DOCUMENT RESUME

ED 138 117 FL 008 543

AUTHOR Doyle, Vincent
TITLE A Critique of the Northwest Regional Educational Laboratory's Review of the Mat-Sea-Cal Oral Proficiency Tests.
INSTITUTION Idaho Univ., Moscow. Coll. of Education.
SPONS AGENCY Center for Applied Linguistics, Arlington, Va.
PUB DATE Oct 76
NOTE 27p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Bilingual Education; Bilingualism; *Bilingual Students; Evaluation; *Evaluation Criteria; *Language Proficiency; *Language Tests; Measurement Techniques; Scoring; *Speech Communication; Test Bias; Test Construction; Testing; Testing Problems; Test Interpretation; Test Reliability; *Test Validity; Verbal Tests
IDENTIFIERS Language Dominance

ABSTRACT

This paper presents a critique of the Northwest Regional Educational Laboratory's (N.W.R.E.L.) review of the Mat-Sea-Cal Oral Proficiency Tests in their publication, Oral Language Tests for Bilingual Students. That publication was released in July, 1976 as a guide to administrators and program coordinators in the selection of instruments for assessing students' language dominance and oral proficiency (-ies). In rating each instrument, four criteria were explored: measurement validity, examinee appropriateness, technical excellence, and administrative usability. Several questions within each criteria were examined in determining the overall criteria rating. A descriptive review of the Mat-Sea-Cal is presented and the reviewer's rating is summarized in a chart. This critique scrutinizes the evaluations rendered to the Mat-Sea-Cal by the reviewers in each of the four criteria. Discussion is offered on several points. Differences in perception between the author and the N.W.R.E.L. reviewers on the evaluation of the Mat-Sea-Cal are enumerated. (Author/CFM)

A joint endeavor by, the

Center for Applied Linguistics
(Arlington, Virginia)

and, the

University of Idaho's
College of Education
(Moscow, Idaho)

October, 1976

A CRITIQUE OF THE NORTHWEST REGIONAL EDUCATIONAL LABORATORY'S REVIEW OF THE MAT-SEA-CAL ORAL PROFICIENCY TESTS

(as presented in the N.W.R.E.L. publication, Oral Language Tests for Bilingual Students)

by

Vincent Doyle,

Research Associate

Center for Applied Linguistics,

stationed at the University of Idaho,

Moscow

October, 1976

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

INTRODUCTION

This paper presents a critique of the Northwest Regional Educational Laboratory's review of the Mat-Sea-Cal Oral Proficiency Tests in their publication, Oral Language Tests for Bilingual Students. That publication was released in July, 1976 as a guide to administrators and program coordinators in the selection of instruments for assesing students' language dominance and oral proficiency(-ies).

The N.W.R.E.L. reviewers, in rating each instrument, explored four criteria: Measurement Validity, Examinee Appropriateness, Technical Excellence, and Administrative Usability. Several questions within each criteria were examined in determining the overall criteria rating.

A descriptive review of the Mat-Sea-Cal is presented on pages 101-06 of the publication. The reviewers' rating is summarized in a chart on pages 126 - 27 of the booklet.

This critique scrutinizes the evaluations rendered to the Mat-Sea-Cal by the reviewers in each of the four criteria. Discussion is offered on several points. Differences in perception between this author and the N.W.R.E.L. reviewers on the evaluation of the Mat-Sea-Cal are enumerated.

In citing references, where only page numbers are given, the statement is attributed to the N.W.R.E.L. publication. Where outside sources are quoted, the standard author-title-page number format is followed.

THE CRITIQUE

This critique follows the outline presented in the N.W.R.E.L. booklet in chapter three: Evaluative Criteria. This discussion, therefore, begins with Measurement Validity, followed by Examinee Appropriateness, Technical Excellence, and Administrative Usability -- in that order.

