new papers added annually. The final set of papers from the project has been bound together here for more cost-effective dissemination. Those interested in additional papers on topics related to student assessment in general and alternative assessment methods in particular are encouraged to contact CRESST for copies of relevant publications.

Eva L. Baker
Pamela E. Aschbacher

Improving Large-Scale Assessment

PART I

GUIDELINES

For the RFP Process

Pamela E. Aschbacher
Eva L. Baker
Joan L. Herman

January, 1988

PART I: GUIDELINES FOR THE RFP PROCESS

Table of Contents

The Review Process

PREFACE

The Model RFP Project was developed in collaboration with state testing directors as an approach to improve the technical quality of state tests that assess educational performance of students and teachers. To this end, project personnel conducted interviews and surveys, reviewed recent requests for proposals (RFPs) for state testing from many different states, and held a two day meeting at UCLA in May, 1987, which was attended by three representative groups: state testing directors experienced in the RFP process, commercial test companies who bid on such RFPs, and researchers from the academic measurement community.

The project's original objective was to develop model language for a state assessment RFP. However, during the course of the project's activities, an urgency for improving the entire RFP process was revealed, particularly among testing directors and vendors, and the focus of the project expanded.

At the Model RFP Project's 1987 meeting, participants decided to address problems in the generic RFP process as well as issues in specifying standards of technical quality in RFPs. Participants discussed various choice points and options in the RFP process and the treatment in RFPs of such technical concerns as equating, item bias, and content validity.

After a general discussion of RFP procedural problems and three technical issues, participants divided into two groups: one that focused on improving the RFP process, and one that focused on technical concerns.

The RFP process group devoted its time to articulating choice points and options in the RFP process. Members of the group voiced concerns about a gamut of issues, including those related to fairness, quality assurance, cost, and communication

of expectations. The group tried to take into consideration the differential constraints of state regulation and the competitive nature of the RFP process. Part I of this document summarizes the Guidelines developed by this group. The work on technical quality issues is presented later.

We wish to thank the following people who participated in the May meeting and reviewed drafts of this document. Their expertise, time, and overall support for the project were generously provided and have been invaluable.

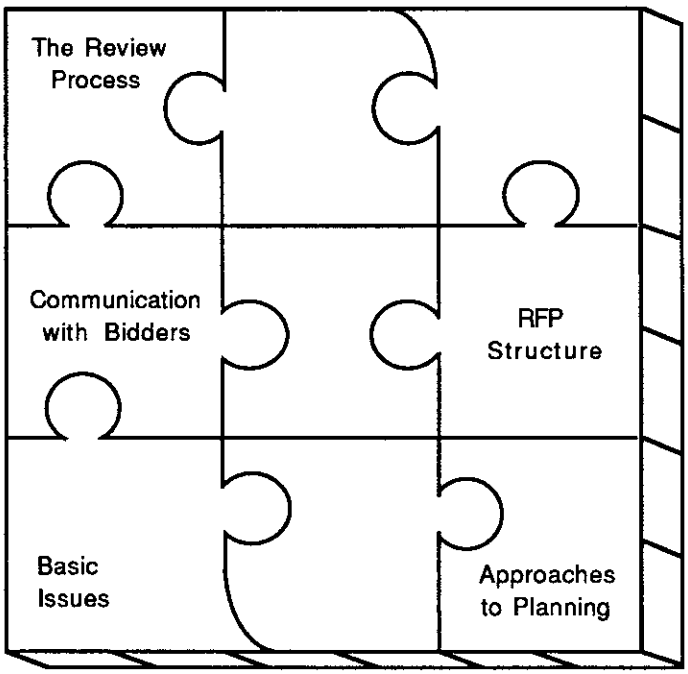
Joan Baron	Stephen Koffler
William Brown	Elaine Lindheim
Leigh Burstein	Wayne Neuberger
Ronald Hambleton	W. James Popham
H.D. Hoover	Edward Roeber
Richard Jaeger	Paul Sandifer
John Keene	Ramsay Selden
Thomas Kerins	Stephanie Zimmermann

We also thank Lawrence Rudner and Tom Fisher for their helpful reviews and suggestions.

Although these guidelines specifically address the RFP process, we recognize that many states and districts use other procedures for obtaining their tests, including in-house development and development through special relationships with state universities or other local organizations. We hope that our guidelines will prove of some use in these settings as well. Despite the fact that we originally intended the audience of our project to be state testing directors, we hope that our efforts will also be useful to district testing personnel.

E. B.
P. A.
J. H.

Guidelines
for an Effective RFP Process



Solving the Puzzle of Large-Scale Assessment

INTRODUCTION

Purpose

These Guidelines for RFPs articulate key choice points, options, and considerations faced by state (and district) testing directors as they solicit services for large-scale assessment from commercial testing companies. Since so many states and districts write Requests for Proposals (RFPs) to procure services from test service vendors, we have focused these Guidelines on the RFP process.

The questions and discussion presented in the Guidelines are intended to broaden the scope of issues and concerns that state and district testing officers consider in preparing RFPs for large-scale testing programs. Many of the same issues and problems are faced by states and districts that use methods other than RFPs to plan and obtain testing services, so we expect the Guidelines to be of some use to them as well.

Prior to our work on this project, many of the issues described here had not been discussed openly among state testing directors and vendors. The frank discussions that developed among experienced testing directors, representatives of major commercial test companies, and measurement specialists during project sessions highlighted ideas and explicated options that should be useful to others involved in procuring large-scale assessment services. We hope that this effort will be particularly useful to new testing officers and those not part of a supportive network. Commercial test vendors may also be interested in these guidelines.

We have tried to present a logical, straightforward approach to planning and developing RFPs for large-scale

assessment. Our approach has been guided by the assumption that one of the primary goals of an effective RFP process is to obtain a reasonable number of proposals that are "on target," that is, which speak directly to the needs of the state or district. However, there are significant tensions inherent in the nature of the RFP process that obstruct the establishment of a set of failsafe procedures. These tensions were referred to frequently during our project meeting to explain states' and vendors' behavior or motivation in a variety of circumstances. A brief overview here of the way these forces operate will delineate our own point of view in developing these materials and will facilitate your implementation of these Guidelines.

Tensions in the RFP Process

The provision of testing services is big business, and this results in competing interests: constrained versus open communication, creative approaches versus use of carefully specified detail and adherence to RFP requirements, and cost versus quality. In addition, the limitations imposed by local or state policies and procedures inhibit implementation of an ideal, logical RFP process. Herein lies the root of many of the decisions made and the difficulties suffered by both testing directors and vendors.

In an attempt to preserve fairness in a very competitive atmosphere, states develop policies that sometimes result in very limited communication between state and bidders. For example, some states do not allow testing directors to talk directly with bidders prior to the proposal deadline. All communication is routed through the purchasing agent. Many states try to maintain fairness by soliciting questions from bidders and then sending a written document of all questions and answers to every bidder before the deadline. While this practice would be expected to facilitate fair communication, it

often fails to do so. Because of the intense competition, bidders tend not to ask significant questions that might reveal their approach to a problem or clarify for other bidders as well as themselves important aspects of the state's needs. Some bidders also hesitate to ask questions for fear of revealing their poor comprehension of the project at issue. In addition, by the time bidders receive the written answers, it may be too late to use this information to modify their proposals.

The communication problem is exacerbated by the fact that many RFPs are quite vaguely worded for any of several reasons. Sometimes RFP authors are uncertain about the purposes or details of a new program, especially during the early conceptual stages of the program. In some cases, testing directors are faced with such long timelines in getting RFPs authorized, reviewed and accepted by various state officers (up to a year in some cases) that they must dovetail tasks very closely in order to meet a deadline. This may result in their having to write the RFP before some of the essential groundwork has been completed. For example, they may not be able to describe the number of objectives and items to be developed because the objectives committee has not finished composing the list.

Another cause of vaguely-worded RFPs is the desire to enhance competition as a means to obtain better test services for less money. Some states have provided limited specificity to bidders, especially in terms of the level of effort and scope of work expected. The notion is that less information will spur the creativity and competitiveness of the bidders. The consensus of our meeting, however, was that this notion is a myth. Explicit information in an RFP about the expected cost and scope of work facilitates rather than inhibits good proposals.

Vagueness, on occasion, does yield a highly creative, reasonable bid. However, bidders' solutions to vaguely-stated requirements may be so diverse in scope, quality, and cost that bids cannot be fairly compared. Furthermore, this situation may open the door to legal challenges of the bid process by one or more of the bidders.

Since cost is so important to states, budgetary concern often shapes the focus of the planning effort and the RFP itself, sometimes to the detriment of technical quality issues that need to be addressed. For example, much attention may be devoted to specifying the number of meetings to be held, where they will be held, who will attend, and how much is budgeted for lunch, but no specifications may be provided for the *purpose* of the meetings (e.g., how to establish the content validity of the test). If the RFP emphasizes process over purpose, the resulting tests may suffer in quality as a result.

Concern for costs may also directly affect district or state policy, impacting bidders and in turn the alternatives available to the state or district. If it is known that a state contract is likely to go to the lowest bidder, because of either official or practical policy, vendors with possibly better plans but higher costs often will not bother to submit bids. States represented at our Model RFP Project meeting said their RFPs typically solicited bids from only three to four vendors. Given the size and importance of many projects, most states would prefer a greater choice of proposed services. If none of the proposals is adequate, additional time and money may have to be spent to issue a revised RFP, or the state may have to shelve the project.

Although states tend to think of competition as resulting in lower costs, they must also be aware that quality may be compromised. There are many reasons why a vendor, large or small, may provide a low price, and this may or may not benefit the state. For example, a test company may

underbudget a proposal for an early test development effort in a state (taking a loss on that contract) to position itself favorably for the larger, related testing contracts that may occur there in the future. The first contract may yield a high-quality, low-cost test—a real bargain. However, the bargain may be balanced later by the cost of future contracts, either with the same vendor who must recover his costs eventually, or with another vendor whose costs may be higher because he was not involved in the original work on that program. In any case, the state should attempt to protect the quality of its programs by specifying its expectations for quality in the RFP, providing for reference checks for vendor performance, and carefully examining the consequences of its low-bid policies.

Organization

The following Guidelines for the RFP Process are organized into five sections and are arranged in an order that follows the chronology of the RFP process.

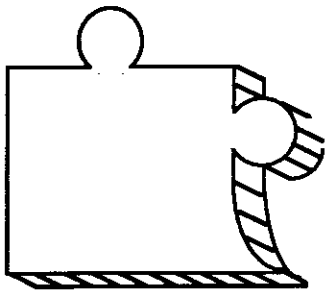
The first section covers basic issues that you should consider prior to any RFP writing: the amount and type of money available, the amount and flexibility of time available, the degree to which the project calls for innovative approaches, whether the bid is to be competitive, and whether there will be one or more phases of RFPs required to accomplish the entire project.

The second section discusses the pros and cons of two methods that can be used to facilitate the planning of new programs before you write the RFP: concept papers and planning meetings. These methods involve obtaining expert advice on how to handle a new project or approach a particularly thorny technical issue.

The third section describes methods to facilitate bidders' understanding of what is required in the RFP: letters announcing upcoming RFPs, bidders' conferences, and methods of handling bidders' inquiries.

The fourth section discusses the major sections of an RFP and emphasizes the importance of articulating your needs and priorities to assure a top-quality and cost-effective product. Topics include the introduction of the RFP, expected cost of contract, the scope of work, technical design and report, expected services and products, personnel loadings, the RFP budget, and quality control and scheduling.

The fifth and last section is about structuring the review process and pre-contract negotiations to assure an economical product of the highest quality. Included here are discussions of specification of criteria and process, relative weights of technical quality and cost, members of the review panel and their qualifications, usefulness of oral presentations, and factors to negotiate before the contract is finally let.



Basic Issues

This section considers five fundamental aspects of the testing project that will significantly influence the planning of the RFP process and document:

1. Money
2. Time
3. Type of Project
4. Type of Bid
5. Single vs. Multiple Phases

1. MONEY

How much money is available for the testing project? Is the money fixed, flexible, or a combination?

The amount of money available for a project is a critical factor in what the project can hope to accomplish and thus in the scope of work set forth in the RFP, so test directors have a responsibility to budget well before issuing an RFP. It is important to consider whether funding has been granted for the entire project or only for a portion of it. The amount of funding and how it is scheduled will influence the number of RFPs that you will write for a project. Such information will also be important to communicate to bidders and should be included in the introductory section of the RFP itself, if that is

allowed in your state. In addition, certain contracts can be awarded in whole or in part.

Whether the money is fixed, variable, or a combination will also influence how you write the RFP. If the money is variable and the final cost of the project turns out to be more than expected, you may have to deal with the added effort, expense, and bother of rewriting the contract. One strategy to avoid rewriting when actual costs are unknown is to estimate the biggest possible number of students to be tested, materials to be printed, and so forth. Then the needed money will be available without having to rewrite the contract. Contracts should have a provision that payment will be for work done. For example, if the RFP calls for budgeting on the basis of 100,000 students, and only 85,000 are tested, the vendor's bill for that work ought to reflect the 85,000 students.

For many projects you may want to include both some fixed costs and some variable ones, such as scoring costs per student or optional reports paid for by the state that individual schools might elect to receive. In the end, the cost to the state will be the fixed costs plus the variable costs minus any credits (given by the vendor for contracted work that you mutually agreed to omit in return for a credit) and minus any penalties (such as those assessed for late delivery of materials or services). Requests for Proposals can have an options section with items to be included should funds become available. This can eliminate the need for new procurements.

2. TIME

How much time is available? Is there any way to increase the flexibility of time schedules?

There are several major constraints that may affect the time available for a project and its degree of flexibility: legislative mandates, funding tied to the fiscal year, contingency of part of a new project on previous work having been completed, and the time required by purchasing agents and others with whom you must work to let the contract. These constraints will affect your plans for completion of the project and will determine a portion of the information communicated to the bidders prior to issuing the RFP and in the RFP itself.

When a project is mandated by the state legislature, the timeline is usually non-negotiable. Your best bet, where feasible, is to try to influence the generation of the legislation *before* it is actually passed. Since mandates differ in their level of prescription, it is obviously to your advantage to encourage a less prescriptive mandate in which you may be able to set up at least some of your own timelines and may be able to deviate from them when it proves necessary. Failing that, you can attempt to work with vendors to educate legislators, governors, and their aides regarding what sort of timeline would be minimally adequate for accomplishing a high quality project. (CRESST hopes to address this problem in the near future by including policymakers in the dialogue for improving the nature of large-scale assessments and by providing states with materials or other means to communicate to policymakers the importance of quality in large-scale assessments, and the importance of having adequate time and resources to achieve the desired quality.)

The contract dates for some testing programs are set by the state's Department of Education (DOE), and are usually tied to the fiscal year. In this case you must help the DOE set reasonable timelines with sufficient flexibility before you write the RFP. Scheduling flexibility is further enhanced by the ability to carry over funds from one year to the next. Project schedule, cost and quality are in a delicate balance, and it is important that contracting agencies and vendor organizations be aware of this. When schedules are compressed, costs may increase because of the use of additional staff, overtime pay, courier services, etc., and the number of quality assurance steps may be reduced or eliminated.

Large-scale assessments rarely exist in isolation. Such testing programs sometimes resemble a very complex puzzle comprised of many small parts that must be integrated in terms of time. When writing an RFP for part of such a complex situation, it is important that you consider what work may need to be accomplished before following portions can be done and then structure the timelines and RFPs accordingly.

In some cases, a great deal of time must be budgeted to shepherd the RFP through a variety of required administrative procedures, such as gaining authorization for the RFP (which may take up to eight months), writing and reviewing boilerplate sections of the RFP, scheduling when the RFP will be issued, scheduling a pre-bid meeting, listing potential vendors, scheduling the review committee, preparing insurance forms, and so forth. Sometimes Commissioners of Education or Assistant Commissioners can help to expedite this process.

Time must also be budgeted for the bidders to respond to the RFP, and thoughtful responses require time. Typically, about four to six weeks are allowed, but two to four months (depending on project complexity) would be desirable. Scheduling pressures are often increased by deadlines, such as

the need to start the project before the end of the fiscal year or the need to field the assessment by a certain time in the school year. Unfortunately, some of these tasks and deadlines may be outside the control of the testing director. In fact, you may need to work with purchasing agents, accountants, and others who know little about testing services.

In this situation it is imperative to anticipate these constraints and to organize and dovetail tasks to minimize negative effects on the program. In some cases you may be forced to use language in an RFP that is more general or vague than you would like simply because there is not sufficient time to wait until you know all the details. When there is little or no flexibility in your time schedule, you may be able to gain some flexibility by carefully wording the RFP and fully using pre-RFP opportunities to communicate with vendors.

3. TYPE OF PROJECT

Is the testing project conventional or innovative? Does the technology for solving the problem exist or does it need to be invented? How committed is the state to a particular approach or specific solution?

There are three very different types of testing projects: (a) those in which the new project is to be an extension of an ongoing program or to replicate a model program; (b) those in which the new program is to differ significantly from what has been done in the past; and (c) those in which a new program is to be implemented where no program existed in the past. It is critical to tell the bidders which sort of program you want.

In the first case, in which a previous program or approach to a problem is to be replicated, it may be important

to tightly specify what has been done in the past so that it can be repeated, right down to the number of items and the number of test booklets. If you are modeling a program after one used in another state, it is particularly useful to describe the similarities and differences between the two programs. Testing directors need to be careful, however, not to imply that a project is more complicated than it really is. The more complex a project appears, the more vendors tend to budget for it. A misleading description may result in the state being overcharged.

In the second and third cases, where significant change or innovation is called for, you should be as specific as possible about where the innovation is desired. If you want a creative approach in one area, such as type of test item, but are committed to a particular approach or solution in another area, such as type of analysis, it is imperative to communicate that to the vendors. It is also important to carefully state the criteria for judging proposals, distinguishing between what is "required" and what is "desired but not necessary." Remember that if the review process goes by the numbers, *requiring* a creative approach means eliminating proposals that may have good but not "creative" approaches.

It is also essential to outline for the bidders any givens, decisions, or constraints within which creative solutions must work. This should include a detailed description of any previous related projects, particularly those aspects which should be avoided and those which might offer clues to success in new approaches. Validation and development costs will probably be higher with innovative projects, and implementation will probably take longer.

Many states include language in their RFPs stating that all ideas in the submitted proposals become the property of the state. States may then use any good ideas in the proposals

without having to contract with the vendors who proposed them. While this outcome may be useful to the state, the provision may restrict the expression of good ideas in proposals and discourage some vendors from bidding at all. Since certain states require that all proposals become the property of the state, placing them in the public domain, you may have no control over this situation.

4. TYPE OF BID

Should the bid be sole source or competitive? Are there a number of vendors who could do the project well or really only one?

For most large-scale assessment programs there are many vendors who might be able to do the project well, and states are usually well served by the competitive bid process. In fact, it often would be premature for the state testing director to decide that only one vendor could or should have the contract. In some states, directors do not have this option. However, it is occasionally quite obvious that only one vendor (who possibly is subcontracting part of the work) is in a position to do the desired project, and then it is better and more efficient to work directly with that company if possible. Other vendors will not lose the time and resources involved in making a bid that would not have been seriously considered.

In some states you may not have to even write an RFP if there is a sole source for the project. In the case where a sole source must still submit a bid, you may be able to help them put their bid together, which will improve the eventual contract.

Some states feel that continuity with a single contractor over several years of a continuing program is important. In fact, the RFP process requires so much internal effort that at least one state is moving toward more five-year RFPs. However, one testing director recommends one- or two-year contracts in the beginning to avoid trouble through lack of experience.

5. SINGLE VS. MULTIPLE PHASES

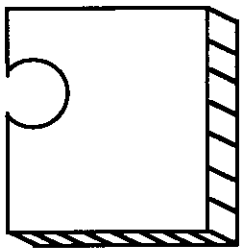
Should there be more than one phase of an RFP to accomplish the project?

In some cases a single RFP will be all that is necessary to solicit the work to accomplish a program. However, many programs may be better served by multiple phases of RFPs. For example, a very large, complex, or innovative program may be best served by a multiple phase proposal process in which later development or implementation is dependent on an effective design or prototype produced during the first phase of the project. Each phase of the project may differ considerably in the amount of time and money available, the degree or type of innovation desired, and its suitability for sole source or competitive bids.

