CUSTOMIZED TESTS AND CUSTOMIZED NORMS

CSE Technical Report 325

Robert L. Linn & Ronald K. Hambleton

UCLA Center for Research on Evaluation,
Standards, and Student Testing

July 1991

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

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 exacerbates 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 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 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. 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.

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.

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 • Publishers' CST reporting may be inappropriate and/or costly to change	• Wilson and 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 the 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 • face and content validity are enhanced	• 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 (cont'd)

Model	Description	Advantages	Disadvantages	Examples
NRT-Based		• 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)		
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	• usefulness of test results for curriculum review is enhanced (emphasis is 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) • Cleveland & Columbus, Ohio • Indiana, New Mexico, Tennessee

Table 1 (cont'd)

Model	Description	Advantages	Disadvantages	Examples
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 is reduced by an amount that depends on many factors, including the design used in equating, the frequency of equating, and CST-NRT content equivalence • there is a tendency for a positive bias to be present in the predicted NRT scores • the cost of test development and time needed to do the job well can be substantial • test equating is a 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)

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. 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 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 safe guards 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 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 that are of central concern for a CST type of assessment, as well as the normative interpretations of the scores, that are central to an NRT assessment.

Model assumptions. As 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, for example, 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 of 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 (e.g., 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 et al., 1987, p. 11).

This conclusion is supported by research with 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 et al., 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 et al., 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 three or four of the six 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, Forsyth, & Hoover, 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 et al. (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 depends 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; 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 percent 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 percent of the NRT items correspond to its curriculum and for that 50 percent 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: 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).

When a district selects an NRT it is 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, for example, 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 it is 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 Harris (1987), Allen et al. (1987), Ansley et al. (1989), and Way et al. (1989) involved calculations for subsets of items covering some but not all content categories of an NRT. Those results along with those 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 6 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. 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 et al. (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), Green (1987), and Hambleton and Martois (1983) involved combinations of CST and selected NRT items. In both the Dungan and Green studies IRT calibration was used to place locally-constructed CST items on the NRT scale and then the two sets of items are used together to obtain norm-referenced estimates. In the Hambleton and Martois 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, and 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 5 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 3 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 10 cases (5 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 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 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 seem 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 et al. (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 et al., 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. 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 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 with 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 (e.g., 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 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 the resulting norm-referenced interpretations as well as 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 Rep. 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.

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.

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 of the National Council on Measurement in Education, New York.

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. (Chair). (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 Tech. Rep. No. 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: Kluwer Academic.

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.

Hieronymus, A.N., & Hoover, H.D. (1986). *Iowa Tests of Basic Skills forms G/H. Manual for school administrators, levels 5-14.* Chicago: Riverside.

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.

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 meeting of the American Educational Research Association, San Francisco.

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

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.

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. *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 Rep. No. 17-TR-21, Chapter 6, pp. 87-109). Princeton, NJ: Educational Testing Service.