CRITERION: MEASUREMENT VALIDITY

Seven questions (pp. 30-2) were considered in determining an instrument's measurement validity. However, ratings were given for only two categories (pp. 126 - 27). Judging from the evaluation chart point scale (p. 126), questions #a through #e (pp. 30-2) appear to have been combined into the category "content and construct" (p. 126). The second category within this criterion, "concurrent and predictive", apparently is composed of questions #f and #g (p. 32).

Of note, the authors of the N.W.R.E.L. publication have presented no rationale within the text for condensing seven discrete questions into two evaluation categories.

Content and Construct. In this category the Mat-Sea-Cal received five of a possible seven points. This was the highest rating achieved by any of the eleven instruments reviewed. In fact, the Mat-Sea-Cal was the only instrument to be awarded over half of the maximum possible points alloted for content and construct validity (p. 126).

This rating is significant in psychometric terms. Content and construct validity are deemed as the initial, critical stages of instrument

development. Content validity addresses the issue of whether an instrument samples the universe it purports to measure. Construct validity explores the question of whether performance on the sampled items reflects an accurate measure of the respondee's knowledge of, or competence in, the theoretical constructs being tested (Cronbach, in E. L. Thorndike, Educational Measurement, p. 446).

As the initial phase of instrument development, demonstration of content and construct validity is, therefore, a pre-requisite to conducting further analyses. That the Mat-Sea-Cal was reviewed favorably by an independent agency, N.W.R.E.L., is a manifestation of the instrument's quality.

Concurrent and Predictive. The Mat-Sea-Cal was awarded no points for concurrent and predictive validity in the N.W.R.E.L. publication (p. 126). Concurrent and predictive validity are both correlation analyses between the instrument under development and a measure of criterion proficiency.

Specifically, concurrent validity entails the correlation of the test and the outcome data at very nearly the same time. The correlation between the two is the measure of concurrent validity. Emphasis is placed on selecting an appropriate outcome measure, as the new instrument will be correlated with "whatever the outcome measure tests." (Cronbach, Essentials of Psychological Testing, pp. 104, 108, and 117; see also Thorndike, p. 484).

The primary interest of predictive validity is finding an accurate measure of a future outcome. A score on the instrument under development is checked against a criterion measure. The aim of testing is to predict this criterion, and the merit of the instrument is judged by the accuracy

of its prediction. Normally, several months time elapse between testing and data gathering on the criterion measure. Success in predicting the criterion ususally results in a statistical decision theory model. That is, a formula is used in assigning students to different treatments in the future, based on the predictive validity findings. Here too, the practical "worth" of the decision formulas rests on the selection of a valid criterion measure of the competence or performance on which prediction is desired (Cronbach, pp. 108 and 117; see also Thorndike, pp. 303, 443 + 44, and 484).

Related to concurrent validity the N.W.R.E.L. review noted that significant correlations (at the .01 level) were established between the Mat-Sea-Cal and the S.R.A. Achievement Survey (pp. 104 - 05). These findings were originally published in Dr. Matluck's A.E.R.A. paper in April, 1976 (Matluck and Matluck, The Mat-Sea-Cal Instruments for Assessing Language Proficiency, p. 10). Inspection of Dr. Matluck's sources reveals that these correlations were largely between .30 and .70. Based on the N.W.R.E.L. rating scale (p. 32), the Mat-Sea-Cal would, therefore, deserve one point for concurrent validity.

Since no points were awarded (p. 126), one must conclude that the reviewers did not pursue Dr. Matluck's original sources. As his A.E.R.A. paper was delivered in early April and the N.W.R.E.L. publication not released until July, ample time for checking sources existed. In the light of such shortcomings, doubt must be raised with the reviewers stated intent that, "an effort was made to obtain as much descriptive material as possible on each instrument" (p. 44).

Regarding predictive validity, the Mat-Sea-Cal was again awarded no points by the reviewers. To date, no a-priori predictive investigations with the instrument have been undertaken. Thus, this rating appears justified.