Sometimes you may not be able to find the type or quality of work you want for all aspects of a project at an affordable price from a single vendor. In this case, you may want to break the project into parts, each part to be done by a different company. For example, one vendor might develop the test and provide a camera-ready copy and a second vendor might provide shipping, printing, data analysis and reporting. Dividing a program between two or more vendors makes the specification of responsibilities and timelines of each part of

the program critical. It is also critical to state in the RFP which tasks may be split among vendors, since bidders often base costs, quality, and schedule on integrated processes.

Dividing the RFP into multiple phases can be problematic. Coordination strategies must be put in place to integrate multiple contingencies, monitor compatibilities, and prevent critical aspects of the project from being overlooked. The more people involved in the project, the more deadtime is necessary at the beginning to establish mandatory coordination. Geographic separation of the companies usually makes coordination more difficult, expensive, and time consuming. In addition each separate contract multiplies the red tape and amount of time required to develop the RFP itself. To minimize some of these problems, you can encourage the main vendor to subcontract.



Approaches to Planning

This portion of the Guidelines discusses a couple of methods to assist the planning of new programs prior to writing the RFP, including:

1. Concept Papers or Individual Comments
2. Planning Meeting

1. CONCEPT PAPERS OR INDIVIDUAL COMMENTS

Should there be some sort of pre-RFP communication among state testing officers and others (such as measurement specialists and vendors) that invites concept papers about proposed testing plans or individual comments on a rough draft of the RFP? Should an outside consultant be hired to help with the RFP? How can inequities in distribution or in competitive advantage be avoided?

When considering a significant new program or new technical approach, you may want to gather expert input while conceptualizing the project, before finalizing the RFP. Unfortunately, this choice is often precluded by lack of time. When time permits, input from measurement specialists and vendors may be quite useful. You could consult them about state-of-the-art technical approaches to such areas as bias, validity, equating, or the scoring of writing samples. However, some vendors may be reluctant to give away their good ideas.

Pre-RFP concept papers or requests for individual comments could also be used as an initiation to the bidding process to identify qualified bidders. You could send your ideas for a new type of test to a number of vendors for their suggestions and comments, then invite some of them to respond to your RFP if you liked their initial response. To avoid charges of unfairness, the criteria used to select the initial grouping of responses should be specified ahead of time. Note that in some states it may be illegal to request pre-bid concept papers or invite only some vendors to respond to an RFP.

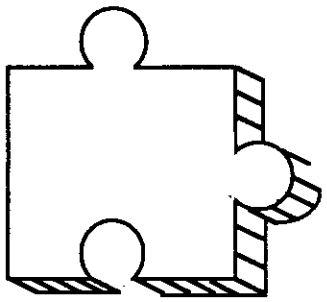
When requesting papers or comments, you should be open about (a) who can respond with the first comments or concept papers, (b) whether the ideas expressed in the comments or concept papers will belong to the state to possibly use in its ensuing RFPs, and (c) whether the eventual RFP bidding will be open to anyone or only to those who have been selected as "qualified" during the comments phase.

2. PLANNING MEETING

Should the state hold a planning meeting with vendors or test experts on how best to handle a new project or a thorny technical issue? Will the state pay for the cost of the meeting?

Another approach to gathering expert input prior to writing the RFP is a planning meeting involving measurement specialists and/or vendors. Group interaction can evoke many perspectives and provide insightful solutions; however, don't expect consensus. A meeting will usually raise more questions than it answers, but raising these questions early may preclude major problems later in the process.

Measurement experts may be able to provide useful ideas on solving problems, conceptualizing the issues, and recommending procedures, but they need to be people familiar with the complexities of real world testing, not just academic ideas. Vendors also may be able to provide good insights about the problem or issue, but they may not wish to share these ideas at meetings where competitors are present. A written response, such as a concept paper, may suit them better. However, the testing project under consideration would have to be quite major to induce vendors to spend this much time on it without any assurance that they would get the final contract.



Communicating with Bidders

This section deals with methods of informing bidders about the RFP to enable them to fully understand what is required. The methods are:

1. Letter to Announce Upcoming RFP
2. Bidders' Conference
3. Bidders' Inquiries

1. LETTER TO ANNOUNCE UPCOMING RFP

Should an introductory letter be used to announce an impending RFP?

An introductory letter sent about a month in advance of an RFP has a couple of advantages. Perhaps most important, it lengthens the response time to an RFP, effectively doubling it in many cases. This permits vendors more time to consider whether and how to respond to an RFP. It allows them time to plan ahead for possible staffing allocations, to shift priorities, and to organize their time, all of which is important, particularly for small vendors. An introductory letter also provides a crucial period when vendors can ask clarifying questions in those states where communication is virtually cut off (other than through a purchasing officer) once the RFP is issued.

2. BIDDERS' CONFERENCE

Should a bidders' conference be held? Should attendance be mandatory?

A conference theoretically informs vendors about the state's needs, priorities, and commitments. This knowledge would be particularly helpful to vendors when the RFP is not very specific on certain points or when the project is expensive, complex or very innovative. The resulting proposals should more likely be on target.

Few conferences, however, are actually as useful as they could be. A bidder's main reason for attending is often to see who else is bidding. The usefulness of a conference in clarifying important details of an RFP tends to be limited by the competitiveness of vendors. They tend to be very guarded about the types of questions they ask to avoid giving away information to their competitors. Some meetings have been as short as ten minutes because no one wanted to ask any questions. Nonetheless, conferences can serve a useful purpose. Regardless of how routine the testing project might be, new vendors may want to bid, past years' practices may be changed, and so forth, and the meeting can be the source of important information.

Attendance at conferences may be mandatory or optional. A mandatory conference may help a state to weed out vendors who are not interested enough to send someone to a required meeting. However, the cost to vendors of sending staff to a conference, particularly one far away, must be recovered by future business. Thus, the states, as consumers, will inevitably pay for any vendors' costs associated with such conferences. Some small vendors may simply opt not to respond to RFPs that require attendance at such meetings, so the effect may be to limit the number of vendors who send proposals to a state

that uses this procedure. Some states hold conferences at which attendance is optional, allowing the vendor to weigh the advantages and disadvantages of participating.

Taping a conference is recommended for several reasons. The taped record can clear up misunderstandings, help reviewers, and protect the state from court action if necessary.

If the state knows exactly what it wants (such as continuing an ongoing project), and what it wants is fairly routine, a meeting is probably a waste of time and money for both vendor and state as long as the RFP is quite explicit.

3. BIDDER INQUIRIES

How can bidders' inquiries be fairly handled?

States handle bidder inquiries in a variety of ways, depending in part on state regulations. Their strategies range along a continuum from forbidding the testing officer to talk to any bidders during the RFP period to allowing any and all communication between a bidder and the testing officer at any time. One middle-ground approach is to send all bidders a written list of all questions raised and the answers before the proposal deadline. Each approach has its advocates and its advantages and disadvantages. Three illustrative approaches are described below:

Approach 1: All questions are referred to the purchasing officer; the testing officer is not allowed to talk directly to any bidders during the RFP period.

The purchasing officer should be a conduit, receiving calls from vendors, passing them on to the test director and relaying the answers. In some circumstances the purchasing officer may try to shorten the circuit and answer testing questions himself. He may not be knowledgeable enough to answer testing questions, and if he does not seek out answers in a timely fashion, the bidders will be left in the dark. However, this approach leaves all bidders in the same boat, so it is "fair" to all.

Approach 2: Questions may be asked, sometimes only in writing, and all answers are sent in writing to all bidders before the proposal deadline.

This approach is fair to all bidders; however, vendors may ask few substantial questions in order to avoid cueing competitors about their plans or the state's needs. In this approach, it is the test director's responsibility to minimize turnaround time, allowing bidders the opportunity to use the information sent to them before proposals are due.

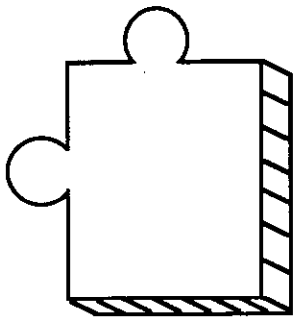
Approach 3: The testing officer can talk to anyone at any time (and may or may not send out written answers to all).

This approach leaves room for favoritism, particularly if written answers are not sent to all bidders. In this case the testing director should be careful not to give information to one vendor that would help him write a better proposal than anyone else. Proponents of this approach feel that it allows maximum communication between state and vendors, which may benefit both. It also rewards vendors who are savvy enough to ask good questions.

Opponents of this approach suggest that it is all too likely to result in litigation by vendors who don't get as much information as others. Even if all vendors do get the same

information, the appearance of bias can have a very negative effect.

The importance of communication between state testing officers and vendors may be underscored by an anecdote. A vendor, hoping to win a contract by supplying a superior quality bid in a highly competitive situation, wanted to include some techniques for reducing item bias in his proposal although the state's RFP did not mention the subject. The vendor feared that if he included techniques for dealing with item bias in his proposal and the state did not carry them out for some reason, the state's failure to do so might be held against it in court should there be litigation in the future (which was probable in this circumstance). The vendor's dilemma was whether to propose their best work and expose the state to some risk, or to do a lesser job, risk losing the contract, and not jeopardize the state. In this situation, communication between the vendor and state was critical to both parties.



RFP Structure

This section provides a description of the major portions of an RFP:

1. Introduction
2. Expected Cost of Contract
3. Scope of Work
4. Technical Design and Report
5. Expected Services and Products
6. Personnel Loadings
7. Budget
8. Quality Control and Scheduling

This section stresses the importance of structuring the RFP to articulate needs and priorities to assure a quality, cost-effective product. The proposal review portion of an RFP is discussed later in this document.

1. INTRODUCTION

Does the introduction clearly define the purpose or problem to be addressed by the RFP? Does it provide suitable detail on the programmatic context, background, and relevant legislation?

The introduction should contain several important pieces of information and provide the "flavor" of what you are trying to accomplish. First, provide a clear, concise statement of the testing program involved and the services and materials solicited. Be sure to specify whether you want to replicate a previous program or create something totally different. Clarity cannot be overemphasized. For example, if you use the term "edit," do you want someone to redesign a test for you or just to make minor wording and format changes? This summary statement of purpose at the beginning makes it easier for vendors, who must read many RFPs, to quickly decide whether to respond.

Second, be as complete and accurate as possible in describing exactly what you want and what constraints, decisions, and commitments a contractor will have to deal with. When some important decisions have not been made yet, be explicit about what the decisions entail, when they are likely to be made, and who will make them (e.g., the state legislature, state board of education, state purchasing agent, state testing director). You may refer to attached documents that provide important information about what is to be accomplished, such as a paper on a particular design approach that is desired or a description of the new core curriculum for the state.

Third, discuss the "big picture," the context into which the proposed project will fit. Describe the relationship of the new project to other current or planned state programs, but do not

bother the reader with description of unrelated programs. Information about related existing programs lets the bidders know that they do not have to address those problems in this project.

2. EXPECTED COST OF CONTRACT

Should or can the expected cost of the contract be made explicit?

Some RFP writers feel that not mentioning the expected cost of a project will result in lower cost bids and save the state some money. However, if vendors lack information about expected cost, they may propose approaches that are much too grand for the state's budget or just the reverse. If all proposals received are too expensive, the state is left in a difficult position. On the other hand, if a vendor underbids and wins, the state may be pressured into awarding the contract to someone who may not be prepared to do all that needs to be accomplished. As a result, the contract may have to be revised to accommodate additional expenses and the process may entail several months of lost time for the project. In addition, providing expected cost information helps vendors concentrate their efforts and resources on proposals for projects that they can ably handle.

In general, when expected costs are provided in the RFP, the state is more likely to obtain proposals that match its cost restrictions. Although the proposals may not match the state's desired technical quality, the state can focus its evaluation on the quality of the proposals in relation to what the expected budget can buy. The state can concentrate on getting the best product it can afford.

Specifying the probable cost of a project may reduce the number of proposals that state receives by weeding out the more expensive ones. Unfortunately, this may deprive you of seeing some good ideas and receiving feedback on your costing and scheduling, which may lead you to underestimate the effort actually needed to accomplish projects. Accurate projections of costs are especially difficult for new testing officers who must work with short timelines, low budgets, and high expectations. Feedback on costs and schedules can provide insight that helps these officers revise impractical estimates.

If your state law prohibits the provision of even ballpark figures, you may be able to provide relevant portions of a similar contract from a previous year, including the cost, which will probably be public information. This is particularly useful with large projects. Another possible strategy is to specify the expected cost in terms of the expected number of hours the project will need for completion.

If you are unsure how much of a flexible budget may be eventually allocated to a particular project, it is helpful to the vendor for you to be straightforward about the situation. Perhaps you can give high- and low-end figures. Vendors always have the option of bidding beneath the low end if they wish.

3. SCOPE OF WORK

Does the level of detail specified in the required scope of work match your understanding of task requirements?

When you know exactly what you want, say so. As simple as it sounds, this dictum is not universally followed. It is unlikely that vendors will be able to intuit expectations precisely. Furthermore, if a task is not mentioned in the RFP and thus is not a part of the contract, the state cannot compel the vendor to do it. Products, particularly, need to be specified as carefully as possible. Technical processes may be specified in detail if you are sure they are technically sound, but it is best to clearly state that bidders may suggest improved methods. This would allow vendors to use newer and better technical approaches with which you may not be familiar.

When you do not know what you want (which may happen in the early stages of a developmental project) it may be hard to be very precise. Loosely worded RFPs are usually intended to evoke vendors' creativity, but such RFPs often result in a group of proposals that differ so significantly that they are difficult to compare.

Vagueness, especially when occurring in the RFPs of a state without free communication between testing officers and vendor, can make the proposal writing very difficult for the vendors since they have little way of discovering what the state wants. Some good vendors may decide not to respond to such unclear RFPs, and the state may lose the advantage of the vendor's competition and good ideas.

When you are in the early developmental or conceptual stages of a project, you are better served by:

(a) a pre-RFP planning meeting with consultants and possibly with potential vendors, to help clarify what you want so that the RFP can be precise and detailed, or

(b) a planning or design RFP, which can help you decide how to proceed with writing an RFP for the actual test development and other required services.

4. TECHNICAL DESIGN AND REPORT

Does the RFP require that vendors completely specify and justify all major elements of their technical design? Does the RFP include specifications for a full technical report? If specific technical requirements are included in the RFP, are they technically sound?

There are several reasons why many state testing officers prefer to allow bidders to propose their own technical methods rather than requiring specific techniques. At the time the RFP is written, you may not know exactly what you want in terms of technical design and analyses or if what you want is technically sound and state-of-the-art. Allowing the bidders to propose their own technical suggestions may give you information about different options, thus helping you select the best methods to ensure the technical quality of your project.

In addition, if you provide very specific technical requirements in the RFP, vendors' proposals may need only parrot the RFP, making it difficult for you to judge the depth of their technical comprehension and expertise without seeking additional information. If you do allow bidders to propose their own technical suggestions, you should require that bidders completely specify the approach and rationale for all major elements of their design. This information will help you

compare approaches when reviewing the proposals and negotiate changes in proposed approaches with the vendor finally selected.

A disadvantage of allowing bidders to propose their own technical approaches (which may differ significantly) is that you may have a difficult time comparing value for cost.

Specifying in the RFP that the vendor will provide full technical reports lays the groundwork for later monitoring of the project. You may want to use technical "watchdogs" (experts) to review vendors' proposals or to review the vendor's technical work after the contract has been awarded. In the latter case, the expert functions as an outside consultant, hired by and working for you but paid for by the vendor. This strategy is particularly useful if advanced and/or complex statistical and technical procedures will be proposed by vendors. Technical assistance may be invaluable.

5. EXPECTED SERVICES AND PRODUCTS

Does the RFP address all expected services and products, and specify the quantity and quality of each that are expected? Does it specify which of them belong to whom?

As with other elements of the RFP, clarity and specificity are critical here. The cost of a project is often greatly affected by the number of tests and reports to be created, printed, delivered, and so forth. It is wise to address potential costs or savings of contract revisions. Sometimes vendors give "credit," which may be used in later phases of the same project. Obviously this is a poor arrangement if you will not be conducting business with this company in the future. In

addition, the credit offered by a vendor may not adequately represent the true cost of the omitted work, thereby depriving the state of full value. In some situations you may be able to trade certain tasks or products for others as the project progresses and priorities change, or the vendor may reduce the final billing on a project.

The RFP should also specify exactly who will constitute the group(s) to be tested and how special populations (e.g., Spanish speakers, visually impaired) are to be addressed. For example, should special forms or special administration procedures be developed? RFPs should also specify which products will belong to the state and which to the vendor, so possible disputes can be avoided.

6. PERSONNEL LOADINGS

Does the RFP require bidders to justify personnel loadings by task and relevant qualifications? Does the state have veto power over proposed personnel and/or changes in critical personnel?

Sometimes it is important to have more than just the top people specified for a project. Changes in other key staff may have a profound effect, particularly if these people have some special expertise or knowledge of the project or related programs. You will want to use a "key personnel" clause to protect the state from both intentional ("bait and switch" tactics) and unintentional changes in personnel.

You may want to require bidders to provide a list of their proposed staff who are already committed to concurrent projects and bids in order to see the spread of key personnel should the vendor win other outstanding bids. You could also

ask to be updated immediately prior to proposal review. It should be helpful to see the percent of key personnel's commitment to other projects on a monthly basis during the timeline of your project so that you can judge whether their availability will be adequate to your needs, especially at critical periods of your program. A task loading chart can help identify how serious the bidders are about various aspects of their proposal.

In fairness to vendors, each organization may be structured differently in regard to support resources, so the percent of time a key person in one organization requires to meet contract specifications may differ significantly from the percent of time required by another. Level of experience may also impact the percent of time required to do a job. Since the appropriate commitment of key personnel can greatly influence the success of a project, checking the track record of the bidders may be one of the best ways to ensure adequate allocation of key resources.

7. BUDGET

Does the RFP request a budget at the task level? Does it specify a payment schedule or request that bidders propose one?

It is very important to provide a standard format for all bidders to use in presenting their budget proposals so that you can adequately compare their bids, particularly if little time is available for this comparison. If you must or want to select the lowest bidder, it is imperative that you be able to judge who is truly the lowest. A standard budget format should apply to both the summary page, which is very useful to reviewers, and to the details of the budget. If details are to be compared, it is

important to have all bidders break the details down in the same way. If the project extends for more than one year, it is also helpful to have all bidders' budgets broken down by year at the same level of detail; the first year's budget should have the greatest detail.

A detailed budget is useful for a variety of reasons. It allows you to decide which services or products to omit if it is necessary to cut costs, or what credit to expect if part of a project is cancelled. Detail is particularly necessary if you allow variable costs in the proposals, but it can also be helpful with fixed costs. It allows you to check that all requested products and services appear in the bidder's budget and protects the state from problems that may result when something is inadvertently omitted from the RFP and/or the vendor's budget.

A caution: Excessive budgetary detail required by an RFP may drive away some potential bidders who do not find the effort worth their while. Be sure you can substantiate your need for all the figures you require in the proposals. The level of detail required ought to be in proportion to the size of the project, with greater detail for bigger projects.

8. QUALITY CONTROL AND SCHEDULING

How can quality control and scheduling requirements be assured? Should there be penalties for failing to complete scheduled work on time?

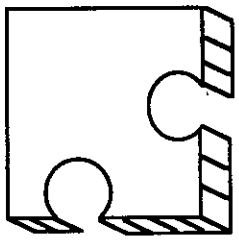
The RFP can spell out deadlines for completing interim job tasks and can request interim or progress reports and drafts of final reports to be reviewed by the state office prior

to the final version. It also may be helpful to specify the turnaround time for state review of documents.

Requirements such as progress reports protect the state in the "worst case scenario" and may be relaxed in practice as is appropriate. Interim reports may be weekly, biweekly, monthly, or other. They usually summarize the work done and pending, critical decisions to be made by the state, and information to be provided by given dates. In addition to serving these important managerial functions, such reports can serve as documentation.

Many states include a paragraph in the RFP about penalties for failing to complete scheduled work on time or failure to meet certain other terms of the contract, usually in terms of a certain amount of money per day. A ceiling for penalties should be stated for the protection of both parties. This gives the state a time frame beyond which they can attempt to salvage a project by seeking alternate sources of services funded by the penalties from the contracted vendor. It gives vendors the ability to assess in advance the extent of financial risk involved. If no ceiling is given, some vendors may choose not to bid.

States sometimes require performance bonds to protect themselves when dealing with small bidders about whom they know little. A few states have required bid bonds. Vendors, especially small ones, say they tend not to go to this expense unless very motivated.



The Review Process

This section discusses several aspects of the review process and the importance of pre-contract negotiations. It covers the following topics:

1. Criteria and Process
2. Weighting the Criteria
3. Reviewers
4. Oral Presentations
5. Pre-contract Negotiations

1. CRITERIA AND PROCESS

Are the selection criteria and review process clearly specified in the RFP?

Vendors need to know how their proposals will be judged. The more specific the criteria, the easier it is for the vendors to prepare their proposals and for the reviewers to pass judgment on them. Criteria are often stated in such general terms that it is difficult for a bidder to know just how the proposal will be judged.

States vary in the process they use to select a proposal. Many use a two-stage process in which they first rate the work plan and then look at the budget. A few do not even look at the budget if the work plan is not adequate. This approach avoids the problem of being tempted or forced to accept a

"bargain" bid for a basically inadequate proposal. However, other states argue that it is valuable to consider cost along with other criteria.

RFPs tend not to reveal the actual process by which reviewers come to a decision, such as whether they vote independently and then compute an average score for each proposal, use a consensus system, or use some other approach. Vendors, of course, would like as much information about the process as possible. However, in some states, testing directors cannot discuss areas controlled by general state bidding policies.

2. WEIGHTING THE CRITERIA

What weight should be accorded to the various components of the bid (e.g. technical merit, staff quality, corporate capability, and so forth)? How does the state balance technical quality and budget issues in making selection decisions?

Some states apportion a total of 100 points among the various criteria and then assign points to the proposals for each criterion. An advantage of this approach is that it indicates to the bidders the relative weight of the criteria. It is most helpful when the criteria themselves have been explicitly stated. For example, it is instructive to know that the state will give a certain number of points for proposals that provide a solution to a particular technical problem.

In states where the review process is strictly followed, it is important to state criteria carefully and to distinguish between "necessary" and merely "desired" attributes of a proposal. For example, a state once called for bidders to

propose a creative approach to a task and then was forced to eliminate vendors who proposed good but not creative approaches because the proposals were technically "incomplete." Now this state indicates the necessary basics in the criteria section of its RFP, and after selecting a vendor, the state may request the use of desired creative techniques, for which it allows extra time and money. If the vendor is unable to do this, the state will subcontract part of the project to another company.

Vendors particularly appreciate candid information about the relative importance of technical merit and cost so that they can develop their proposals accordingly. If cost is an overriding factor, be straightforward about it. There is no sense in gathering technically grand proposals that cannot possibly be funded. When cost is particularly important, the vendors who know they cannot compete against the lowest cost operations will probably not submit bids, and they would prefer to make this decision than waste their efforts. If, on the other hand, technical considerations are very important (and cost is just "normally important" as opposed to "critical"), be candid about it. Avoid the false economy of implementing a bare bones plan that fails to provide needed quality. An efficient RFP process is to the state's advantage.

3. REVIEWERS

Who can/should serve as reviewers (internal and external), and what qualifications should they possess?

It is useful for the reviewers to represent a variety of perspectives. For example, a technical expert may be familiar with the vendors' level of technical expertise, and a school

district representative has probably had to deal directly with the local consequences of good and bad work by various test scoring and reporting vendors. In some states the purchasing office has requirements regarding who can serve as a reviewer, what they may be paid, and whether/how you may be able to "train" them.

Unfortunately, it is possible that a reviewer may have a grudge against a particular bidder and it may not be apparent until it is too late. Your best protection is to avoid persons who are overly opinionated or narrow-minded and to search for a balanced panel of reviewers. In addition, in order to validate or refute reviewers' input, you can obtain references for key personnel from the directors of previous projects on which the personnel have worked.

Reviewers are seldom trained despite the importance of their task, they sometimes serve without pay, and they are not usually held accountable for their decisions. To ameliorate the situation you may be able to discuss with potential reviewers what you hope to accomplish, who you expect will bid, what problems are to be solved, and so forth, in order to "educate" the reviewers and to detect their biases ahead of time.

If possible, reviewers should have input into the development of the RFP. At the very least, they should be given a copy of the RFP, any previous proposals or planning documents, and any other information well in advance of reviewing the proposals.

Some states use an ongoing technical committee comprised of technical experts from school districts and universities around the state to review proposals and monitor the progress of programs over several years. This arrangement allows the committee members to feel ownership of the program and to remain involved; they are most likely to make

responsible decisions in guiding the program. In order to avoid any hint of bias in the review process, people who join the committee sign an agreement not to be on retainer to vendors whose proposals the committee may review.

4. ORAL PRESENTATIONS

Will oral presentations be possible or required? Will they occur before or after a proposal is selected? Will there be other opportunities for clarification of bids?

Oral presentations can benefit both vendors and states. They can allow the vendors to more fully explain their proposals and explore issues or priorities with the state. Orals can, in turn, provide the state with valuable details about the proposals and information about key personnel. If there is no clear winner among the proposals, oral presentations may help the reviewers select a contractor. It may also be useful to have the option to schedule an oral with the apparent winner of a contract before the final letting of the contract.

Since in many states the RFP and proposal become part of the contract, it is imperative to resolve any discrepancies between the two documents. The expense and effort involved if orals are mandatory may burden small, distant vendors, and some of them may choose not to bid, thus reducing the choices available to the state. This problem can be avoided by making the orals optional, allowing bidders the choice of oral or written clarification of their proposal. However, reviewers may strongly prefer to question bidders in person, particularly if the oral is to be with the final bidders or with an apparent winner.

It is important to recognize that a great deal of effort may be required ahead of time to prepare reviewers to be objective and to avoid being swayed by the slickness of some vendors' presentations.

5. PRE-CONTRACT NEGOTIATIONS

Are pre-contract negotiations referred to in the RFP as part of the contracting process? What factors should be negotiated?

Pre-contract negotiations and specifications of special conditions are invaluable tools for contracting agencies. After selecting a vendor to provide the required test services, and before or while the contract is drawn up, a critical period exists during which you should negotiate a number of important aspects of the project with the contractor. Although the RFP and proposal include explicit timelines, work plans, and so forth, reality may differ from good intentions. For example, the purchasing office may have taken longer than expected to accomplish its tasks, necessitating a compressed timeline if a given deadline is still to be met. You will want to clarify or negotiate the details of the *actual* timeline, plus the work plan, staff assignments, schedule of meetings, involvement of outside groups such as curriculum committees, number of objectives to be represented on the test, and so forth.

Some states use this pre-contract period for oral screening of the selected vendor before any award is made. This opportunity to examine needed changes provides a safety valve for both sides—allowing vendors to change their minds or back out at the last minute, thereby averting larger problems down the line. Referring to this procedure in the RFP informs bidders of what to expect.

Improving Large-Scale Assessment

PART II

SAMPLE RFP OUTLINE

for Large-Scale Assessment

Pamela E. Aschbacher

February, 1989

PART II: SAMPLE RFP OUTLINE FOR LARGE-SCALE ASSESSMENT

Table of Contents

PART II: PREFACE

Part II contains a sample RFP outline for large-scale assessment. The purpose of the outline is to provide a fairly comprehensive checklist of the types of information to be included in state RFPs. It combines features of RFPs for both test development and administration, based on a compilation of many different, successful RFPs from across the country. The sample outline is not meant to be prescriptive. Every RFP is unique and thus may need to include only some of the topics covered here or some new ones.

Work on the outline originally began in a task force of the National Council on Measurement in Education during Richard Jaeger's tenure as president of that organization. The purpose of the task force, chaired by Bob Heath, was to develop a model RFP that would be helpful to states and other agencies as they contracted for testing services. The group collected many RFPs from across the country and began to draft a compiled outline of the best RFPs.

In the fall of 1987, the NCME Model RFP Task Force merged with the MITEI Project Task Force at CRESST to become the joint NCME/CRESST Task Force on Large-Scale Assessment, chaired by Pamela Aschbacher. At that time, the Joint Task Force reviewed the original goal and agreed that an expanded sample outline would be more helpful than a "model" RFP. The group reviewed the outline and submitted many helpful suggestions for expansion, which were then edited into the present document. It is hoped that this outline will be a useful tool in developing more effective RFPs for large-scale assessment.

P. A.
E. B.

SAMPLE RFP OUTLINE FOR LARGE-SCALE ASSESSMENT

I. INTRODUCTORY INFORMATION
   A. PURPOSE AND INTENT
      1. Clear statement of purpose of test; rationale (e.g., legislation)
      2. Content areas to be covered
      3. Grade levels
      4. Approximate number of students to be tested per grade level
      5. Time of year tests are to be administered
      6. Special considerations (e.g., bilingual or handicapped students to be tested)
      7. Any tasks or subtasks to be bid separately
   B. KEY DATES
      1. Dates during bid process
         a. Bidders' conference
         b. Bidders' inquiries
         c. Bids due
         d. Contract awarded
      2. Dates during contract period
         a. Scheduled start date
         b. Completion date
   C. BIDDING INFORMATION
      1. Issuing office and address, contact person and phone number
      2. Number of copies due
      3. Bidders' conference
         a. Mandatory?
         b. Place and time
         c. Recorded?
         d. If, when, and how minutes will be available
      4. Questions and inquiries
         a. How to ask (e.g., in writing only?)

b. Whom to ask
c. Responses shared with all?

5. Revisions to RFP
   a. When issued
   b. Who will receive revision information
6. Level of effort
   a. Expected cost
   b. Fixed and variable costs
   c. Funding amount and schedule set by legislature
   d. Contract awarded in whole or in part
7. Bonding
   a. Performance bond required?
   b. Bid bond required?
8. Subcontracting
   a. Allowed?
   b. Subject to approval by state/district
   c. Information about subcontractor to be provided
      1) Company name, address, officers, contact person
      2) Organization support and experience
      3) References
   d. Who is responsible for which tasks
9. Particular requirements of state (e.g., Equal Employment Opportunity (EEO), percent minority staff, special consideration to in-state companies)

D. CONTRACT INFORMATION

1. Project monitoring
   a. Planning documents after contract is let
   b. Progress reports
   c. Project officers and assistants working for state department of education (DOE) and contractor
   d. Technical advisory committee (e.g., who, when meet, functions)

e. Other advisory or oversight committees
f. Schedule of reviews and approval of materials (e.g., who, when, length of review period)
g. Penalties and contact person for late work
h. Extension (e.g., possible length, how to notify contractor, how contractor must respond)

2. Prime contractor responsibilities
   a. Proposal, RFP contents, and minutes from bidders' conference become part of any contract awarded as result of RFP
   b. Can contractor assign or transfer responsibilities without state's/district's approval?
   c. Conditions under which contract may be terminated
   d. Period for which accounting records are to be kept and made available
   e. Effort required beyond scope of this RFP
      1) Hearings, meetings, etc.
      2) Conditions (when, who, how) to determine that a new contract is needed
      3) Costs (a part of contract or additional fee?)
3. Ownership of materials, data, documentation: what belongs to state/district and what to contractor
4. Invoicing
   a. When rendered to a state/district
   b. When due and payable by state/district

E. PROPOSAL FORMAT AND CONTENT

1. Definition of "non-responsive" proposals
2. Contents
   a. Technical proposal
   b. Organization support and experience
      1) Personnel qualifications and loading

2) Organizational capabilities: previous experience with projects of similar scope (give name of company of project officers)
3) External consultants
4) References

c. Cost proposal

3. Format
   a. Proposal required to use same organizational structure as RFP?
   b. Specifications for cost proposal
      1) Under separate cover?
      2) At the task level?
      3) Standard format

F. EVALUATION OF PROPOSALS
   1. Evaluation criteria
   2. Point values or other indication of weight/importance
   3. Open to creative approaches to particular problem?
   4. Oral presentations
      a. Mandatory/optional?
      b. How request/assign date and time

II. <u>BODY OF THE RFP</u>

A. BACKGROUND INFORMATION
   1. Relation of proposed assessment to related past, present, and future programs
   2. Salient features of or quotes from relevant legislation
   3. Important (e.g., legislated) dates

B. SCOPE OF WORK (Specify products and processes, let bidder recommend, or do both)
   1. Specification of assessment type
      a. Content area and grade levels to be assessed and when
      b. Test objectives (e.g., provided or to be developed and how)

c. Assessment strategies (e.g., census testing, matrix sampling, duplex design)
d. Criterion-referenced, norm-referenced assessment, or both
e. Speed or power assessment

2. Composition
   a. Item development (e.g., all original? number of items per objective)
   b. Item review and editing (e.g., who, where, when, cost)
   c. Bias control (e.g., statistical and/or subjective review; who, when, what)
   d. Response mode(s) (e.g., essay, multiple choice, performance)
   e. Relationship or role of state committees
   f. Timelines
3. Trial testing
   a. Pilot and field testing
      1) Purpose
      2) Contingent on review/approval
      3) Supporting administrative procedures (e.g., training sessions)
      4) Design (e.g., when, minimum number of responses per item, number of items per test form, minimum amount of test time per student, security)
      5) Who decides on sampling plan and selects schools (DOE or contractor)
   b. Contacts with schools
      1) Liaisons
      2) Who administers tests (DOE, contractor, Local Education Agency)
4. Developmental analyses: what, when, design (RFP may specify particular procedures or request that bidder describe proposed procedures, rationale, and types of statistics to be obtained)

a. Item analysis
b. Calibrations
c. Reliability of test forms
d. Validity
e. Demographic data desired
f. Procedure for setting critical scores (i.e., cut scores, standards)
g. Forms (number of equivalent or parallel)
h. Norming
i. Equating to other tests or forms (e.g., anchor form?)
j. Sampling of items
k. Scaling

5. Distribution of pretests and final form
   a. School-year timing
   b. Delivery and return (e.g., who, when, where, number, overage, whom to contact for shortages and problems)
   c. Packaging
   d. Security
6. Data collection
   a. Registration of examinees (if required)
   b. Test administration
   c. Training
   d. Security
   e. Quality control
7. Operational analyses
   a. Scoring (formulas or plans)
   b. Data processing
      1) Data cleanup
      2) Documentation
      3) Hardware
      4) Software
      5) Required turnaround
8. Deliverables
   a. Planning document (after contract let)

b. Reports (progress and final)
c. Tests
d. Manuals
    1) Test administration
    2) Interpretation
    3) Technical
e. Training materials
f. Computer tapes

9. Reporting
    a. Audiences
    b. Formats
    c. Publicity requirements
10. Cost proposal (note: RFPs may require that this be in the body of the proposal or in a separate document)
    a. Organization
        1) Budget at the task level?
        2) Summary
    b. Standard format

III. <u>LICENSING, COMPLIANCE, CERTIFICATION, AND AFFIRMATION STATEMENT</u>

Improving Large-Scale Assessment

PART III

TECHNICAL QUALITY ISSUES

to Consider in RFPs

Ronald K. Hambleton
H. D. Hoover
Richard M. Jaeger

February, 1989

PART III: TECHNICAL QUALITY ISSUES TO CONSIDER IN RFPS

Table of Contents

PART III: PREFACE

Part III of this document contains three papers on critical technical issues to consider when drafting requests for proposals (RFPs). The technical papers presented here are based on discussions held at the 1987 MITEI RFP Project meeting attended by several state testing directors, representatives of major testing services, and academic measurement and evaluation experts.

During the meeting, the group discussed equating and item bias at length, and was able to discuss content validity briefly. Unfortunately, there was insufficient time at the meeting to discuss other important topics, such as reliability, passing scores, and details of the test development process. Please note that this exclusion was a result of time constraints and in no way suggests that these issues are of lesser importance. In fact, the group expressed the wish to address some of these issues at a later date.

During the group discussions, members agreed that the standards of technical quality for tests should be explicitly addressed in both the body of the RFP and in the criteria for judging proposals. There was consensus that some states have required too little of vendors to assure the technical quality of the tests. Other states have sometimes required inappropriate practices, such as asking vendors for equating studies with expectations far beyond what measurement experts believe to be psychometrically sound practice.

The group disagreed, however, about which specific technical strategies had greatest merit within the areas of equating and item bias. Because of this legitimate and significant disagreement, and because of difficulties in anticipating specific data conditions, the group was unable to provide step-by-step directions to states or model RFP

language. Instead, the group felt it appropriate to encourage the development of RFPs that require vendors to identify decision rules that should be used at critical choice points and to be as specific as possible in stating and justifying their chosen technical approach.

In addition, members of the group felt it important for states to be aware of the experts' methodological disagreements before developing RFPs, evaluating proposals, and contracting for technical services.

After the meeting, the authors of the three papers presented here agreed to write up their views of the group's discussions, submit those drafts to the rest of the group for review, and revise their papers. Hence, each paper reflects both the group's consensus on major points as well as the individual author's own perspective on the issue.

The paper on equating was written by Professor Richard Jaeger of the University of North Carolina at Greensboro. It defines and discusses test equating and calibration, recommendations for the test equating specifications that should be provided in RFPs, and the evaluation of equating quality.

The paper on content validity was written by Ron Hambleton, Professor of Education and Psychology at the University of Massachusetts at Amherst. The paper includes a description of the types of content validity evidence typically needed, questions to ask at each stage of test development, information needed in an RFP, appropriate scheduling to maximize the usefulness of the evidence, and composition of review committees. We would like to thank reviewers Richard Jaeger, Bob Linn, and James Popham for their particularly valuable suggestions for this paper.

The paper on item bias was written by H.D. Hoover, Professor of Education and Statistics at the University of Iowa and co-author of the Iowa Test of Basic Skills. It incorporates discussions of the nature and control of item bias, fairness, specific recommendations for item bias issues relevant to RFPs, and a discussion of areas of disagreement with reviewers.

We offer special thanks to the authors of these papers for all the time and expertise they devoted to the task. We also appreciate the insightful reviews and helpful suggestions from the continuing members of our Task Force as well as several new members:

William Angoff
Stan Bernknopf
Sharon Johnson Lewis
Carol Robinson

P. A.
E. B.

ISSUES TO BE CONSIDERED IN THE EQUATING PORTIONS OF REQUESTS FOR PROPOSALS FOR LARGE-SCALE ASSESSMENT PROGRAMS

Richard M. Jaeger

University of North Carolina at Greensboro

Because of test security problems and the evolution of school curricula, large-scale assessment programs require the creation of multiple forms of tests. For a variety of reasons—such as ensuring that each examinee has an equal opportunity to evidence his or her achievement, or a desire to examine growth or other temporal trends in the average achievement of students in schools or school systems—it is essential that multiple forms of tests used in large-scale assessments be placed on the same score scale. The process used to place multiple test forms on the same scale (and thus make the forms interchangeable, useful for comparing the performances of examinees who are tested with different test forms, and useful for examining trends in average student achievement) is termed *test equating*.

Developments in measurement theory and advances in computer technology and statistical software over the past 20 years have made routine equating of multiple test forms far more feasible than was the case several decades ago. In addition, the development of mathematical models that provide specific descriptions of examinees' performances on test items has greatly increased the range of available test equating procedures. However, these models are based on strong assumptions and provide accurate and durable equating only if their assumptions are met.

Strictly speaking, tests that are to be equated must be psychometrically parallel. Frederic Lord (1980) has noted that two tests are parallel, and thus capable of being equated, only if it is a point of indifference to any examinee which test he or she completes. Although the score scales of any two measures can be made to appear the same (through a process called *calibration*), the process will not result in equating unless the measures are parallel. To illustrate this point, consider two contrived examples.

First, suppose you were to weigh two random samples of adult men. The first sample is weighed on a scale that measures in English units (pounds), and the second sample is weighed on a scale that measures in metric units (kilograms). Suppose also that the first scale had been adjusted so that it added one pound to every person's weight, whereas the second scale had been adjusted so that, on average, it showed correct weights. Weights produced by the two scales could easily be equated (placed on the same score scale). If the samples of men were large enough, the formula needed to convert weight on the scale that weighs in kilograms to the scale that weighs in pounds would be estimated correctly as follows:

Weight in Pounds = 1 + 2.2046(Weight in Kilograms).