However, discussion may precede from whether the demonstration of predictive validity is within the scope of field test instruments. Predictive validity is usually the final hurdle of instrument development. It is undertaken after other investigations (e.g., reliability, item analysis, concurrent validity, test revision, etc.) have proven successful. In fact, demonstration of predictive validity indicates that an instrument is ready for commercial distribution.

Summary (Measurement Validity). The Mat-Sea-Cal Tests were awarded five of eleven points for measurement validity. The reviewers thereby classified the Tests as "poor" on this criterion.

Two issues belie this rating. First, in the concurrent validity section the reviewers failed to pursue source documents. As a result, they overlooked significant correlations which would have entitled the Mat-Sea-Cal to an additional rating point.

The second question pertains to whether field-test instruments should be rated on predictive validity in a manner similar to commercial tests. Demonstrating predictive validity would appear to be more the domain of commercially marketed measures.

In sum, the Mat-Sea-Cal should be credited with six rating points for measurement validity. Then, employing the nine or eleven point scale (excluding or including the predictive validity rating), the

instrument should be reclassified as "fair" on this criterion, or "good" on a nine-point scale..

CRITERION: EXAMINEE APPROPRIATENESS

Thirteen questions (pp. 33-6) were considered in determining the examinee appropriateness of an instrument. Points were awarded for twelve of the considerations. No rating (only a description) was given for "mode of examinee response," and no reasons were tendered for this exception.

Like the measurement validity section, some questions were combined into single rating categories (of which there are nine). Questions #b and #c (pp. 33 - 4) form the category "item relevance" (p. 126). Questions #d, #e, #f, and #g make up the category "instructions." The remaining questions are retained as individual evaluation entities. No discussion is provided as to how and why some queries were aggregated, while others were retained as individual items.

As with the previous criterion (measurement validity), the maximum points awarded per category in examinee appropriateness vary. The point scale ranges from zero to four in some instances, while a zero-one alternative is the choice in others. Are the concerns rated in one category four times as important as those rated in another? The reviewers provide no enlightenment.

Summary (Examinee Appropriateness). The Mat-Sea-Cal Tests received fourteen of fifteen possible points on the criterion of examinee appropriateness. This rating earned the instrument a classification of "good" for this criterion.

Only in failing to require test administrators to inform examinees of the test's purposes (i.e., "justification") did the Mat-Sea-Cal not receive the maximum points in any category. This rating preference assumes that primary-age children's performance is positively affected by knowing why they are being tested. In actual settings this information most likely motivates some youngsters, while creating anxiety in others.

CRITERION: TECHNICAL EXCELLENCE

Four questions (pp. 37 - 8) were evaluated in determining an instrument's technical excellence. Unlike each of the previous two criteria (validity and appropriateness), each question was rated as a separate entity. However, the maximum point value within each category varied: from one to three points. Again, these scale differentials remain unexplained.

Several points related to the evaluation of technical excellence need to be made. First, the category of "replicability" appears to be an administrative matter (the next criterion) rather than a technical concern.

Of a more serious nature is the reviewers' collective knowledge of the concept of reliability. Their bias favors instruments capable of simultaneously exhibiting three types of reliability: alternate form, test-retest, and internal consistency. In doing so, the N.W.R.E.L. reviewers failed to address whether each reliability type was congruent in nature to that of the instruments being evaluated. Also, under certain circumstances, one type of reliability computation yields coefficients

virtually identical to those calculated by a second method (Guilford and Fruchter, Fundamental Statistics in Education and Psychology, p. 410). As another example, the creation of an alternative form for the purpose of presenting a second reliability coefficient is not a practice advocated by educational psychometricians (Stanley, in Thorndike, pp. 404 + 5). Such state-of-the-art positions, however, have been ignored by the reviewers in listing their reliability ratings for instruments.

Alternate form. The Mat-Sea-Cal was awarded no points for alternate form reliability. As only one form of the instrument (per language) exists, this finding was expected.

However, with power tests alternate form and internal consistency estimates of reliability "can be used almost interchangably" (Guilford and Fruchter, p. 410). Power tests are those in which examinees have ample time to answer all questions. (Standard educational measurement texts list specific requirements - - Thorndike, p. 192; see also Guilford and Fruchter, pp. 406 - 07.) By instruction and as demonstrated in actual administration, the Mat-Sea-Cal permits sufficient time for all examinee responses. Thus, the rating of the Mat-Sea-Cal for alternate forms duplicates the evaluation for internal consistency (which is detailed later).

Furthermore, construction of an alternate form solely to demonstrate a second reliability coefficient would be an unnecessary depletion of test development resources. Creation of a second form places additional requirements on test development.

First, the amount of variation in content and format between forms must be skillfully balanced, if they are to be truely comparable. If the alternate forms differ too distinctly, the correlation between them will underestimate the desired reliability. By contrast, if the two forms overlap to an excess, the obtained correlation will overestimate the forms reliability (Stanley, in Thorndike, pp. 404 - 05).

In addition, care must be taken to insure that items selected for the "second" form are representative of the respective universe. Further, the manner in which items are chosen for inclusion in the "alternate" form must be equivalent, and not reflect increased skill in item writing. Violation of either requirement would have a deleterious-effect on the instrument's content and construct validity (Ibid.).

Finally, item statistics and correlations would need to be comparable, Therefore,

> "if only a single form of a test is needed for the research or practical use to which the test, is to be put, it seems unduly burdensome to prepare two separate tests in order to obtain an estimate of reliability" (Ibid., pp. 405 - 08).

Another type of alternate form reliability is the "instant readministration", or split-half technique. Here, the instrument is divided into halves (randomly, or in-pattern), administered once; but a second administration is considered to have occurred "instantaneously." The two halves are separately scored, then correlated. However, as reliability is a function of test length (to a point), the obtained coefficient is spuriously low. It is, therefore, adjusted by the Spearman-Brown estimation formula. In addition to being an estimate of an underestimate, the split-half coefficient is regarded as a one-form reliability correlation (Ibid., p. 369).

In sum, an evaluation of the Mat-Sea-Cal for alternate form reliability appears unnecessary. The instrument exhibits high standards for reliability on an internal consistency measure (as will be documented shortly). The instrument is a power test, there its internal consistency coefficients would be comparable to alternate form computations. Thus, though the instrument in its present form cannot show alternate form coefficients, this cannot be deemed detrimental to its overall psychometric quality.

Test-Retest. The Mat-Sea-Cal received no points for test-retest reliability from the reviewers. No test-retest studies with the instrument have been conducted, to date; thus, the rating is as expected.

However, it should be noted that the test-retest technique is not readily applicable to the Mat-Sea-Cal. The test-retest technique indicates the stability of examinee responses over time. High test-retest coefficients are associated with the rank ordering of examinee scores on the tested constructs remaining fairly constant.

Obtaining a retest coefficient with a one form instrument that measures oral proficiency would be difficult. If several months (or even weeks) elapsed between the two administrations, score differences are likely to be confounded by the effects that schooling and individual maturational patterns have on children. In essence, the reliability coefficient would be affected to an unknown degree by factors beyond the testing situation. (Ibid., p. 407; see also Guilford and Fruchter, pp. 407 - 08).

On the other hand, allowing only a short interval between administrations (e.g., a few days) introduces a memory effect. The examinees are likely to

recall specific questions and their responses to them. In such instances, it is recommended that the test-retest procedure be avoided (Thorndike, pp. 407 + 08).

Thus, for technical reasons the Mat-Sea-Cal's reliability should not be computed by the test-retest method. The respondees' performance on the instrument would not likely remain static over a long interval. By contrast, a short test-retest cycle introduces a memory effect.

Internal Consistency. The Mat-Sea-Cal received zero points for an internal consistency rating on the N.W.R.E.L. evaluation chart (p. 127). Internal consistency was to be demonstrated by either a split-half technique or by a Kuder-Richardson formula coefficient (p. 37).