The 1 appears in the formula because the scale that measures in pounds adds a pound to everyone's weight, and the 2.2046 appears in the formula because it is the number of pounds in one kilogram. Now suppose that you wanted to apply this equating formula to the weights of two samples of women, half of whom had been weighed on the English-unit scale and half of whom had been weighed on the metric-unit scale. The equating formula derived from the data on men's weights would produce perfectly comparable scores for the women, just as it did for the men, because the two measurement instruments (the scales) measure the same variable and are

thus parallel instruments. Only if measurement instruments (e.g., tests) are parallel, will the equating formula developed using one sample of examinees apply correctly to other samples or populations of examinees. Our second example illustrates the converse situation:

Suppose you had weighed all of the sampled men, using the scale that measures in pounds, and that you had then measured their heights in inches, using a tape measure. You could use the height and weight data for the men to develop a calibration formula that would convert the men's weights in pounds to the scale of their heights in inches. Any of several calibration methods could be used. The simplest approach would be to calculate the mean ($\underline{M}_W$) and the standard deviation ($\underline{S}_W$) of the men's weights and the mean ($\underline{M}_H$) and the standard deviation ($\underline{S}_H$) of their heights. These statistics would be used in the following conversion formula:

$$\text{Height} = (\underline{S}_H/\underline{S}_W)\ (\text{Weight} - \underline{M}_W) + \underline{M}_H.$$

This formula would put the weights of the men on the same scale as their heights, in the sense that, on the new scale, the men's weights and heights would have the same mean (average value) and the same standard deviation. Since the distribution of weights and heights of men follow a bell-shaped curve (are approximately normally distributed) in the adult population, creating score scales that had the same mean and standard deviation would make the score scales comparable at every score value.

If you followed this process, you would have calibrated the scale (measuring weight) and the tape measure (measuring height) for the sample of adult men—the numbers these measurement instruments produced when applied to the sample of men would be on the same score scale. However, you *would not have equated* the scale and the tape measure

because they measure different variables; that is, they are not parallel. To verify this conclusion, you would merely have to apply your calibration formula to the heights and weights of a sample of women. Since the relationship between height and weight is different for women than for men, the calibration formula for men would not produce converted heights for women that were anywhere near their actual heights. More to the point, the mean of the height values produced by using the men's conversion formula would not be the same as the women's actual mean height, and the standard deviation of height values produced by using the men's conversion formula would not be the same as the actual standard deviation of women's heights. Not only would the conversion formula for men result in converted scores (heights) that were wrong for most individual women, but the average converted score would be wrong as well. Although this example is contrived, and admittedly extreme, it applies directly to two tests that measure different psychological functions, and are therefore not parallel.* The scales of such tests can be made comparable for a single sample of examinees by creating a conversion formula, but the tests cannot be equated. The conversion formula will not produce trustworthy score conversions for other samples or populations of examinees when the tests are not parallel, regardless of the test equating method used.

**Parallel* is used here to mean test forms that measure, within acceptable limits, the same psychological function. The operational definition of parallelism, according to Angoff (1984) is: "Two tests may be considered parallel if, after conversion to the same scale, their means, standard deviations, and correlations with any and all outside criteria are equal." It is the last requirement that would be violated in the second contrived example (conversion of weights to heights) cited earlier.

Test Equating Specifications for RFPs

This section contains recommendations on the test equating specifications that should be provided in requests for proposals (RFPs). The recommendations are necessarily general because specifics depend on the nature of the test forms or tests to be equated, and the constraints that govern collection of data for equating.

Since the psychometric literature is replete with methods for equating tests (cf. Angoff, 1984; Petersen, Kolen, & Hoover, in press) and none has been demonstrated to be universally superior, RFPs should specify a particular equating procedure only if the issuing state strongly prefers that equating procedure. In the latter case, proposers should be permitted to specify use of an alternative equating procedure, provided the specification is supported by a thoroughly-developed rationale.

RFPs should include the following three sections pertaining to test equating: "Rationale," "Procedures," and "Evaluation," as described below.

Rationale for Test Equating

If prospective bidders are to respond appropriately and completely, they must be fully informed about the purposes of test equating in the context of the assessment program operated by the issuing agency. The RFP must contain a detailed narrative description of the tests to be equated and the state's objectives in requesting that tests be equated. Among several potential objectives, listed in order of increasing problems and difficulties, are the following:

1. equating psychometrically parallel, multiple forms of a test;

2. equating a slightly customized norm-referenced achievement test (a test that incorporates some new development of item content specifications or some new item formats, but with at least three-fourths of the customized test identical in content specifications, psychometric item specifications, and item formats, to the standard norm-referenced test) to a nationally normed standard form;

3. equating a moderately customized norm-referenced achievement test (a test that incorporates new development of item content specifications or new item formats, but with at least half the customized test identical in content specifications, psychometric item specifications, and item formats, to the standard norm-referenced test) to a nationally normed standard form;

4. equating an extensively customized norm-referenced achievement test (a test that incorporates substantial new development of item content specifications or substantial use of new item formats, with less than half the customized test identical in content specifications, psychometric item specifications, and item formats, to the standard norm-referenced test) to a nationally normed standard form;

5. equating a curriculum-tailored, criterion-referenced test to a nationally standardized norm-referenced test; and

6. placing multiple levels of a test intended for different grade levels or age levels of students on a continuous, longitudinally-interpretable scale.

Authors of RFPs should realize that the current state of measurement science does *not* support the use of test equating for purposes 2 through 6 listed above. As noted earlier, it is widely known that test equating is not robust when applied to (a) tests that differ substantially in content, (b) tests that differ substantially in difficulty or reliability, (c) tests that are targeted to groups that differ substantially in ability, and (d) tests that assess a multiplicity of constructs that are differentially sensitive to instruction. The greater the differences among tests on any of these factors, the weaker will be the generalization of equating results to populations that differ in composition from the equating sample. If tests differ substantially in what they measure, the result of using equating procedures will be calibration, rather than equating, as described in the hypothetical example considered earlier.

Although previous research has shown that pre-equating of test items (purposefully selecting test items for a new form that are similar in content, format, and difficulty to items in the old form that is to be replaced) is generally not sufficient to ensure equivalent test forms in operational use; every attempt should be made to construct test forms that are as nearly parallel in content distribution and psychometric properties as is possible. Careful attention to content parallelism and psychometric parallelism should be required in RFPs that call for the development of multiple forms of assessment instruments.

Equating Procedures

RFPs should require that proposals include detailed discussion of the procedures to be used in equating tests or test forms to achieve each purpose specified in the RFP. Among the procedures that should be discussed in bidders' proposals are the following:

1. the data-collection design to be used, including plans for sampling examinees and plans for the administration of tests or test forms to be equated;

2. the sizes and composition of samples of examinees to be used in the equating study, including specification of the sampling frames to be used, the sampling units to be used, and backup sampling to compensate for nonresponse; and

3. the analytic equating methods to be employed, including discussion of the use of anchor tests or items (if any), and the specific statistical procedures to be used in constructing a comparable score scale for all tests and forms to be equated.

The RFP should require that the proposal contain a detailed justification of the data-collection design, sampling procedures, and data-analytic methods proposed for each equating purpose, including reasons for selecting the proposed design and methods instead of viable alternatives.

Evaluation of the Test Equating

The RFP should require that the proposal contain a detailed discussion of the methods to be used to evaluate the quality of the equatings that result from the data collected and the analytic procedures employed. In particular, the proposal should describe methods that will be used to estimate the degree of random equating error overall, at the mean, and at various points on the score scale including values at or near any cut-off scores that the contracting state intends to use in classifying or selecting individuals on the basis of test scores. In situations where equating is to be applied to a sequence of

tests over a period of years, methods to be used to estimate the resulting degree of scale drift should be described and justified.

The RFP should also require that the proposal include a description of procedures the prospective contractor will use to obtain an independent validation of the equating, so as to verify its accuracy and the appropriateness of all procedures used to collect and analyze equating data.

References

Angoff, W.H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

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

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 221-262). New York: Macmillan.

ISSUES TO BE CONSIDERED IN THE CONTENT VALIDITY PORTIONS OF REQUESTS FOR PROPOSALS FOR LARGE-SCALE ASSESSMENT PROGRAMS

Ronald K. Hambleton

University of Massachusetts at Amherst

According to the AERA, APA, and NCME *Standards for Educational and Psychological Testing* (1985), content validity evidence requires reviewers to "assess the degree to which the sample of items, tasks, or questions on a test are representative of some defined domain of content" (p. 10). Expert judgment is the main mode of investigation of a test's content validity (Messick, 1989). In assessing *content validity*, test content is matched to the content specifications for the test.

In preparing content validity specifications for a Request for Proposal (RFP), the RFP writer has the choice of (a) asking bidders for a content validation plan, or (b) providing details of the types and nature of content validity evidence which are of interest. Four categories of content validity evidence are typically needed to support the uses of tests in large-scale assessments:

1. *Objective Representativeness*—Are the objectives that are selected for inclusion in the test representative of the objectives included in the domain of content of interest? For competency tests, normally the domain of content of interest is based upon a state curriculum or an agreed-upon set of state objectives. The objectives themselves are often reviewed for appropriateness by a committee. Appropriateness can be assessed by judging how well the set of selected objectives covers the most important parts of the state's objectives or

provides an adequate sampling of the full set of objectives. In the case of professional exams, the domain of content of interest may be based upon the results from job analyses or role delineation studies. Another possibility is that the content is based on a review of college curricula in required courses.

2. *Item Representativeness*—Are the items measuring each objective in the test representative of the domain of content defined by the objective? To address this category, well-developed objectives, such as those that highlight a model test item, content specifications, and distractor specifications (with multiple-choice items), are commonly used (e.g., see Popham, 1978). The set of test items can be judged for their representativeness by asking reviewers to comment on how well the set covers the full domain of items spanned by the item specifications for the objective.

3. *Item-Objective Congruence*—Is the item a valid indicator of proficiency of the objective to which it is matched? Does successful performance on the test item require the same cognitive processes as those specified in the objective the item was prepared to measure? Measurement specialists can be especially helpful here. Unlike (2), which focuses on the assessment of sets of test items, (3) refers to the evaluation of individual test items.

4. *Technical Adequacy of Items*—Do the items satisfy standard item writing principles? Are the chosen item formats appropriate to permit valid assessments of the objectives of interest? Measurement specialists are well-qualified to comment on the suitability of the item formats. In some cases, empirical evidence would be desirable.

It is common to address the four categories of evidence using rating forms. Four examples from Hambleton (1984) are provided in Appendices A, B, C, and D. Interested readers are

referred to Hambleton (1984) for more information about these categories of content validity evidence and approaches for addressing the categories.

In preparing the content validity section of an RFP, the point must be made with prospective bidders that when building a test, amassing content validity evidence should not be viewed as a one-shot activity carried out at the completion of the test development process. Rather, content validity evidence should be compiled throughout the test development process and used in a timely way to make adjustments to the items in the test and items that are selected. Content validity evidence should be collected and used to guide the test development process at several important places. Some important places and appropriate questions to ask at each place follow:

1. At the *item development stage*, are the items representative of the domains of content they were intended to measure? Is *each* item technically sound? Is there evidence of item-objective congruence? When the answer to one or more of the questions is *no*, revisions can be made to the test items, or, in some cases, they can be discarded.

2. At the *item tryout stages*, is there evidence of technical adequacy of items as reflected by the results from an item analysis? Comments from the field may also be useful.

3. At the *final test development stage*, are the topics, subtopics, or objectives that have been selected for inclusion in the test, representative of the domain of content of interest? If not, new content selections can be made. Similarly, item representativeness with respect to each objective should be assessed at this stage.

4. At the *final test development stage*, were content validity considerations used in test development? How? And what evidence is there concerning the content validity of the test? Documentation of content validity is handled at this stage.

At each stage in the test development process, content validity evidence can guide the item writing process (where are items needed to meet needs?), item-writing training, and item selection.

A few additional points concerning content validity studies follow:

1. *Representativeness* means assessing the more important or critical objectives, and reflecting the proportional size of the domains of content for objectives. In other words, for the representativeness criterion to be met in content validity studies, objectives which are more important or broader in scope than others need to be emphasized in test construction.

2. Judging item or objective representativeness may involve stratifying the domain of content prior to obtaining the reviewers' ratings. For example, in organizing a set of mathematics objectives, categories such as "computations," "measurement," "geometry," and "problem solving" could be useful for stratifying the objectives, prior to evaluating the representativeness of the set selected for inclusion in the test.

3. Content validity studies are technical in nature, but the evidence can also meet political agendas as well. Designers must therefore seek out not only groups who can comment on content validity concerns, but also groups who are apt to raise concerns about the test if they have not had the opportunity to

review and influence the choice of test content early in the test development process.

4. Minority representation on item review committees is particularly important in conducting meaningful content validity studies. Therefore the RFP should make this point.

5. On some occasions, the number of test items may be too large for judges to review in the time available to complete the work. (There is also a practical limit on the number of test items that judges are willing to review.) On such occasions, a sampling plan must be developed to ensure that each test item is reviewed by an acceptable number of judges. Obviously, more judges will be needed when the number of items to review is large.

6. In the early stages of the test development process, judges should be encouraged to offer editorial changes to test items when they see shortcomings. At the final stages, editorial changes may be less useful because the proposed changes would need to be reviewed, and time may not be available to carry out these reviews. Less than ideal items can be withheld from the test and reviewed again later for inclusion in a future form of the test.

7. The composition of review committees should be given considerable attention. Technical as well as political considerations must be addressed in the selection of reviewers for committees.

Possible details to request from prospective contractors in an RFP include proposed methods for selection and training of judges or reviewers, the number of judges to be used, the intended review process and sample rating forms, methods for resolving conflicts, intended data analyses, and approaches for

reporting and using content validity data. These details will be addressed again in the next section.

Information Needed in an RFP

A well-written RFP should address six parts of a content validity study:

1. Ask for the types of content validity information that bidders feel are needed and why. Alternately, the state may wish to tell prospective bidders the nature and/or scope of the content validity studies they want.

2. Ask for details on the group or groups of persons who will be involved in the item and objective review tasks, along with desired numbers, and how persons will be selected and by whom.

3. Ask for details on the nature and amount of training for reviewers.

4. Ask for examples of item rating forms and approaches for data analysis and reporting.

5. Ask for details on the timing of content validity studies (in relation to the stages of test development) and how the available data will be reported and used.

6. Ask for details on analysis of content validity data.

Of course, a prior question before writing the content validity phase of the RFP is for the state to review its own resources (available time and expertise) to determine its role in the content validity process. The state may vary its involvement from essentially none (except observing the

content validity meetings) to total involvement. State departments of education normally have the technical knowledge on staff to carry out content validity studies without assistance from contractors. Seldom, however, do the departments have sufficient numbers of staff and the time to direct the work themselves. Assuming sufficient resources, the main argument against total state involvement is the question of conflict of interest. Some might argue that a state department of education has too much at stake to identify a test as lacking in content validity—the state's judgment in selecting a competent contractor would be questioned, and relations with the contractor would become very difficult. On the other hand, the contractor may not be the best agency either. Contractors know the test best, but they have the most to gain from a positive review. It is hard to imagine a contractor who would design a study to show its test lacked content validity. An intermediate position might involve the formation of a neutral committee under the direction of (say) an independent consultant. Ben Shimberg, George Madaus, and others have called for the formation of an independent auditing agency that could conduct validation studies which would include content validity evidence in the scope of their work.

Also, it is important for state departments of education to ensure that a contractor schedules the collection of content validity evidence at a time in the test development process when changes to the test can still be made. Normally, this time would be (a) following the item writing phase, (b) following the pilot-testing, and (c) following the subsequent construction of the test but prior to printing the test.

To this point in the report, we have described the content validity evidence that is needed during the test development process. On some occasions, an "off-the-shelf" test may be proposed for use in a large-scale state assessment (e.g., selecting one of the major standardized achievement tests may

be of interest). Here the review task shifts to judging how well the test content matches the state's objectives for assessment and the intended curriculum and instruction. Again, bidders need to be instructed to provide complete details on their plan for reviewing test items and for making a final test selection.

Additional Research and Development Issues

At least four aspects of content validity studies require additional research:

1. Guidelines for helping to decide when a sufficient amount of content validity evidence to support the intended use of the test scores has been collected would be helpful (e.g., see Smith, 1985). The particular test use and the feasibility of collecting the criterion data are important considerations.

2. Guidelines for documenting (reporting) content validity evidence would be helpful.

3. More research on the actual procedures for carrying out the four types of analyses described above are needed. Content validity evidence is greatly valued, but the process of collecting the relevant data, unlike the standard-setting problem for competency tests, for example, appears to be understudied.

4. Extensions to the methods proposed in this report for collecting content validity evidence are needed to handle subjective item formats such as performance items (e.g., writing assessments).

The author is grateful to Richard Jaeger, Robert Linn, and Jim Popham for providing evaluative comments on an earlier draft of this report.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: APA.

Hambleton, R.K. (1984). Validating the test scores. In R. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199-230). Baltimore, MD: Johns Hopkins University Press.

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

Popham, W.J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Smith, I.L. (1985). Content validity study of the AASPB item bank. *Professional Practice of Psychology, 6*, 233-250.

Appendix A

An Example of a Judge's Item Rating Form

Item Content Review Form

Reviewer: _____ Date: _____ Content Area: _____

First read carefully through the lists of domain specifications and test items. Next, please indicate how well you feel each item reflects the domain specification it was written to measure. Judge a test item solely on the basis of the match between its content and the content defined by the domain specification that the test item was prepared to measure. Please use the five-point rating scale shown below.

Poor	Fair	Good	Very Good	Excellent
1	2	3	4	5

Circle the number corresponding to your rating beside the test item number.

Objective	Test	Item Rating					Comments
1	2	1	2	3	4	5	
	7	1	2	3	4	5	
	14	1	2	3	4	5	
2	1	1	2	3	4	5	
	3	1	2	3	4	5	
	8	1	2	3	4	5	
	13	1	2	3	4	5	
3	4	1	2	3	4	5	
	6	1	2	3	4	5	
	12	1	2	3	4	5	
4	5	1	2	3	4	5	
	9	1	2	3	4	5	
	10	1	2	3	4	5	
	11	1	2	3	4	5	

Appendix B

An Example of a Judge's Summary Sheet for the
Items/Objectives Matching Task

Item/Objectives Matching Task

Reviewer: _____ Date: _____ Content Area:_____

First read carefully through the lists of domain specifications and test items. Your task is to indicate whether or not you feel each test item is a measure of *one* of the domain specifications. It is, if you feel examinee performance on the test item would provide an indication of an examinee's level of performance in a pool of test items measuring the domain specification. Beside each objective, write in the test item numbers corresponding to the test items that you feel measure the objective. In some instances, you may feel that items do not measure any of the available domain specifications. Write these test item numbers in the space provided at the bottom of the rating form.

Objective	Matching test items
1	
2	
3	
4	
No Matches	

Appendix C

Instructions for Using the Multiple-Choice Item Review Form

1. Obtain a copy of the objective and the test items written to measure it.

2. Place the objective number, your name, and today's date in the space provided at the top of the Item Review Form.

3. Place the numbers corresponding to the test items you will evaluate in the spaces provided near the top of the Item Review Form. The numbers should be in ascending order as you read from left to right. (This must be done if the processing of your data along with the data from many other reviewers is to be done quickly and with a minimum number of errors.)

4. Read the objective statement carefully.

5. Read the first test item carefully and answer the first 15 questions. Mark "✔" for "yes"; mark "**X**" for "no"; and mark "**?**" if you are "unsure."

   The last question requires you to provide an overall evaluation of the test item as an indicator of the objective it was written to measure.

   There are five possible ratings:

   5 – Excellent
   4 – Very Good
   3 – Good
   2 – Fair
   1 – Poor

6. Write any comments or suggested wording changes on or beside the test item.

7. Repeat the rating task for each of the test items.

8. Staple your Item Review Form, objective, and copy of the test items together, and return to the coordinator.

Appendix D

An Example of a Technical Review Form for Items

Item Review Form
(Multiple Choice)

Objective No.: _____ Reviewer: _____ Date: _____

Test Item Characteristics (Mark "✔" for Yes; mark "X" for No; and mark "?" for Unsure)	Test Item Numbers					
1. Is the readability level of the test item stem and answer choices suitable for the examinees being tested?						
2. Does the item stem describe a single problem for an examinee?						
3. Is the item stem free of ambiguities and/or irrelevant material?						
4. Is the content of the test item matched closely to the goal statement, objective, or task?						
5. Are all negatives underlined?						
6. Do the item stem and answer choices follow standard rules of punctuation, capitalization, and grammar?						
7. Are the answer choices arranged logically (if such an arrangement exists)?						
8. Is there one correct or clearly best answer?						
9. Is the placement of the correct answer made on a random basis?						
10. Are the answer choices free of irrelevant material?						
11. Are numbers or letters used to label the answer choices?						
12. Is any material provided in another test item that will provide a clue to the correct answer?						
13. When pictorials, tables, or figures are used, are they printed clearly and labelled correctly?						
14. Can the test item be answered by simple logic or common sense?						
15. a. Have words that give verbal clues to the correct answer, such as "always," "may," "none," "never," "all," "sometimes," "usually," "generally," "typically," etc., been avoided?						
b. Have repetitious words or expressions been removed from the answer choices?						
c. Will the distractors be plausible and appealing to examinees who do not know the correct answer?						
d. Are the answer choices approximately the same length?						
e. Has the use of "all of the above" or "none of the above" as answer choices been avoided?						
f. Are four or five answer choices used?						
g. Have double negatives been avoided?						
h. Have "clang" associations with the stem been avoided for the correct answer?						
i. Have distractors that mean the same thing or are opposites been avoided?						
j. Are the answer choices for an item similar in type, concept, and focus so that they are as homogeneous as possible?						
k. Is the correct answer stated at the same level of detail as the other answer choices?						
16. Disregarding any technical flaws which may exist in the test item (addressed by the first 15 questions), how well do you think the content of the test item matches with some part of the content defined by the objective? (Remember the possible ratings: 1=poor, 2=fair, 3=good, 4=very good, 5=excellent)						

ISSUES TO BE CONSIDERED IN THE ITEM BIAS PORTIONS OF REQUESTS FOR PROPOSALS FOR LARGE-SCALE ASSESSMENT PROGRAMS

H.D. Hoover

Iowa Testing Programs

The University of Iowa

The question of test bias as it applies to various social and cultural groups is a multifaceted one that is somewhat different for achievement tests than it is for aptitude or ability tests. In the United States, similarities in schools, language and common culture transmitted through the mass media, and population mobility make common nationwide or statewide achievement testing a meaningful endeavor. At the same time, it is also clear that there are significant curriculum differences among schools, that language differences exist across regions and cultural groups, and that the common culture is supplemented by many rich and unique cultural experiences. Some of the implications of this diversity for large-scale statewide assessment programs are as follows:

1. Tests should focus primarily on the common experiences of all students.

2. Special efforts must be made to avoid content unfamiliar to the experience of special groups and to balance familiarity of content for the various major cultures of the state or country.

3. In interpreting test scores, emphasis should be placed on the individuality of pupils and the unique cultural circumstances that affect educational development.

4. Norms provided with the tests should adequately represent this cultural diversity.

Some methods that have been used by test publishers and other researchers to minimize cultural bias have included the following:

1. employing contributing test authors with diverse cultural backgrounds;

2. selecting materials that reflect the varied interests of pupils from a wide range of cultural backgrounds and experiences;

3. reviewing materials at all stages of preparation for unfairness or lack of relevance for diverse groups;

4. conducting item tryouts in culturally diverse groups, analyzing results for potential item bias, and using this information in item selection and revision;

5. conducting research on relationships between cultural background and such factors as academic aptitude, achievement, social acceptance, persistence, and extracurricular participation;

6. conducting research on educational and testing needs for different groups; and

7. conducting research on differential item functioning across groups.

A distinction should be made between the potential bias that is a characteristic of the measuring instrument per se and bias resulting from the process of fallible human beings making decisions based, at least in part, on test evidence. Bias in test instruments may be more or less equated with lack of relevance. A test or test item which more nearly meets the individual needs of one pupil rather than another is less relevant for the latter and might be said to be biased against him or her. If differences in interest and motivation are considered to be biasing factors, all tests, or all experiences, may be said to have a certain amount of bias. A certain reading passage or language item might be more interesting and motivating for a girl than for a boy, for someone who is sports-minded, for someone from an urban environment rather than a rural environment, or for someone who is interested in science rather than in literature. A test that requires a pupil to do creative thinking is thus biased against a pupil who is not accustomed to thinking creatively. Examples of these subtly biased situations are all much easier to find than items that favor one ethnic group over another. Differences in motivation, interests, and values are extremely variable in all subcultures. This variability may explain in part the low reliability exhibited by most statistical methods used to detect biased items (e.g., Hoover & Kolen, 1984).

Thus, in a sense, a given item or passage or even a whole test might be fairer for one pupil than another. If bias is defined in this way, it is difficult to conceive of a test that does not present some advantage for a given pupil or group of pupils. If all "bias" of this kind were to be removed, it would result in the elimination of all that is interesting, clever, novel, challenging, and creative. Such a test would be bland, uninteresting, and irrelevant for everyone. (The attempt by textbook publishers to protect themselves from similar allegations of unfairness or "bias" is considered by many to have been a major contributing factor in the "dumbing down"

of textbooks.) While it is obvious that situations likely to be unfamiliar to a vast majority of students should be avoided, it is probably more important that the totality of items in a test exhibit balance across as many dimensions (gender, region of country, race/ethnicity, etc.) as reasonably possible.

Another situation sometimes cited as a potential source of bias is one that results from asking a given pupil a question based on something the pupil has never had an opportunity to learn. This could be a situation in which the knowledge is relevant, possibly even critical, but the school or society has not provided the opportunity to obtain it. One might reasonably contend that a test containing items of this type is not fair to the pupil, or even that the test is not valid in that it is not measuring what has been taught. However, if the purpose of testing is to improve instruction, it is exactly in this situation that a test has the potential for greatest usefulness because its use should lead to the provision of such experiences.

The preceding discussion focused on differences among individual students for the sake of illustration only. Strictly speaking, bias as it relates to test development and test use is a characteristic present for rationally defined groups, not individuals. A comprehensive definition of bias representing this view covering all aspects of test development and use is illustrated by the following quote from Cole and Moss (1988):

> An inference from a test score is considered sufficiently valid when a variety of types of evidence support its plausibility and eliminate primary counterinferences. An inference is biased when it is not equally valid for different groups. Bias is present when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or

> implications for the remainder of the test takers. *Thus, bias is differential validity of a given interpretation of a test score for a definable, relevant subgroup of test takers.* (p. 205)

This definition implies that a test item, or test, might be biased in one use but not in another. For the purposes of this discussion, the following somewhat simpler definition should be sufficient: *An item or test is biased if examinees of equal ability from different groups exhibit differential performance.*

Responsibility for Fairness

The responsibility for ensuring that tests used in large-scale assessment programs are as free from bias as is reasonably possible is one shared by the issuers of the RFP (hereafter referred to as the "state" for simplicity's sake) and the bidder or vendor. The degree of responsibility is related to the use and ownership of the final test. As the ownership shifts to the state, so does the responsibility. Uses of a test not explicitly recommended by the publisher also shift responsibility to the state. The following examples should help clarify the nature of this shared responsibility:

1. *An RFP calling for the use of a nationally standardized achievement battery (or shelf test) in the fall of the year where the testing program's primary focus is on the improvement of instruction, rather than accountability*—In this case, the responsibility would lie nearly totally with the vendor, since this use is one explicitly recommended for the test by the publisher. The vendor should be able to furnish evidence pertaining to efforts used to ensure fairness, such as the judgmental review and field-testing of items described earlier. The state must judge from this evidence whether these efforts

were satisfactory. In such situations little if any cost should be added to the vendor's bid.

2. *An RFP calling for the use of a nationally standardized achievement battery to determine promotion from Grade 8 to Grade 9*—Such a use is likely to impact various racial/ethnic groups differentially. In this case the responsibility should probably be shared roughly equally between the state and the vendor. Since such a use is not one normally recommended by the publisher, evidence in addition to that described in the first example would be required. While the state might initially appear to have primary responsibility in this application, it should be kept in mind that the vendor is apt to benefit from the additional data obtained on the test battery pertaining to validity and bias. The sharing of cost, along with the responsibility for fairness, would seem reasonable in this context.

3. *An RFP requesting a criterion-referenced test that is to become the property of the state; a test tailored to its curriculum and intended for use in a high stakes decision, similar to that of the second example*—In this situation both the cost and the responsibility would lie predominantly with the state. The contracting of item review, sampling, and analysis procedures to a vendor would not abrogate this responsibility. The RFP should be quite explicit with respect to the methods used to ensure equity in such an application. In fact, a separate RFP dealing only with item or test bias might be preferable.

These examples indicate that in some situations the responsibility for fairness lies almost solely with the vendor and in others it lies more with the state. In those cases where the primary responsibility is the state's, it may still be reasonable for the state to contract this responsibility to the vendor. However, the procedures for the vendor to follow in these situations must be made explicit by the RFP.

If the test is used for high stakes decisions affecting individual students, the state assumes more responsibility. Such use necessitates careful attention to item bias issues.

Item Bias Specifications in RFPs

There are two commonly used ways of screening potentially biased items from intact tests or from pools of test items. With *judgmental* methods, experts evaluate the fairness to various groups of the item development process, the presentation format, and the content of potential items. With *analytical* methods, item data obtained from relevant subgroups and indices sensitive to differential performance by these subgroups are computed. Judgmental methods are especially helpful in dealing with perceived fairness issues such as balance and unintended stereotyping. While they may also be of some help in minimizing differences in item performance among groups, most studies comparing judgmental and analytical methods have found the two to be essentially uncorrelated. For this reason, the use of test scores of individuals in high stakes decision-making requires some attention to analytical methods.

Specific recommendations related to each of the two methods follow:

1. *Judgmental*—If a judgmental review of items is required, the RFP should document the process to be followed in item development to ensure fairness, provided the RFP requires "new" items. If shelf items are to be used, the RFP should ask for procedures used by the vendor in item development. Any additional procedures required for content or linguistic review should also be made explicit in the RFP. If judges representative of specific racial/ethnic groups are expected to be a part of this process, it should be stated.

However, it must be kept in mind that if the state expects shelf items from the vendor, highly restrictive specifications may keep most, if not all, potential bidders from responding. While it would be expected for judges representing specific groups to focus their evaluation on characteristics of items they are likely to be most qualified to judge (e.g., women assessing gender fairness), all judges should be informed that representation balance for an entire set of items, or a completed test, is critical. The chapter by Alpert, Gorth, and Allan (1988), "Bias Concerns in Test Development," is an excellent resource for instructing judges regarding balance and other concerns. Another excellent source is a book by Maggio (1987), *The Nonsexist Word Finder.*

2. *Analytic*—If empirical procedures requiring the use of "item bias indices" are included as a part of the RFP, a number of issues must be kept in mind. As is the case for equating procedures, a number of alternative item bias indices exist, but none has been shown to be universally superior. However, procedures utilizing only differences between groups in average percents correct have been shown in general to be inappropriate and should be avoided. Given adequate sample size, some of the procedures based on item response theory (IRT) appear promising. However, because of the uni-dimensionality assumption underlying IRT models, it is unresolved as to whether items identified by such methods as being "biased" might simply be indicating differences in dimensionality among groups (Linn & Harnisch, 1981). For example, in many states, a majority of the students of a given racial/ethnic group may be enrolled in a limited number of the school districts of that state. If curricula in these districts differ appreciably from those in the rest of the state, racial/ethnic differences in performance are nearly totally confounded with curriculum differences. Many people would still argue that tests or items measuring different things for different cultural groups are by definition biased. However, as

was pointed out earlier, if the primary purpose of testing is to improve instruction, this is exactly the situation in which a test is most useful.

Even though somewhat in disagreement with the definition of bias we adopted earlier, the preceding example on the confounding of group membership with instructional differences would indicate that differential performance does not necessarily imply item bias. In high stakes testing applications (e.g., promotion or certification), it is strongly recommended that data on possible curriculum and educational background differences be obtained. Such information is sometimes referred to as "opportunity-to-learn" data.

Another major consideration in the use of item bias indices is their demonstrated low degree of reliability and subsequent low predictive validity. This low reliability becomes especially apparent when curriculum differences are controlled (Hoover & Kolen, 1984). In general, it is recommended that item bias indices not be the sole criteria for decisions regarding test item fairness, but that they be used in conjunction with other relevant information, including judgmental review of items.

If data on racial/ethnic differences in performance are to be obtained as a part of the item development process, the RFP should state the following: (a) which racial/ethnic or linguistic groups are to be sampled; (b) which sampling design is to be used, and what stratification variables will be furnished by the state; (c) whether opportunity-to-learn data is to be gathered (strongly recommended for high stakes decisions); and (d) which item bias method is preferred by the state.

The preceding discussion of issues and recommendations are primarily directed toward how to deal with item bias prior to the operational use of a test. It is recommended that states

carefully analyze group differences in performance on intact tests and individual items after their first administration. This information should be used to improve tests for subsequent use. If it is expected that these analyses are to be performed by the vendor, it should be stated in the RFP.

The author would like to thank William Angoff, Sharon Johnson-Lewis, John Keene, Tom Kerins, Steve Koeffler, Wayne Neuberger, Ed Roeber, and Ramsey Selden for comments on an earlier draft. Hopefully the revision does justice to their excellent, but sometimes conflicting, suggestions.

References

Alpert, R.T., Gorth, W.P., & Allan, R.G. (1989). Bias concerns in test development. In R.T. Alpert, W.P. Gorth, & R.G. Allan (Eds.), *Assessing basic academic skills in higher education: The Texas approach* (pp. 177-228). Hillsdale, NJ: Lawrence Erlbaum.

Cole, N.S., & Moss, P.A. (1989). Bias in test use. In R. Linn (Ed.), *Educational measurement*, (3rd ed.) (pp. 201-219). New York: Macmillan.

Hoover, H.D., & Kolen, M.J. (1984). The reliability of six item bias indices. *Applied Psychological Measurement, 8*, 173-181.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Maggio, R. (1987). *The nonsexist word finder*. New York: Oryx Press.

Improving Large-Scale Assessment

Part IV

ISSUES IN MEANINGFUL REPORTING

July, 1990

PART IV: ISSUES IN MEANINGFUL REPORTING

Table of Contents

PART IV: PREFACE

It has been said that anyone can collect data; the real question is how to interpret it. Meaningful interpretation of results, in fact, is of critical importance both to those who compile assessment reports and to those who read and act on them. Part IV of this document contains two papers on issues in meaningful interpretation and reporting of large-scale assessment results. They are based on research projects conducted by CRESST and supported by the Office of Educational Research and Improvement.

The first paper by Bob Linn and his colleagues addresses the use of normative comparisons in reporting state and district results: How are normative comparisons used and misused? What does it mean to be "above the national average"? What can be done so that reports fairly represent achievement results?

The second paper, by Bob Linn and Ron Hambleton, focuses on the validity of interpretations and uses of customized tests and customized norms.

We thank the many state and district directors of testing and other staff who generously cooperated with CRESST projects to provide much of the data for these papers. We also thank the authors and reviewers for generously contributing their time, effort, and expertise to this notebook project.

COMMENTS ON THE USE OF NORMATIVE COMPARISONS IN REPORTING STATE OR DISTRICT ACHIEVEMENT TEST RESULTS*

Robert L. Linn

M. Elizabeth Graue

Nancy M. Sanders

University of Colorado-Boulder

The trustworthiness of state and district reports comparing the achievement of their students to the national norm was called into serious question in the fall of 1987 by the publication of a report by Dr. John J. Cannell entitled "Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above Average." Not only did Cannell claim that no state was reporting below-average test performance in the elementary grades on nationally normed tests, but he went on to conclude that "standardized, nationally normed achievement tests give children, parents, school systems, legislatures, and the press inflated and misleading reports on achievement levels" (p. 3). Cannell blamed test publishers and educators for practices that he claimed inflated the test results. He also charged that test publishers and educators are misleading people with their reports of results.

Is it really the case that all states and nearly all districts claim that their students are performing above the national average on achievement tests? If so, does this imply that the

---

* A version of the Linn, Graue, & Sanders (1990a) study on which this report is based will be published in *Educational Measurement: Issues & Practices*, (1990b).

results are "inflated and misleading"? How should such a finding be interpreted?

The purpose of the present paper is to discuss some of the findings and conclusions of a recent study by Linn, Graue & Sanders (1990a) designed to address the questions raised by Cannell's claims and the possible influence of the changing meaning of norms. The first section below provides some background on the nature and use of normative comparisons. The second section summarizes the findings, discusses the context and several explanations for the results, and discusses why annual norms are not an appropriate or effective solution. The final section discusses conclusions and recommendations.

How Are Normative Comparisons Used and Misused?

Standardized achievement tests have long been used by schools to report student achievement to parents, policy makers, and the general public. In recent years, however, the increased emphasis on the use of test results for purposes of accountability has made questions of test quality and the trustworthiness of interpretations of major concern to educators and policy makers.

A major, albeit not the only or necessarily the best, way of providing the various audiences a means of interpreting test scores is to compare achievement test scores for a school building, a district, or a state to national norms. Slightly over half of the states and a substantial majority of the school districts rely on off-the-shelf, standardized achievement tests, for which normative comparisons provide a primary basis of interpretation. These comparisons take on a wide variety of forms, including the average grade equivalent score, the average normal curve equivalent score, the median percentile rank or percentile rank of the mean, the proportion of students

scoring above the "national average," or more precisely, the national median, and the proportions of students with "below average, average, or above average" scores where the three categories correspond to stanines 1 thru 3, 4 thru 6, and 7 thru 9, respectively. In each of these examples, national norms provide the primary basis of comparison.

Norms, of course, are not the only basis for interpreting test scores. Some states and districts rely on criterion-referenced interpretations of either publisher- or locally-developed tests. Such interpretations are referenced to the domains of knowledge which students may possess, and they often gain meaning in comparison to past performance. For example, trends in the proportion of students passing a minimum-competency test, the proportion of students mastering specific objectives, or the average number of objectives mastered provide a means of comparing the current year's achievement with a benchmark. Trends may also be important in the interpretation of norm-referenced results, but the national norm still provides the major frame of reference for expressing the scores. Even states with locally-developed or customized assessment programs sometimes also use comparisons to national norms, obtained thorough special equating studies or item response theory links, to aid in the interpretation of their achievement test results.

The pros and cons of normative comparisons have been discussed on many occasions. Discussions of appropriate and inappropriate normative interpretations are provided, for example, by Angoff (1971), Petersen, Kolen, and Hoover (1989), and in several introductory texts on educational and psychological measurement. Good discussions of appropriate and inappropriate uses and interpretations of norms may also be found in the technical manuals and interpretive guides provided by the publishers of the major standardized achievement tests.

Despite these discussions, normative interpretations continue to be misused and misinterpreted. The distinction that Angoff (1971) and others have made between the statistical meaning of "normative" which refers to "performance as it exists" and the use of the term to refer to "standards or goals of performance" (p. 533) is too often overlooked. The fact that norms for school average or district averages differ markedly from norms for individual students is too often ignored or given insufficient emphasis in interpretation. Because a school average is based on a range of student scores it necessarily falls somewhere in between the score of the highest scoring individual student and that of the lowest scoring student. Consequently, the distribution of school average scores is less variable than the distribution of individual scores. The average achievement score that corresponds to the 70th percentile using school building norms, for example, may correspond to only the 60th percentile using norms for individual students.

Further, it is widely believed that some tests have "easier" norms than others. If the norms of test A are easier or less stringent than those of test B, then a given level of achievement would be expected to appear better (e.g., result in a higher percentile rank or a larger proportion of students scoring above the national average) with test A than with test B. Note that the difficulty of norms is different than the intrinsic difficulty of test items. A test that asked easy questions could have hard norms because the norming sample was unusually able in the content area of the test. Conversely, a second test that asked relatively more difficult questions could have easier norms because the norming sample for the second test included a disproportionate number of low achieving students. The relative difficulty of norms for a particular school, school district, or state may also depend on the degree to which the test content matches the curriculum at the building or classroom levels.

The meaning of norms depends fundamentally on the definition of the reference population, and secondarily on the adequacy of sampling, the level of participation, and the motivation of the students in the norming sample, among other considerations. The year in which the norms were obtained is one of the important properties that define the reference population, and it is clearly the case that norms become dated. If achievement is improving nationally, then the use of old norms will make a district or state appear to be doing better relative to the nation than would the use of current norms which provide a higher standard of comparison.

Should Normative Comparisons be Trusted?

Although the above concerns about the use of norms are hardly new, questions about the meaning and trustworthiness of normative comparisons that states and districts are using to communicate test results to policy makers and the public have recently taken on increased importance. The increased importance is due, in part, to escalation in the stakes involved in testing. Concerns about normative comparisons were also exacerbated by the sharp criticism and dramatic language of Cannell's report (1987).

However, Cannell was not the first to notice that states were reporting results that were above the national norm in greater numbers than would be expected based on past experience or common-sense notions of the likely relative standing of particular states. In 1984, the Southern Regional Education Board (SREB) reported that 9 of 11 SREB states with norm-referenced test results for elementary grades were at or above the national average (SREB, 1984). Two years later, "[i]n June, 1986, SREB first described this situation in which student achievement in nearly all states was reported to be at or above the national averages as the 'Lake Wobegon effect'—descriptive

of Garrison Keillor's mythical town where all children are above average" (Korcheck, 1988, p. 3). However, it was the Cannell report that placed the issue in the national limelight and the focus of considerable discussion at national meetings and in professional journals concerned with issues of educational achievement and measurement.