Interestingly, the N.W.R.E.L. description of the Mat-Sea-Cal (p. 105) lists Kuder-Richardson coefficients of .94 and .91 for the English and the Spanish tests, respectively. Thus, the reviewers offer a new and intriguing evaluation system! They describe necessary criteria (p. 37), report coefficients meeting the criteria (p. 105), yet refuse to award the rating points (p. 127)?!

Such minor oversights reach an unpalatable level when N.W.R.E.L. solicits U.S.O.E. endorsements, to the effect that,

> "Administrators, teachers, and other school personnel involved in planning bilingual/bicultural programs. . . will find this document an invaluable aid. . . in providing objective, comprehensive evaluation of these tests in order to facilitate the selection of appropriate measurement instruments" (pp. 5 - 6).

Practicing educators are rarely trained linguists or psychometricians. They are apt to rely on organizations such as N.W.R.E.L. and U.S.O.E. for up-to-date, factual information on technical matters. Misinformation,

such as the above, does little to enhance the quality of technical input on which educational decisions are often based.

Correcting the oversight in the evaluation chart (p. 127) would credit the Mat-Sea-Cal with two points for internal consistency (the maximum allowed within this category). For reasons unstated, the reviewers permit a maximum of three points for alternate form, or test-retest reliability, while two is the maximum for internal consistency.

Replicability. The reviewers gave the Mat-Sea-Cal no points in this category. Replicability dealt with whether testing procedures outlined in the administrator's manual could be duplicated in other situations. Two items deserve mention in critiquing this category.

First, as defined above, replicability is an administrative consideration, not a statistical/technical matter. The major portion of this, the technical excellence, criterion dealt with measurement (specifically, reliability). In fact, nine of the ten evaluation points in the criterion were reserved for reliability. Furthermore, replicability appears more congruent to the next criterion, administrative usability.

Second, items evaluated as replicability (pp. 37 - 8) are rated throughout the administrative usability section. For example, "administrative details" are evaluated in questions #a through #c (of administrative usability). "Scoring" is rated in items #d through #f. "Interpretability" is the focus of concerns #h, #i, and #n. "Standardization" is reviewed in items #k and #j.

In short, replicability is both misplaced, and a duplication of the ratings in other sections.

Summary (Technical Excellence). On technical excellence the N.W.R.E.L. reviewers rated the Mat-Sea-Cal as "poor". This rating was a product of the reviewers marginal expertise of the concept of reliability, and an omission related to internal consistency. As a result, the rating as "poor" on technical excellence is unsubstantiated.

The Mat-Sea-Cal Tests demonstrated high internal consistency coefficients. A discussion as to the comparability of these coefficients to alternate form coefficients was provided. Similarly, the inappropriateness of the test-retest method for an instrument such as the Mat-Sea-Cal was presented.

Finally, doubt was cast as to whether replicability belonged with technical excellence or the administrative usability criterion. Further, it was noted that considerations within the replicability category were rated elsewhere.

In conclusion, the Mat-Sea-Cal has demonstrated high internal consistency coefficients as measures of reliability. The instrument has, thus, met the major concern of the technical excellence criterion. Therefore, a rating of "good", not "poor", is justified.

CRITERION: ADMINISTRATIVE USABILITY

Fourteen questions were considered in determining an instrument's administrative usability. Each question was evaluated independently, and points were awarded in fourteen separate categories. The point scale ranged from zero-two with four of the considerations, zero-one for the remaining ten questions. Again, no discussion was provided for the difference in scaling range.

The Mat-Sea-Cal was awarded the maximum point value on four items: training of administrator, number of administrators, range of the test, and diversity of skills measured. As the maximum point value on these items was achieved, no additional discussion of them is provided here. Instead, comments are directed toward the remaining ten considerations.

Clarity of Manual. The Mat-Sea-Cal was awarded no points in this category. Aspects considered in evaluating test manuals included:

> "discussion of purpose, uses, and limitations of the test; clear administering and scoring directions; and description of test development and validation" (p. 38).