Reviewers of the Cannell report (e.g., Drahozal & Frisbie, 1988; Koretz, 1988; Lenke & Keene, 1988; Phillips & Finn, 1988; Williams, 1988) identified several shortcomings of the Cannell study and interpretations. The failure to distinguish between group and individual student norms in interpretations, aggregation bias that results when the percent of *districts* with average scores above the national median is used to make inferences about the percent of *students* with scores above the national median, and the treatment of the percent of students at the 4th stanine or above as if it were an indicator of the percent of students above the national average are among the misleading analyses and interpretations that were identified.

Despite these and other limitations, some reviewers concluded that Cannell's major findings are still probably correct. Stonehill (1988), for example, states simply that "Cannell's evidence is compelling" (p. 23). Others were more circumspect. Koretz (1988), for example, noted that "Dr. Cannell's errors are to some extent beside the point...for they are not sufficient to call into question his basic conclusion" (p. 11) and Phillips and Finn (1988) stated that in the absence of "evidence to the contrary" they generally concurred with "the central finding of Dr. Cannell's report" (p. 10).

Linn and several colleagues conducted a study to collect data not only about the norm-referenced test scores that are reported by states and districts, but also on a variety of related issues, including the way in which test results are used (e.g., public reporting, grade retention, school incentives), when and

why the uses were initiated, how and when the tests were adopted, and policies regarding test administration, test security and the preparation of students for taking tests. The Linn, Graue and Sanders report (1990a), highlighted here, is focused on the test results and the possible influence of changes in the stringency of norms over time. Other aspects of the project data are addressed elsewhere (e.g., Baker, 1989; Burstein, 1989; Shepard, 1989).

Linn, Graue, and Sanders conducted a review of published reports, a mail survey, and telephone interviews with directors of testing in the 50 states and a stratified random sample of school districts throughout the nation. Their study focused on testing and reporting practices over a three-year period, 1986-1988. The results of their study are discussed below, followed by explanations, implications, and conclusions.

Are All States and Most Districts Really "Above Average"?

The results of the Linn, Graue, and Sanders study (1990a) provide support for Cannell's general finding, but their analyses lead to conclusions that are different, and certainly less sensational, than the ones Cannell reached.

Results Above the "National Average"

The study results suggest that for the elementary grades almost all states and a majority of school districts *are* reporting norm-referenced achievement test results that are above the national norm. Weighted estimates from the district sample suggest that at least 57% of the students in grades 1 through 6 are obtaining scores above the national median on norm-referenced reading tests. The corresponding figure for

mathematics is 62%. The comparable figures for grades 7 through 12 are lower, but still somewhat greater than 50%. As can be seen from Figure 1, the state results are quite consistent with the district estimates.

The Context

It is important to put the "above average" findings in context. While the percentages displayed in Figure 1 are generally above the naive expectation of 50%, many individual students are receiving scores that are "below average" even in districts or states that are reporting substantially more than 50% of their students are "scoring above the national average." For example, when a district reports that 57% of its students obtained reading scores that are at or above the national median, the other 43% of the students obviously scored below the median. It should be emphasized that although most districts report results that are "above the national average," there are still many districts throughout the nation that are reporting results that are below average. For example, one out of ten districts in our sample reported that only about a third of its students at a given grade level scored above the national median in reading.

Furthermore, the percent of districts that have more than half of their students scoring above the national median should not be interpreted as a direct indication of the percent of students across districts who are scoring above the median. It would be possible for a substantial majority of districts to have more than half their students above the median while less than half of all students across districts were above the median. For example, there may be many small districts whose average student performance is above the median balanced by a few very large districts with below-median performance by a great many students.

What Factors May Account for "Above Average" Scores?

Cannell concluded that norm-referenced achievement tests are producing inflated reports from states and districts on the achievement of their students. But the finding that more than half the students are scoring above the national median that was obtained when the norms were established does not necessarily imply that the results are inflated. There are many factors that may lead to the general finding, and three likely ones are discussed below.

Increases In Achievement and the Use of Old Norms

It seems clear that the use of "old" norms is one of the major factors that contributes to the abundance of "above average" scores. There is ample evidence that scores on norm-referenced achievement tests given in the elementary grades have increased substantially for the nation as a whole during the past decade (Congressional Budget Office, 1986, 1987; CTB/McGraw-Hill, 1987; Hieronymus & Hoover, 1986; Wiser & Lenke, 1987). This contrasts with national trends between the mid 1960s and mid 1970s when scores on norm-referenced tests decreased about as much as they have increased during the past decade.

Because of the national increases in test performance, there is a strong tendency for more recent publisher norms to be more stringent than older norms. Consequently, a state or district where the average student scores at the *current* national average will be *accurately reported* to be above the national average defined by norms that are several years old. It appears that a substantial fraction of the "Lake Wobegon" phenomenon may be attributable to the use of old norms. It

should be noted that the use of "old" norms is not purposeful on the part of school districts or states; they generally use the most recent norms available. Since standardized tests are usually normed every seven years, the most recent norms available will, on average, be three or four years old in most school years.

Exclusion of Students From Testing

A second factor that appears to be contributing to the abundance of high test scores is a tendency to exclude some students from test administrations. Districts and states may define and exclude from testing certain handicapped, special education, and other students according to rules or criteria that differ from those used in the norming process (Phillips & Finn, 1988). Exclusion of students from testing, even for educationally sound reasons (e.g., students with limited English proficiency or those with particular handicapping conditions), can result in a majority of the test takers being above the national average even when a majority of all the students in the district or state are *not* performing above that average.

Familiarity With the Test Form

The reuse of the same test form year after year is a third factor that appears to contribute to the apparently high test scores. Wiser and Lenke (1987) compared test scores during norming studies of students in districts that had been using a given standardized test ("users") with those in districts that had not used that test ("non-users") prior to the study. They found that in the elementary grades, test users performed as well or better than non-users across all subject areas, supporting the contention that part of the apparent growth in achievement based on norm-referenced tests may be due to

increased familiarity with a particular test form. However, for grades 7-12, the results were more mixed, with non-users performing better than users in some subject areas. To the extent that new norming efforts rely more and more on users (who are generally more willing to participate in norming studies than non-users, and who may be better prepared for the test because their curricula tend to be more closely aligned with the test), norms may be expected to increase in difficulty.

Another way to examine the effects of test familiarity on achievement is to compare performance on the National Assessment of Education Progress (NAEP) with norm-referenced achievement tests. Since NAEP items are more secure than those on norm-referenced tests and they are administered under conditions that involve much lower stakes than is often the case for norm-referenced tests, NAEP should be clearer window on real performance gains. In fact, NAEP performance has increased less than performance on other tests, suggesting that part of the large increase in scores on norm-referenced tests during the past decade is due to some combination of familiarity with specific tests that are reused year after year and the high stakes that are associated with the test scores.

Are Annual Norms An Appropriate and Effective Solution?

Concerns about dated norms have led to suggestions that publishers should produce current annual norms so that norm-referenced comparisons could be made to a current standard (e.g., Cannell, 1988; Phillips & Finn, 1988). Publishers are now attempting to do this by obtaining weighted estimates of national results from user data. As Shepard (1989) has pointed out, however, annual norms based on user data potentially have several serious defects. If users differ from nonusers in

ways other than those reflected by the demographic variables used for weighting, then user-based annual norms could provide biased estimates. In particular, if test familiarity leads to higher test performance, a state or district that changes publishers and administers a several-years-old test form for the first time would be at a disadvantage when compared to user norms. As a result, the user-based annual norms may be worse than dated norms where there is at least an understood frame of reference.

Furthermore, frequently updated norms represent a moving target "where educational gains or losses would be masked by the relative nature of the information" (Lenke & Keene, 1988, p. 18). The use of the same norms over a period of years enables the test user to demonstrate improvement relative to a constant reference group.

The alternative of conducting special national norming studies every year, or even every other year, is not a realistic or desirable possibility. Norming is not only expensive, but the quality of the results is very dependent on voluntary participation of schools. Current participation rates in norming studies conducted roughly every six or seven years by a publisher are already far lower than would be desired. More frequent attempts to norm tests would surely lower the participation rates still further and thereby degrade the quality of the norms.

Finally, although more recent norms provide a more stringent standard of comparison when scores are going up as they have been during the last decade, they would provide a less stringent standard during periods of decline in scores such as that experienced between the mid 1960s and the mid 1970s (Koretz, 1987). Thus, we do not believe that the use of annual norms is an appropriate or effective way to deal with problems caused by dated norms.

Has Student Achievement Really Improved?

There is ample evidence that scores on norm-referenced tests have been going up in grades 1 through 8 in recent years. But the more important question is: Has student *achievement* improved in recent years? Unfortunately, the answer to the latter question is equivocal.

Achievement test scores are of interest to the degree that they enable valid inferences to be made about broader achievement domains. But little attention has been given to the issue of the degree to which valid generalizations about broad achievement domains can be made from state or district test results.

Comparisons of the changes in norms of standardized tests with estimates of changes in achievement based on NAEP results suggest that test norms may be changing more rapidly than is student achievement as measured by NAEP. The Wiser and Lenke (1987) findings that apparent increases are generally smaller for non-users than for users of a given test series suggest that part of the apparent growth in achievement based on norm-referenced test results may be due to increased familiarity with a particular form of a test. Only part of the apparent gain can be explained in this way, however.

The differences between the gains in performance indicated by NAEP and by norm-referenced tests, and between Wiser and Lenke's total norming sample and their non-users, at the very least, suggest that caution is needed in interpreting gains in norm-referenced test scores as reflections of the amount of improvement that has taken place in achievement, more broadly defined. More direct assessments of the degree of generalizability of results to other tests and to other indicators of student achievement are greatly needed, however.

Hoover's (1989) finding that only about 6%, rather than the expected 10%, of students scored below the 10th percentile in the first year of operational administration of forms G and H of the ITBS suggests that a larger fraction of less able students are excluded from operational test administrations than from the norming studies. This suggests that greater emphasis in reporting needs to be given to the lower end of the score distribution and to the students who are excluded from testing when results are reported by states or districts. It may be quite appropriate, indeed desirable, to exclude students with limited English proficiency or students receiving particular types of special education services from a norm-referenced test administration. Such students should not be ignored, however, when district or state achievement results are reported. At a minimum, the number of such students and the reasons for exclusion from testing should be reported.

The practice of using a single form of a test year after year poses a logical threat to making inferences about the larger domain of achievement. Scores may be raised by focusing narrowly on the test objectives without improving achievement across the broader domain that the test objectives are intended to represent. Worse still, practice on nearly identical or even the actual items that appear on a test may be given. But, as Dyer aptly noted some years ago, "if you use the test exercises as an instrument of teaching you destroy the usefulness of the test as an instrument for measuring the effects of teaching" (1973, p. 89).

Although the desire for accountability is reasonable, it is the opinion of the authors that accountability pressures place too great an emphasis on test scores. It is unlikely that any single test, no matter how well constructed, normed, and validated, can withstand the pressures to serve as both an instrument of instruction and an instrument for measuring the effects of instruction. Making valid inferences about broad

achievement domains from test scores has always been a challenging and difficult undertaking, but is made all the harder by current demands for accountability and the use of standardized test results as primary indicators of accountability.

What Recommendations Can Be Offered?

To recapitulate, results of this study point to a number of ways in which the reporting of test results could be changed to lessen the exaggerated view of performance they can otherwise offer.

1. ***Emphasize the year in which the norms were obtained and explain the implications of using norms that are several years old (in an era when the stringency of norms has been on the rise).***

2. ***Identify clearly the number of students excluded from testing, the proportion of the student population this represents, the definitions of exclusions.***

3. ***Change forms of the test from year to year.***

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.

Baker, E.L. (1989, March). *What's the use? Standardized tests and educational policy.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Burstein, L. (1989). *Looking behind the "average": How are states reporting test results?* (Report to OERI, Grant No. G-86-0003). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends for Education.

Cannell, J.J. (1988a). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice, 7*(2), 5-9.

Cannell, J.J. (1988b). The Lake Wobegon effect revisited. *Educational Measurement: Issues and Practice, 7*(4), 12-15.

Congressional Budget Office. (1986). *Trends in educational achievement.* Washington, DC: Author.

Congressional Budget Office. (1987). *Educational achievement: Explanations and implications of recent trends.* Washington, DC: Author.

CTB/McGraw-Hill. (1987). *Technical report: California Achievement Tests, Forms E and F, Levels 10-20.* Monterey, CA: Author.

Drahozal, E.C., & Frisbie, D.A. (1988). Riverside comments on the Friends of Education report. *Educational Measurement: Issues and Practice, 7*(2), 12-16.

Dyer, H.S. (1973). *Recycling the problems in testing. Proceedings of the 1972 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Hoover, H.D. (1989, June). *Reactions to "Lake Wobegon" one year later: Results from a replication of the Cannell study.* Symposium presented at the 1989 Assessment of the Education Commission of the States, Boulder, CO.

Hieronymus, A.N., & Hoover, H.D. (1986). *Iowa Tests of Basic Skills, Forms G/H: Manual for school administrators, Levels 5-14.* Chicago, IL: The Riverside Publishing Company.

Korcheck, S.A. (1988). *Measuring student learning: Statewide student assessment programs in the SREB states.* Atlanta: Southern Regional Education Board.

Koretz, D. (1987). *Educational achievement: Explanations and implications of recent trends.* Washington, DC: Congressional Budget Office.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15, 46-52.

Lenke, J.M., & Keene, J.M. (1988). A response to John J. Cannell. *Educational Measurement: Issues and Practice, 7*(2), 16-18.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990a). *Comparing state and district test results to national norms: Interpretations of scoring 'above the national average'* (CSE Technical Report 308). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990b). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5-14.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 221-262). New York: Macmillan.

Phillips, G.W., & Finn, C.E., Jr. (1988). The Lake Wobegon effect: A skeleton in the testing closet? *Educational Measurement: Issues and Practice, 7*(2), 10-12.

Shepard, L. (1989). *Inflated test score gains: Is it old norms or teaching the test?* (Report to OERI, Grant No. G-86-0003). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

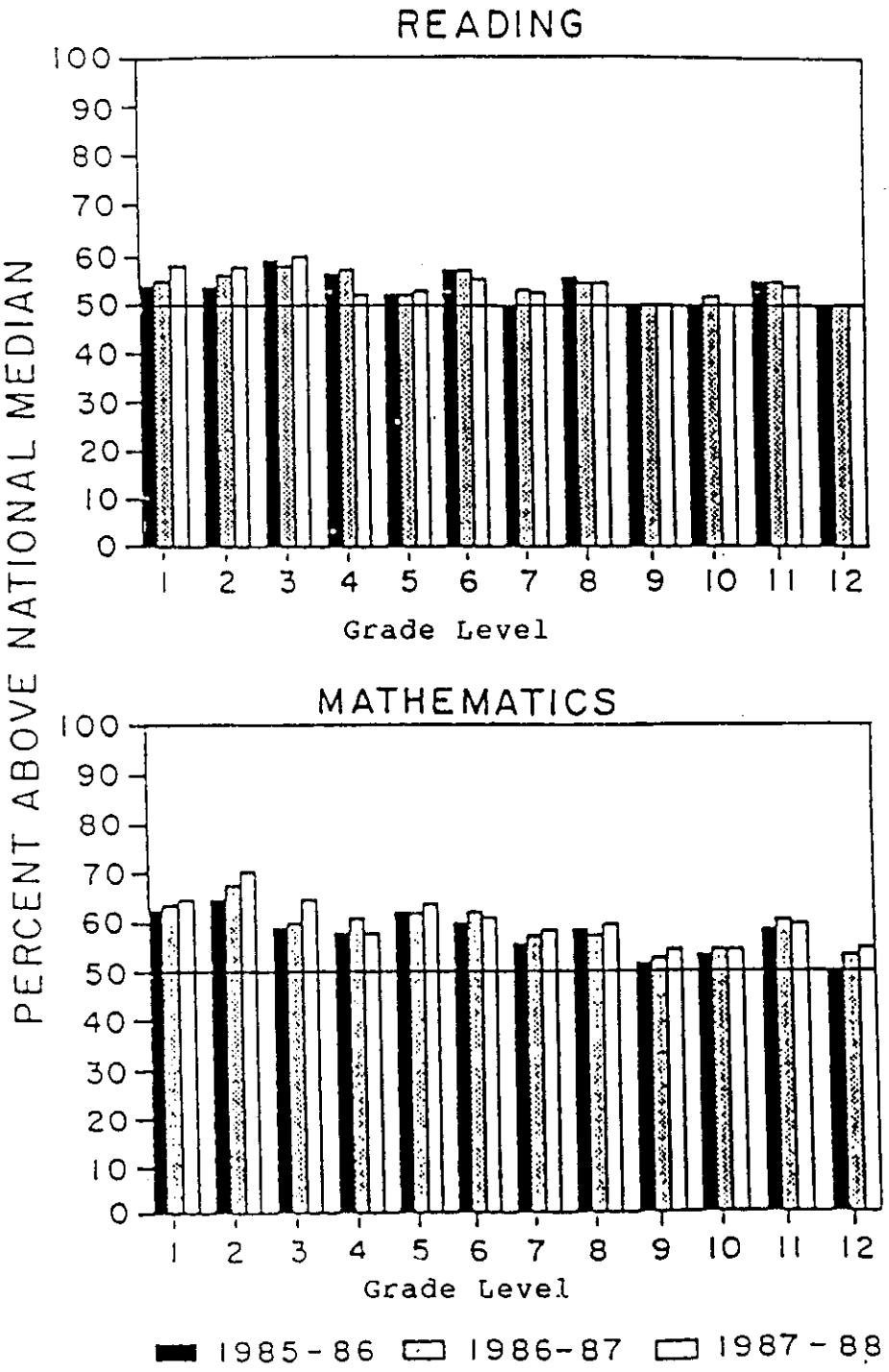
Southern Regional Education Board. (1984). *Measuring educational progress in the South: Student assessment.* Atlanta: Author.

Stonehill, R.M. (1988). Norm-referenced test gains may be real: A response to John Jacob Cannell. *Educational Measurement: Issues and Practice, 7*(2), 23-24.

Williams, P.L. (1988). The time-bound nature of norms: Understandings and misunderstandings. *Educational Measurement: Issues and Practice, 7*(2), 18-21.

Wiser, B., & Lenke, J.M. (1987, April) *The stability of achievement test norms over time.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Figure 1
Percentage of Students Scoring Above National Median
Based on States Reporting (Weighted by Number of Students)



CUSTOMIZED TESTS AND CUSTOMIZED NORMS

Robert L. Linn
University of Colorado-Boulder

Ronald K. Hambleton
University of Massachusetts

Accountability has been a prominent feature of the educational reforms introduced by states and districts during the past decade. Many new testing programs were introduced in the 1980s as part of the accountability movement and existing programs were expanded and made increasingly salient (Pipho, 1985). Tests were not only expected to monitor the effects of reforms, but, in many cases, to be the major mechanism for accomplishing desired changes (Linn, 1987; Madaus, 1985).

Expectations for tests were and continue to be manifold. For example, test results are expected to set more rigorous standards for students, to focus the efforts of teachers, to raise standards for teachers, to provide a means of judging strengths and weaknesses of the curriculum, and to yield comparisons with other districts, other states, the nation, and even other nations. It is hardly surprising that a testing program designed to serve well one of these purposes may do a relatively poor job of satisfying another expectation. The temptation may be to produce several specialized testing programs aimed at particular purposes. But, a proliferation of specialized testing programs, each designed with a particular purpose in mind, has serious drawbacks. A number of observers believe that an excessive amount of time is already devoted to testing (e.g., National Commission on Testing and Public Policy, 1990). The problems of high costs associated with producing and managing multiple testing programs should not be underestimated either.

Hence, there are strong pressures for the development of efficient testing systems that can serve multiple purposes simultaneously.

As Ansley, Forsyth, and Hoover (1989) have noted, "the desire on the part of consumers for more information from less testing time" (p. 1) is not unique to periods of increased emphasis on testing. It is natural, simply on the grounds of cost and efficiency, to want a test to serve multiple purposes. But expanding expectations for testing exacerbate this desire.

Among the many purposes for testing, two stand out with regard to their apparent differences in requirements for a testing program: (a) the need to obtain information about the performance of students relative to the specific aspects of a state or district mandated curriculum and (b) the need to obtain information about the performance of students in relation to a nationally representative sample of examinees. There is considerable demand for both types of information, criterion-referenced as well as norm-referenced information, but the two frequently seem to be in conflict, or, at least, to require separate testing programs. However, in many instances, legislative mandates require that both types of information be obtained from a single assessment.

The need for detailed information about performance of students relative to the objectives of a state or local curriculum requires the development and use of tests that are designed to match the specifics of the curriculum. Such a custom-made test needs to include items that assess performance relative to each of the important outcomes of the curriculum. That is, the test needs to be designed to match the curriculum. Such tests are frequently referred to as objective-referenced tests or criterion-referenced tests (Berk, 1984), but the key feature for present purposes is that they are designed to match the details

of the local curriculum. Hence, we will simply refer to them as curriculum-specific tests (CST).

Although CST results provide an assessment of current performance, and over time, of student progress relative to those specific objectives, they do not provide a basis for answering questions about how local student achievement compares to that of the nation on content that represents those broad areas often taught at particular grades. The latter type of information is obtained by the administration of norm-referenced tests (NRT). The content of an NRT is selected to provide broad coverage of objectives that are common to widely used textbooks and curriculum guides from various states and large school districts, but cannot be expected to match in detail the curriculum of a particular state or district. The California Achievement Tests, for example, are described as being "intended to measure a student's understanding of broad concepts as developed by all curricula rather than the student's understanding of the content specific to any particular instructional program" (CTB/McGraw-Hill, 1987, p. 2-1). An NRT may include some content not included in a particular curriculum or not covered until a later grade, it may exclude some objectives in the local curriculum, and it may differ from the local curriculum in terms of emphasis given to particular objectives.

The dilemma for local and state educators is that the dual needs for curriculum-specific information and national comparisons are met by neither a CST nor an NRT. In principle, national norms could be collected for a CST, but such a solution would be highly impractical for a school system or even a state department of education. As previously noted, the alternative of administering both tests, while possible, is quite time consuming and likely to lead to resistance from those who are concerned about the expense and amount of time devoted to testing. The approach to solving this dilemma that has been

used with increasing frequency in recent years involves a combination of the two types of tests. The resulting combination is called a *customized test with customized norms.*

Customized tests may take any of several forms, but the most common is a nationally normed standardized achievement test that has been modified so that the testing needs of a particular group (e.g., district, state) might be met better. Modifications can include anything from adding a few CST items, to substituting locally constructed items for a few NRT items, to substituting a CST for the complete NRT, and then using equating methods to obtain predicted NRT scores from the CST scores.

Customized tests have considerable appeal. They promise to efficiently accomplish multiple assessment goals. Thus, it is not surprising that they have attracted considerable attention and come to be used with increasing frequency in recent years. However, they also raise a number of questions regarding a wide range of practical and technical issues. Central among the questions that need to be answered are those that concern the validity of interpretations and uses of scores. There are several competing approaches to customized tests and customized norms, and we are just beginning to have the experience and research basis needed to consider the relative validity of the alternatives for particular purposes.

The validity of interpretations and uses of customized tests and customized norms is the focus of this paper. Some elaboration of the basic approaches to customized testing is needed before considering questions of validity, however. Thus, we begin with a brief description of four general customized testing approaches that are in current use. We then turn to a consideration of the fundamental questions regarding the validity of the uses and interpretations of customized test scores. This will lead to a discussion of the most widely used

analytical models and their underlying assumptions and to a review of the available research evidence. Finally, we will close with a set of recommendations regarding the use of customized testing and needed research.

Current Practice

In their desire to provide both curriculum-relevant as well as normative information, school districts and states, with the assistance of test publishers and measurement specialists have generated a plethora of testing programs. Since, in general, the needs and testing priorities in each school district and state are different, it is not surprising that the testing programs that have evolved are very different, too. For example, in some programs, emphasis has been placed on curriculum-relevant information, whereas, in others, norm-referenced information has been emphasized. In fact, one of the ways in which testing programs around the country can be distinguished is in terms of their emphasis on local objectives or normative comparisons. In a number of districts (e.g., New York and Philadelphia) and states (e.g., Connecticut), the assessment needs are being met through the development of customized tests and customized norms. Four general models that differ in terms of the degree of both test and norm customization can be identified. These four models, which are labeled *NRT-Only*, *NRT-Based*, *CST-Based*, and *CST-Only*, differ in terms of primary orientation and involve different levels of customization. Brief descriptions of the four models are provided in Table 1. Also listed in Table 1 are examples and some of the advantages and disadvantages of each model.

The NRT-Only model is one that has been prevalent for some time. This model uses an intact, off-the-shelf, norm-referenced test in the form in which the national standardization took place. Customization only occurs in the

reporting of additional scores for objectives specific to the local curriculum. There is no customization of the test instrument, only a choice or construction of score reports for clusters of test items that correspond to specific objectives. Work reported by Wilson and Hiscox (1984) provides an example of the NRT-Only approach. They used an intact NRT and added to the normally available NRT scores by obtaining percent correct scores for subsets of the NRT items that were selected to match their learning objectives. Of course, it might be said, too, that the NRT-Only Model permits customization to the extent that users can select the NRT that most closely matches their curriculum. And certainly, in many large districts and state adoptions, considerable time is spent by content experts and measurement specialists in reviewing available NRTs for their content suitability.

The NRT-Only approach yields no information about performance on local objectives that are not included in the NRT. Even for the objectives that are included on the test, the precision of the information will depend on the degree to which those objectives are emphasized on the test, which may or may not match the relative importance they are given within the curriculum.

The NRT-Based model, on the other hand, provides a means of responding to these issues of missing topics or a mismatch in emphasis on the NRT. In this model the full off-the-shelf NRT is administered, but additional items are also administered in order to increase the emphasis of content that is sparsely covered or not covered at all on the test but is an important part of the local curriculum (see Jolly & Gramenz, 1984, for an example of the NRT-Based model). The added items, which are usually contained in a separate test booklet, are not used in determining the norm-referenced scores. Norm-referenced scores are obtained in the usual fashion. The customization occurs only in the construction and reporting of

curriculum-specific or objective-referenced scores by combining appropriate subsets of the NRT and add-on items.

Both the CST-Only and the CST-Based models emphasize local objective information rather than normative comparisons. Test items are selected specifically for the local curriculum and customization is used to obtain norms. The CST-Based model is similar to the NRT-Based model in that both CST and NRT items are administered. In the CST-Based model, however, only a selected subset of the items from an NRT are administered. Normative comparisons are derived from special analyses of the selected NRT items alone or from a combination of those items with the CST items. Items from the NRT may be selected to best estimate a norm-referenced score from the subset used. Those items may be embedded into the CST or administered as a separate short test.

In some applications of the CST-Based model combined CST and NRT item pools are constructed. Selection of the NRT items to administer is determined primarily by the content they assess in relation to the local objectives rather than their utility for estimating norm-referenced scores. Estimation of norm-referenced scores in this design is usually based on a combination of the NRT and CST items which make up the assessment. Some results for the CST-Based testing system used in Philadelphia are presented by Green (1987), and descriptions of an application in New York City are provided by Dungan (1988) and by Taleporos, Canner, Strum, and Faulkner (1988).

As its name suggests, the fourth model, the CST-Only model, uses only items that are developed for the local curriculum. In this case the CST is equated to an NRT in order to derive norms for the CST scores. In this way, students receive norm-referenced scores without actually responding to any of the NRT items. The norm-referenced scores are derived

from the relationship between the CST and the NRT that is determined during the equating process. The reading assessment component of the Illinois Goal Assessment Program provides an illustration of the CST-Only model (Illinois State Board of Education, 1988).

A variety of designs and analytical procedures may be used for the equating. One frequently used design requires that both the CST and the NRT to which it is to be equated be administered to the same sample of students. Alternatively, two randomly equivalent groups may be formed and one of the two tests administered to each group. Most commonly, item response theory (IRT) models (e.g., Hambleton, 1989; Hambleton & Swaminathan, 1985) are used to calibrate the CST items and place them on the NRT scale. The IRT calibration provides the basis for generating NRT scale score estimates that can be converted to various types of norm-referenced scores such as percentile ranks, grade-equivalent scores, or normal-curve equivalent scores. Classical equating procedures can also be used, but they do not offer as much flexibility if multiple test forms are to be constructed from an item bank of CST items.

Applications of the models differ not only in the specific design used to obtain curriculum-specific and normative information, but also in the extent of each type of information that is generated and reported. As would be expected, the normative scores reported by the NRT-Only or NRT-Based models are generally more extensive (e.g., math computation, math applications, math problem solving, and total math) than with either of the CST models where norm-referenced scores are apt to be obtained only for total scores of a content area (e.g., total math). The converse is often true with regard to the objective-referenced scores, especially in the case of the NRT-Only model in comparison to either of the CST models.

The four types of models described above form a continuum. At one end is a test built specifically for norm-referenced interpretations from which some curriculum-specific, objective-referenced information is reported. At the other end is a test built specifically to provide information about performance relative to the objectives of a specific curriculum from which norm-referenced information is reported.

The four models represent different compromises between the competing requirements for norm-referenced and curriculum-specific information. At the NRT end of the continuum the information about specific curriculum objectives is incomplete and less than ideal, while the normative information is apt to be less precise and detailed at the CST end of the continuum. When the CST information is incomplete or skimpy for highly valued objectives of the local curriculum, it is generally apparent to the user. If too few items are used to assess an objective or an objective is not assessed at all, the limitations are relatively self-evident to users who are familiar with the curriculum objectives. The limitations can be taken into account to some degree in interpreting the results. Such safeguards are largely lacking in the case of norm-referenced interpretations, however. Lack of precision or systematic biases in norm-referenced scores are apt to be less obvious to users. Consequently, the potential for misinterpretation and misuse is greater in the latter case. Therefore, the validity of normative interpretations of customized test scores deserves particularly careful consideration.

Validity

The widespread use of test and norm customization in recent years has raised concerns about the validity of both the objective-based and the norm-referenced uses and

interpretations of the scores. The purpose of customization is to accomplish multiple assessment purposes efficiently, thereby minimizing the testing time and burden. The question is whether a test can serve multiple purposes and retain an adequate level of validity for each purpose. Validity questions can be raised about test content and inferences about the accomplishment of specific curriculum objectives, which are both of central concern for a CST type of assessment. Validity questions can also be raised about the normative interpretations of the scores in an NRT assessment.

Model Assumptions

As was previously noted, the most promising and potentially powerful approaches to customized testing rely on IRT models for item calibration and the conversion of student responses to the NRT scale. The potential utility of IRT for this application derives from the invariance properties of item parameters and person proficiency values when the assumptions of the IRT model are satisfied. These invariance properties, which Wright (1968) more colorfully described as person-free item calibration and item-free person measurement, are critical not only to the use of IRT for customized testing, but for a number of other applications such as computerized adaptive testing and item banking. Because of the importance of these putative properties of IRT models for customized testing, they deserve some elaboration.

Person-free item calibration implies that items can be calibrated using a sample of students from a local district or a state just as well as with a national sample. If the assumptions of unidimensionality and local independence, on which this property depends, are satisfied, then estimates will differ only due to sampling error. Thus, estimates based on a sample of, say, 1,000 students from a single school district or state who

vary widely in achievement levels would provide just as good a basis for item calibration as a nationally representative sample of 1,000 students with an equally wide range of achievement. This property is potentially valuable for customized testing applications because it means that the CST items can be calibrated together with a subset of NRT items based on an administration within a given state or district and from that calibration the CST items parameter estimates may be placed on the NRT scale.

The item-free person measurement property is equally important for customized testing. This property allows the computation of NRT scores for an individual student from any set of items that are calibrated on the NRT scale. The precision of the scores will depend on the number of items and their parameters, but except for these differences in measurement error, any set of items can produce valid estimates of a student's standing on the NRT scale. As in the case of person-free item calibration, the promise of item-free person measurement depends on the data satisfying underlying IRT model assumptions.

No mathematical model of human behavior is precisely correct, and IRT cannot be expected to be an exception to this general observation. The assumptions of the model are not perfectly satisfied by any set of responses of a large sample of people to real test items. Models do not have to be exactly right to be useful, however. The important question is not whether a model is exactly correct or if all assumptions are perfectly satisfied. Rather, the questions of interest concern the adequacy of the approximations of data to a model, the accuracy of model-based predictions, and the validity of inferences based on applications of the model.

Dimensionality and Content Match

Neither the typical NRT nor the typical CST is unidimensional (see, for example, Linn, 1990; Yen, Green, & Burket, 1987). Indeed, if such tests were unidimensional, there would be no need for concern about content coverage and representation. Unidimensional IRT models may be useful nonetheless for such purposes as the equating of CST and NRT scores. "Multidimensionality does not preclude the use of a unidimensional procedure to produce an accurate equating. However, it is essential that the tests be matched for multidimensionality" (Yen, Green, & Burket, 1987, p. 11).

This conclusion is supported by research with both simulated and real test data conducted by Hirsch & Keene (1989). They constructed simulated NRTs and CSTs that each had two underlying dimensions. Unidimensional IRT equatings worked well with the simulated data when both tests had similar structure, that is, involved comparable weightings of the two underlying dimensions. Large errors in estimated norm-referenced achievement levels derived from CST item sets were found when the structures of the two simulated tests differed substantially, however. Hirsch and Keene (1989) also found that the adequacy of the equatings of the real data sets was closely related to the comparability of the dimensional structures of the tests to be equated.

This notion of matching for multidimensionality is closely related to advice of several authors (e.g., Holmes, 1986; Lenke, 1989; Yen, Green, & Burket, 1987) that the content coverage of a customized test needs to be carefully matched to the content of the NRT to which it is being equated. Customized norms are apt to be distorted when a content category is disproportionately represented on the customized test and students from the state or district where the customized test is

being used do particularly well or particularly poorly on that content (Linn, 1990; Yen, Green, & Burket, 1987).

Effects of Content Mismatch

Several researchers have re-analyzed subsets of items from an NRT to investigate the degree of correspondence between full-length NRTs and norm-referenced estimates obtained from reduced item sets. Harris (1987), for example, constructed three customized subtests by scoring items from either 3 or 4 of the 6 content categories of the Mathematics Test of the American College Testing (ACT) Program. In general, there was relatively poor agreement in estimated scores between the customized subtests and the full-length ACT. Her results add to the caution provided by others that it is important to assure that the customized test and the NRT have proportional coverage of the content categories.

Three investigations related to the issue of content coverage and match with the NRT have been conducted by researchers at the University of Iowa using subsets of items from an off-the-shelf NRT to obtain predicted norm-referenced scores on the full-length test (Allen, Ansley, & Forsyth, 1987; Ansley et al., 1989; Way, Forsyth, & Ansley, 1989). These studies may be thought of as a type of simulated customized testing where the customized test represents only a part of the content of the NRT. They also illustrate a special type of customization where norm-referenced achievement test items that do not match the local curriculum are deleted from the test or from scoring to obtain "curriculum-referenced norms" (Hambleton, Gower, & Rogers, 1989).

In the series of three studies conducted by researchers at the University of Iowa (Allen et al., 1987; Ansley et al., 1989; Way et al., 1989) items measuring particular content areas

were deleted from tests on either the Iowa Tests of Basic Skills (ITBS) or the Iowa Tests of Educational Development (ITED). For example, Way et al. (1989) deleted 18 language expression items and then computed customized norm-referenced Language scores based on the remaining 22 usage items. Similar content-related deletions were made on three other ITBS subtests by Way et al. and on the Quantitative Thinking test of the ITED by Allen et al. (1987). In the third study (Ansley et al., 1989), deletions of items on tests of the ITBS were made based on a comparison to objectives of the Texas Essential Elements. Items on the ITBS tests were deleted if they did not correspond to the stated Texas Essential Elements.

In each of the Iowa studies customized norm-referenced scores were computed based on the reduced item sets and compared with the corresponding norm-referenced scores for the full NRT. Customized and full NRT mean scores were then compared for schools selected to simulate schools that customized an NRT by deleting items that did not match their curriculum in two of the studies (Allen et al., 1987; Way et al., 1989). In the third study where selections were based on the objectives of the Texas Essential Elements (Ansley et al., 1989), comparisons were made using data from a large Texas school district.

In all three studies the customized or "curriculum-referenced norms" resulted in scores that were generally higher than those obtained using the full NRT. Ansley et al. (1989), for example, concluded that in "many cases, it would seem that individuals, and consequently school systems, would improve their relative performance considerably by administering a customized test. Although some of the results ... indicated that customized tests produced only slightly different ability estimates, the trends observed, ... together with the results reported by Allen et al. (1987), Gramenz, Johnson, and Jones (1982), and Way et al. (1989), certainly

seem to indicate that the use of customized tests must be undertaken very cautiously" (p. 17).

Perspectives on "Overestimation"

The validity implications of systematically higher scores depend on the interpretations and uses of the scores. If the customized score is used as the basis for reporting how well a student, a school, or a school district performs compared to the nation on the general content measured by an NRT, then the systematically higher score will mislead. That is, the inflation of the scores will be a source of invalidity. The inflation is apt to contribute to an exaggerated notion of achievement.

There is another perspective on this issue, however. Hambleton, Gower, and Rogers (1989), for example, have noted that one of the reasons for wanting customized scores in the first place is that an NRT may cover content not included in a local curriculum or not taught until a later grade. Hence, it may be argued that the inclusion of this untaught content on the standard NRT may lead to an underestimation of student performance on the content that is taught. In this situation, the customized test may be a more valid measure of the local curriculum than the NRT, but lead to less valid NRT scores.

This alternative perspective raises difficult questions regarding the nature of the inferences that can and should be made from customized test results. One possible interpretation is that the score represents the relative standing that would be obtained if the NRT contained only that subset of items that are included in the customized test. To test this interpretation, the national norms would need to be re-computed for the particular subset of items in question. Even if such analyses supported this interpretation, however, one would still be faced with considerable problems in communicating the results.

Consider, for example, two hypothetical school districts, both of which score at the national median on a full NRT. District A creates a customized test using the 80% of the items that correspond to its curriculum and obtains an average score at the 55th percentile according to the customized norms. District B, on the other hand finds that only 50% of the NRT items correspond to its curriculum and for that 50% the customized norms put the school average at the 60th percentile. Which district has the relatively higher achievement? In what sense is either district performing better than the national average? Clearly, simply reporting that District A scored at the 55th percentile and District B scored at the 60th percentile provides an incomplete and probably misleading picture. Such reporting would be likely to exacerbate the "Lake Wobegon" phenomenon, that is, the tendency for almost all states and most districts to report NRT results that are above the national average (Cannell, 1987; Linn, Graue, & Sanders, 1990).

It might be noted that when a district selects an NRT they are commonly advised to carefully review the content of the test in comparison to the school's or district's instructional program and curriculum guidelines. The ITBS Manual for School Administrators (Hieronymus & Hoover, 1986), for example, provides the following advice for selecting achievement tests: "The two most important questions in the selection and evaluation of achievement tests for your school should be as follows:

1. Are the specific skills and abilities required of the pupil for successful test performance precisely those that are appropriate for the pupils in our school?

2. Do the test exercises in themselves adequately define our objectives of instruction?" (p. 74).

Inasmuch as schools or districts follow this advice, there is a process analogous to a limited amount of customization that takes place at the time tests are selected (Good & Salvia, 1988). To use the above example of hypothetical districts A and B, one could imagine that both districts would be at the national average on the joint administration of, say six different NRTs provided by several different publishers, but on the specific NRT selected by District A the district average is at the 55th percentile while on a different NRT selected by District B to better match its curriculum the district average is at the 60th percentile. The questions about which district has the higher relative performance and whether they are indeed above the national average pertain here just as they would in the case of customized norms. Although there are a number of other factors such as the use of old norms and teaching to the test that must also be considered, the selection of tests to match curricula in operational NRT testing programs but not in the development of norms may be one of the factors that has contributed to the "Lake Wobegon" effect (Koretz, 1988; Linn et al., 1990; Shepard, 1990).

The studies conducted by Allen et al. (1987), Ansley et al. (1989), Harris (1987), and Way et al. (1989) all involved calculations for subsets of items covering some but not all content categories of an NRT. Those results along with results reported by Hirsch and Keene (1989), Linn (1990), and Yen et al. (1987) all suggest that to make national comparisons more valid, at a minimum, customized tests need to sample content categories in proportion to the coverage of those content categories on the NRT. Even with proportional content coverage, however, questions remain about the adequacy of estimates that can be obtained by using a reduced-length NRT.