The Mat-Sea-Cal Test administrator's manual reads explicit in regards to purpose, uses, limitations, and directions (Matluck and Matluck, Mat-Sea-Cal Oral Proficiency Tests (Field Test Edition), pp. 2-8). The manual does not provide a full description of the test's development and validation.

As a field-test instrument, development and validation of the Mat-Sea-Cal is not complete. Therefore, the question arises as to whether the in-progress data should be reported in the manual; and if so, how often the manual should be updated. The alternative view holds that preliminary information may be misleading, and be proven partially inaccurate when the validation process is completed. This alternative view would hold for the completion of the development/validation processes, when data would be supplied in their entirety.

The N.W.R.E.L. evaluators prefer the former course, employing a zero-one scale. Thus, they were prevented from awarding points to the Mat-Sea-Cal for the information that is contained in its manual.

Had the reviewers selected a broader point scale, a more informative comparison of instruments would have resulted. Ten of the eleven tests reviewed in the N.W.R.E.L. publication received no points in this category (p. 127).

Also, items covered in the evaluation of test manuals are further scrutinized elsewhere in the review scheme. For example, test development includes item selection methods (questions #a and #b in measurement validity). Instrument validation encompasses concurrent and predictive validity studies (questions #f and #g in measurement validity). Scoring processes are also reviewed in other categories of administrative usability (questions #d and #h).

Scoring. The Mat-Sea-Cal received one of two possible points for ease and objectivity in scoring. The second and third sections of the test do require the test administrator to listen to, or observe, examinee responses. The N.W.R.E.L. reviewers felt that such tasks involve a degree of difficulty which detracts from the scoring ease. The reviewers favored templates and stencils as methods of scoring and conversion.

However, the topic may be broached as to whether quality in scoring is necessarily a function of templates, and the like. Providing more detailed examples of correct and incorrect responses would alleviate doubts related to the objectivity of the Mat-Sea-Cal's scoring process, though.

In addition, scoring is rated in questions #b (p. 38) and #h (p. 40) of administrative usability, and is incorporated into question #d (pp. 37 - 8) of technical excellence.

Training. The Mat-Sea-Cal failed to receive points for the question of who may interpret test scores. Consideration was given as to the extent special training was required for accurately interpreting test scores. Value was placed on regular teaching staff being able to do the interpretation.

Several questions may be posed as to the merit in this judgment. For example, is the typical classroom teacher adequately prepared to determine language dominance or evaluate oral proficiency? Are such determinations always within the grasp of simple score conversions? Is a specially qualified test score interpreter intrinsically less desirable than a classroom teacher manipulating a formula or a template? (Or, how accurately can a typical classroom teacher manipulate a template or formula?)

The answers to such questions depend on the constructs being tested, the depth to which traits are measured, and the implications of decision making based on data interpretation. Language dominance and oral proficiency (the Mat-Sea-Cal's domain) can become intricate issues that require sophisticated interpretation. Decisions based on such data interpretation will affect the learning activities offered to individual students. Furthermore, misinterpretation of test data, in addition to being deleterious to the students, could lead to very nasty legal complications.

By comparison, surveying the extent of home language usage in a community would be a simpler matter. Survey data can be gathered and tabulated, and some questions on community language usage answered. However, conducting surveys also requires considerable expertise, specifically to insure that data are gathered in an accurate and an objective manner.

This does not suggest that one purpose is inherently more valuable than another. It points out, though, that purposes will differ; and that as they differ, so will the means of language assessment and data interpretation. In short, the complexities of language assessment and interpretation are dictated by the depth of information required to meet stated purposes.

However, evaluating who can interpret scores was not viewed in terms of the complexities of the constructs measured. Nor was interpretation considered in respect to the implications that certain test-based decisions might have.

Score Conversion. The Mat-Sea-Cal received one of two points for clarity and simplicity in the conversion of raw scores to interpreted scores. Refinement of the instrument's score conversion techniques would be a desired product of the validation process.