Test Length

Harris (1988) investigated the effect of changing test length while maintaining proportional coverage of the content categories using the ACT Mathematics Test. Shortened tests of length 10, 20, and 30 items were constructed from the full-length 40-item test maintaining the balance of content coverage across the six content categories of the ACT Mathematics Test to the extent possible. Harris found sizeable differences between the reduced length and full-test results, which led her to conclude that "test length, in and of itself, is a potent enough factor to make comparisons between total intact tests and shortened customized tests unwise" (Harris, 1988, p. 14).

Qualls-Payne, Raju, and Groth (1989) used short versions of one form of an NRT (referred to as the "core tests") to estimate the proportion correct scores for the alternate form of the test. The alternate form of the test was treated as if it were a CST and then the national proportion correct scores (p-values) were estimated from a scaling of those items together with the core test items from the first form, and those were compared to the actual national p-values. Items from the core tests of length 10, 20, or 30 items were selected to provide proportional content coverage and average item difficulties that were approximately equal to the full form of the test. Their results indicated that very good estimates could be obtained of the p-values on the alternate form of the test using IRT scaling methods for even the shortest core test.

The Qualls-Payne et al. (1989) results are more encouraging for applications than most of the studies that have been discussed above. It might be noted, however, that the simulated CST items consisted of an alternate form of the NRT and therefore might be expected to have the same basic dimensional structure as the core tests with proportionally

selected content, and it is under these conditions that Hirsch and Keene (1989) found close correspondence between customized and NRT norm-referenced scores. Whether the Qualls-Payne results would generalize to a CST consisting of locally-constructed items with an underlying structure and content representation that differed from those of the core NRT items to a greater degree remains to be determined.

Combined CST-NRT Analyses

With the exception of the Hirsch and Keene (1989) and Yen, Green and Burket (1987) papers, all of the previously discussed studies have involved analyses of NRT items to simulate various customized testing situations. The following three studies conducted by Dungan (1988), by Green (1987), and by Hambleton and Martois (1983) involved combinations of CST and selected NRT items. In both the Dungan (1988) and Green (1987) studies, IRT calibration was used to place locally-constructed CST items on the NRT scale, then the two sets of items were used together to obtain norm-referenced estimates. In the Hambleton and Martois (1983) study, IRT calibration involving a national sample was used to place a large collection of test items on a common scale. One set of 50 items was administered to a nationally representative sample of examinees to produce test score norms. Three customized tests that differed substantially in their difficulty levels were constructed from the same calibrated item bank, then comparisons were made between predicted NRT performance using the customized test and actual NRT performance.

In the study reported by Dungan (1988) samples of grade 4 and grade 6 students responded to the complete Mathematics Tests (95 items) of Form M of the Metropolitan Achievement Test, Sixth Edition (MAT6) together with a short CST in mathematics. At each grade there were five different

CST forms, each consisting of 20 items that were administered to different samples of students together with the MAT6. The CST items were calibrated to the MAT6 scale and then substituted for the 20 easiest MAT6 items within each of the three subtest areas reported for the MAT6 (Concepts, Problem Solving, and Computation) to obtain norm-referenced score estimates. That is, customized norm-referenced estimates were computed as if a student had responded to 75 of the 95 MAT6 items plus the 20 calibrated CST items for a given form. Those customized estimates were then compared to the scores obtained from the intact MAT6. Although the mean of the customized norm-referenced test was higher than that for the complete MAT6 in all ten cases (five CST forms at each grade), the differences between the pairs of means were quite small in every case (ranging from a low of 0.3 to a high of 1.5 scaled score points where the standard error of measurement for a scaled score is approximately 12 points).

The Dungan study controlled both content coverage and test length. However, content coverage was controlled at the subtest level rather than at a more detailed level. Thus, if the 20 items on a CST form consisted of 9 concepts items, 7 problem solving items, and 4 computation items, then the 9, 7, and 4 easiest concepts, problem solving, and computation MAT6 items, respectively, were deleted and replaced by the corresponding CST items to obtain customized norm-referenced scores. Given the difference in difficulty, the results appear quite encouraging for situations where length and general content coverage can be maintained but there is a desire to alter difficulty.

Green (1987) analyzed results for specially selected NRT items and calibrated CST items over a period of three years. The NRT items were selected from a California Test Bureau (CTB) item pool scaled to Form U of the Comprehensive Tests of Basic Skills (CTBS). Locally constructed CST items were

calibrated on the CTBS scale. Two customized norm-referenced reading comprehension score estimates (one based only on CST items and one based only on CTB items) were computed for three consecutive years for students in grades 4 and 6.

Assuming that instruction emphasized the content of the newly instituted CST items more than that of the CTB items, one might expect that the CST norm-referenced score estimates would increase more from year to year than the CTB estimates would. There was some limited support for this expectation at grade 6 where the difference in median scaled scores was -1.2, 0.9, and 1.3 in years 1, 2, and 3, respectively, where positive numbers indicate that the CST median is higher than the CTB median. However, these differences are all quite small in comparison to the standard errors of the individual median scores which ranged from 2.2 to 3.4. Furthermore, the differences between the medians for the three years (-0.5, -5.7, and 1.4 for years 1, 2, and 3) revealed no such pattern.

Both the CST and CTB based norm-referenced score estimates went up substantially from year 1 to year 3 (about 20 scale score points at the median for grade 4 and 10 at the median for grade 6). Since Green did not have an intact NRT for comparison, it is unknown whether comparable increases would be obtained using an off-the-shelf test. It is also unclear that instruction was focused more heavily on the CST items than on the CTB items. Despite these unanswered questions, Green's results are encouraging for applications that derive norm-referenced estimates from a combination of selected NRT and calibrated CST items.

In the Hambleton and Martois (1983) study, examinees took one of three customized tests (assigned at random) and a norm-referenced test which were all linked to a common achievement scale. The customized tests were matched in content and length to the norm-referenced test but they

differed in their difficulty. Customized tests were constructed to be considerably easier, considerably harder, or similar in difficulty to the norm-referenced test. The study was carried out in three content areas (reading, language arts, and mathematics) and two grade levels (2 and 5).

Interest in the analysis was centered on the comparison between the actual norm-referenced test scores in each subject area and the predicted test scores obtained from one of the these customized tests (easy, medium, difficult) drawn from the item bank. Results of this study were promising. Predictions from the customized tests showed almost no bias. Differences in the difficulty level of the tests seems to adversely affect prediction accuracy, but not to a substantial degree. Overall, prediction errors were not much larger than the standard error of measurement for the NRT.

Yen, Green, and Burket (1987) supported the testing design used by Hambleton and Martois as one that produces norm-valid scores, provided the item statistics are properly estimated and the content covered in the customized test is proportional to the content covered in the normed test.

Context Effects

A potentially important issue that is not addressed in any of the previously discussed studies is the influence of context on estimated item parameters and examinee scores. If item parameters are influenced by the sequential order in which they appear or the specific surrounding items, then misleading estimates of performance may result when NRT items are selected and administered in a context different from the one for which norms were obtained.

Leary and Dorans (1985) reviewed research on context effects. Much of the early research was largely focused on examinee scores and, in cases where items were considered, classical item statistics. As Leary and Dorans indicated, the early studies yielded mixed results. Item position was found to have some effect on item difficulty for some tests but not others. Some item types appear to be more sensitive to context effects than others. Items associated with reading passages, for example, tend to be more difficult when the passage and items are located toward the end of a test section than when they are located near the beginning.

Using two IRT models with items from the California Achievement Tests, Yen (1980) found that item parameters were substantially affected by context. These effects appeared to be at least partially the result of item position. Wise, Chia, and Park (1989) also found that IRT item parameters varied as a function of item position. The effects were strongest when tests are relatively difficult for the group of examinees for which the items are calibrated. Based on the findings of Yen and of Wise and Park, it seems wise to maintain the relative position of NRT items when constructing customized tests.

Changes in the context in which items were presented contributed along with several other factors to the anomalous results obtained for the 1986 National Assessment of Educational Progress (NAEP) in reading (Beaton, 1988; Beaton & Zwick, 1990; Haertel, 1989). Zwick's (1990) conclusion that "common-item equating procedures should not be assumed to be appropriate" (p. 109) when there are changes in item position or context is particularly relevant for customized testing applications where items from an NRT item pool are sometimes embedded in a CST.

Concerns about context effects contributed to the conclusion that it is important to use "intact blocks of items for

purposes of scale equating in NAEP" (Zwick, 1990). It would seem prudent to take similar precautions in customized testing applications. That is, it would be desirable to control item position and where possible to use an intact section of an NRT when calibrating CST items.

Conclusion and Recommendations

Customized tests and customized norms can yield valid information about performance both in relation to specific curriculum objectives and in relation to national norms. This has been successfully demonstrated in a number of studies, albeit under special conditions, notably similar context configurations. There are many threats to validity of the normative interpretations, however. Cautious application and frequent checks on the validity of the norm-referenced interpretations are needed in order to avoid potentially misleading inferences about student achievement. It is in this light that the following recommendations are offered:

1. The content of a customized test should be closely matched to the content of the norm-referenced test. That is, if CST items are substituted for selected NRT items, proportional coverage of content categories is needed for the CST items to be used for computing normative scores. Likewise, if a CST is substituted for an entire NRT, validity of score interpretations will be enhanced if the CST matches the content specifications of the NRT it replaces.

2. Additional content areas or extra coverage of content that is sparsely covered by the norm-referenced test may be added and used for other purposes, but should not be part of the calculation of norm-referenced scores.

3. Test length and test difficulty of the customized test should be similar to that of the norm-referenced test. In general, the more the customized test parallels the NRT in content and statistics, the fewer the concerns about valid score interpretations.

4. When subsets of norm-referenced items are embedded in a customized test, the position of each norm-referenced item should be similar to its position in the original norm-referenced test.

5. Where feasible, in the CST-Based Model we recommend using intact blocks of norm-referenced items or what Wainer and Kiely (1987) have called testlets rather than individual items in order to reduce the likelihood of context effects.

6. Equating results should be investigated periodically, say every two or three years, to verify that the relationship between the customized test and the norm-referenced test has not changed.

7. Additional research is needed on a number of topics related to customized testing, including, for example, (a) differential effects of curriculum and test content match, (b) content coverage and dimensionality match effects, (c) strengths and weaknesses of alternative approaches to customized testing, (d) context effects, (e) analysis of estimated normative scores for low-, middle- and high-achieving examinees, and (f) evaluation of equating designs and IRT models for customized testing.

In summary, when making a decision about whether or not to customize a test to meet the goals of a multi-purpose test program, in addition to the costs and time required to complete

the work, the validity of both the resulting norm-referenced interpretations and the CST scores must be considered. The NRT-Only and NRT-Based models preserve the validity of the norm-referenced interpretations but the validity of the CST scores in these models, in general, is lower than with one of the CST models. The gap can be closed in the NRT-Based model by choosing the NRT wisely and adding necessary items in an additional test booklet administered with the NRT.

On the other hand, the CST-Based and CST-Only models are likely to provide users with better curriculum-relevant information but the validity of the derived NRT scores and associated norm-referenced interpretations will generally be lower than in one of the NRT models. The magnitude of the loss in validity of the derived NRT scores will depend on the test customization approach that is used. The recommendations above provide guidelines for minimizing the loss of validity in norm-referenced interpretations associated with the CST-Only and CST-Based models.

References

Allen, N.A., Ansley, T.N., & Forsyth, R.A. (1987). The effect of deleting content-related items on IRT ability parameters. *Educational and Psychological Measurement, 47*, 1141-1152.

Ansley, T.N., Forsyth, R.A., & Hoover, H.D. (1989, March). *Test customization: Can we have our cake and eat it too?* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Beaton, A.E. (1988). *The NAEP 1985-86 reading anomaly: A technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A.E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (NAEP Report No. 17-TR-21). Princeton, NJ: Educational Testing Service.

Berk, R.A. (1984). *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends of Education.

CTB/McGraw-Hill. (1987). *California Achievement Tests, Forms E and F, Technical Report.* Monterey, CA: CTB/McGraw-Hill.

Dungan, L.A. (1988, April). *Norm-referenced test customization: Validation of individual score interpretations.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Goldsby, C. (1988, April). *Norm-referenced test customization: Curricular considerations.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Good, R.H., & Salvia, J. (1988). Curriculum bias in published norm-referenced reading tests: Demonstrable effects. *School Psychology Review, 17*, 51-60.

Gramenz, G.W., Johnson, R.C., & Jones, B.G. (1982, March). *An exploratory study of the concept of curriculum-referenced norms using the Stanford Achievement Test - 6th edition.* Paper presented at the meeting the National Council on Measurement in Education, New York City.

Green, D.R. (1987, April). *Local versus national calibrations.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Haertel, E. (1989). *Report of the NAEP technical review panel on the 1986 reading anomaly, the accuracy of NAEP trends and issues raised by state-level NAEP comparisons* (National Center for Education Statistics Technical Report CS 89-499). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Hambleton, R.K. (1988) Customized norm-referenced testing: Some comments. *Proceedings of the National Association of Test Directors, 4*, 58-66.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 147-200). New York: Macmillan.

Hambleton, R.K., Gower, C., & Rogers, H.J. (1989, March). *Customized norm-referenced testing: A review of issues and methods.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Hambleton, R.K., & Martois, J.S. (1983). Evaluation of a test score prediction system based upon item response model principles and procedures. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 196-211). Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic Publishers.

Harris, D.J. (1987, April). *Estimating examinee achievement using a customized test.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Harris, D.J. (1988, April). *An examination of the effect of test length on customized testing using item response theory.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hieronymus, A.N., & Hoover, H.D. (1986). *Iowa Tests of Basic Skills Forms G/H. Manual for School Administrators, Levels 5-14.* Chicago, IL: Riverside Publishing Company.

Hirsch, T.M., & Keene, J.M. (1989, March). *An examination of the effects different dimensional test structures have on test equating.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Holmes, S.E. (1986, April). *Multi-purpose tests: A solution to test proliferation.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Illinois State Board of Education. (1988). *The 1988 Illinois Goal Assessment Program: Technical manual.* Springfield, IL: Illinois State Board of Education.

Jolly, S.J., & Gramenz, G.W. (1984). Customizing a norm-referenced achievement test to achieve curricular validity: A case study. *Educational Measurement: Issues and Practice, 3,* 16-18.

Keene, J.M., & Holmes, S.E. (1987, April). *Obtaining norm-referenced test information for local objective-referenced tests: Issues and challenges.* Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15, 46-52.

Leary, L.F., & Dorans, N.J. (1985). Implications of altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413.

Lenke, J.M. (1989, March). *Norm-referenced scores for customized tests: Issues and solutions.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Linn, R.L. (1987). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1*, 181-198.

Linn, R.L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education, 3*, 115-141.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*, 5-14.

Madaus, G.F. (1985). Public policy and the testing profession—You've never had it so good? *Educational Measurement: Issues and Practice, 4*, 5-11.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: Boston College, National Commission on Testing and Public Policy.

Pipho, C. (1985). Tracking the reforms, part 5: Testing—Can it measure the success of the reform movement? *Education Week*, May 22, p. 19.

Qualls-Payne, A.L., Raju, N.S., & Groth, M.A. (1989, March). *Accuracy of the estimation of national item p-values of a customized test as a function of core test length, sample size, and IRT model.* Paper presented at the annual meeting

of the American Educational Research Association, San Francisco.

Schattgen, S.F., & Osterlind, S.J. (1989, March). *The validity of norm-referenced information obtained from an objective-referenced test using the ORT-Only model.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Shepard, L.A. (1990). Inflated test score gains: Is it old norms or teaching the test? *Educational Measurement: Issues and Practice, 9,* 15-22.

Taleporos, B., Canner, J., Strum, I., & Faulkner, D. (1988, April). *The process of customization of the Metropolitan Achievement Test (MAT-6) in mathematics for New York City public school students.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Way, W.D., Forsyth, R.A., & Ansley, T.N. (1989). IRT ability estimates from customized achievement tests without content sampling. *Applied Measurement in Education, 2,* 15-35.

Wilson, S.M., & Hiscox, M.D. (1984). Using standardized tests for assessing local learning objectives. *Educational Measurement: Issues and Practice, 3,* 19-22.

Wise, L.L., Chia, W.J., & Park, R.K. (1989, March). *Item position effects for test word knowledge and arithmetic reasoning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wright, B.D. (1968). Sample free test calibration and person measurement. In *Proceedings of the 1967 ETS Invitational Conference on Testing Problems* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Yen, W.M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297-311.

Yen, W.M., Green, E.R., & Burket, G.R. (1987). Valid normative information from customized achievement tests. *Educational Measurement: Issues and Practice, 6*, 7-13.

Zwick, R. (1990). Adjustment of 1986 reading results to allow for changes in item order and context. In A.E. Beaton & R. Zwick (Eds.), *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (NAEP Report No. 17-TR-21) (pp. 87-109). Princeton, NJ: Educational Testing Service.

Table 1. Features of the Major Models for Test and Norm Customization

	Model Description	Advantages	Disadvantages	Examples
NRT-Only	• An intact off-the-shelf NRT is selected. • Normative scores are reported as well as any CST scores of interest which can be formed from the available NRT items. • Items measuring inappropriate skills can be eliminated from the CST reporting.	• No additional testing or costs. • NRIs are not compromised.	• CST information is limited to skills/items in the NRT. • Skills of interest may not be measured. • Important skills may be measured with only an item or two. • Publisher's CST reporting may be inappropriate and/or costly to change.	• Wilson & Hiscox (1984) provide steps for districts to follow in identifying items of interest on the NRT. • Oklahoma School Testing Program (Keene & Holmes, 1987)
NRT-Based	• An intact off-the-shelf NRT is selected (like the NRT-Only). • Additional items prepared/selected by the district or state are (usually) placed in a separate test and administered to examinees at the same time as the NRT. • Usually the additional (customized) items are *not* included in the NRT scores.	• Validity of NRIs is identical to the NRT-Only model (as long as customized items are administered as separately timed sections and are not used in compiling NRT scores). • CST reporting which involves combining items from the NRT and the additional items as appropriate, provides curriculum-relevant information. • Utility of the testing program is enhanced. • Both face and content validity are enhanced. • Even if replacement items are used in the NRT, if the number of replacement items is small, the threat to NRT validity is likely to be small also. • Added costs are likely to be small (compared to CST and CST-Only).	• Testing time and costs are increased (relative to the NRT-Only model). • Doesn't address a district's or state's concerns about inappropriate content in the NRT. • Extra time and cost is involved in preparing the local curriculum-specific items. • If the additional items are used in the NRT itself, original test standardization and time limits are violated with an unknown influence on the validity of NRT scores.	• Palm Beach County, Florida (Jolly & Gramenz, 1984) • Hawaii State Testing Program (Keene & Holmes, 1987)

Table 1 (continued)

	Model Description	Advantages	Disadvantages	Examples
CST-Based	• CST items are substituted for NRT items *or* CST items are added to the NRT and used in NRT scoring. • This model places more emphasis on test content/ match to the curriculum than normative information. • Several variations exist for computing NRT scores: 1. predicted from the remaining NRT items only; 2. predicted from a combination of available NRT and CST items; 3. customizing norms by basing them only on the NRT items in the customized test—known as "curriculum-referenced norms."	• Usefulness of test results for curriculum review is enhanced (emphasis on content issues rather than normative scores). • Teachers and administrators are less threatened by results. • Curriculum is less likely to be revised to match the test content.	• Value of NRT scores is reduced (to an extent that depends upon many factors including the amount of test customization and the match of content in the CST-Based test and the original NRT).	• New York City (Taleporos, et al., 1988) • Philadelphia (Green, 1987)
CST-Only	• A completely customized CST is constructed—NRT scores are obtained by equating the CST to the NRT (a common method of equating might involve including an anchor test of items from the NRT in the CST or as an add-on to the CST). • Test content is central.	• CST scores are available on skills of interest with the numbers of items of interest. • Under certain conditions valid NRT scores are available.	• Validity of NRT scores reduced by an amount that depends on many factors, including the design used in equating, the frequency of equating, and CST-NRT content equivalents. • There is a tendency for positive bias to be present in the predicted NRT scores. • Both cost of test development and time needed to do the job well can be substantial. • Test equating is another difficult and expensive activity.	• Illinois Goal Assessment Program (Illinois State Board of Education, 1988) • Districts and other states in their Chapter I reporting (e.g., Connecticut, Missouri) (Schattgen & Osterlind, 1989)