The concerns rated in this section were also examined in questions #b (p. 38) and #d (p. 39).

Interpretation. The Mat-Sea-Cal was recipient of no points for ease of interpretation of test scores. Value was placed on scores that yielded binary judgments, grade equivalents, percentiles, and the like.

Arguments, here, with the review would be similar to those made in the section on training. Not all considerations in the linguistics field reduce to binary, yes-no conclusions (proficiency and dominance, as examples). Ease of interpretation needs review in terms of constructs and depths measured.

Norm data (grade equivalents and percentiles) are outcomes of a completed validation process. For the Mat-Sea-Cal, sample-specific norms were referenced in Dr. Matluck's A.E.R.A. paper, and its source documents.

Overlap of evaluation topic is also evident between this category and items #d (p. 39) and #e (p. 40).

Validating Group. The Mat-Sea-Cal, like every test rated, failed to receive points for representativeness of the validation group (p. 127). Five concerns (p. 41) were examined in determining representativeness.

The N.W.R.E.L. reviewers claimed that neither the sample sizes nor their characteristics were reported in Mat-Sea-Cal studies to date. In fact, a thorough inspection of documents cited in Dr. Matluck's A.E.R.A. paper would refute the reviewer's claim. Those sources enumerated upon sample sizes, geographical representation, and population characteristics of the examinees. Data analyses employed by those studies also examined the sample groups for the effects of such qualitative variates.

Thus, exception may be taken with the evaluation of the Mat-Sea-Cal in this category.

Racial, Ethnic, and Sex Representation. The Mat-Sea-Cal failed to receive either of two points in this category. Four considerations (p. 41) were used to determine the rating.

Examination of sources quoted by Dr. Matluck's paper would, again, refute the N.W.R.E.L. rating. Ethnic, racial, and sex characteristics have been detailed in all Mat-Sea-Cal studies to date.

Thus, the instrument is deserving of a rating of two points within this category.

Can Decisions Be Made. The Mat-Sea-Cal was awarded one of two points on the issue of whether test data was useful in making decisions concerning individual examinees. The reviewers presented examples of statements (p. 42) which they considered to be evidence of decision making prowess. The inclusion of similar statements in test manuals resulted in favorable ratings.

The reviewers did not specify the extent to which 'decision statements' had been verified by support data (p. 42). Generating decision statements from test scores strongly implies the presence of predittive validity. As such decisions affect the educational opportunities offered to learners, confidence in pursuing the recommended decisions must result from empirical evidence.

Without requiring support evidence, the reviewers would be encouraging unsubstantiated hypothesizing in test manuals. On the other hand, rating predicitive validity (in this category) again raises the question as to what psychometric extent field test and commercial instruments may be permitted to differ. The expectation is that commercial measures exhibit more concrete evidence. But is it reasonable to evaluate field test instruments (and give them lower ratings) employing the standards used to judge commercial measures?

In any case, predictive type concerns have been previously rated elsewhere in the review schematic.

Alternate Forms. This category is an extension of the reviewers' bias in favoring tests for which alternate forms have been developed. Comments offered in the technical excellence section (referring to alternate form and test-retest reliability) would apply equally as well here.

Form Comparability. This section also extends the reviewers' preference for tests having at least two forms. No provision is made in the rating scale for one form instruments, except that they awarded no points (i.e., their overall administrative usability rating is lowered).

Summary: Administrative Usability. The Mat-Sea-Cal received seven of eighteen (possible) points for administrative usability. This resulted in the reviewers classifying the instrument as "poor" on this criterion.

This evaluation is questionable on three counts. First, the reviewers failed to pursue source documents in obtaining information related to certain rating categories. This oversight denied the instrument rating points, and a more accurate and favorable evaluation in those categories. Second, the reviewers essentially rated the same topics over, and over again (alternate forms, being the most glaring example). Third, the reviewers insist on evaluating field test instruments with the same standards used for commercial measures. This persistence prevents an evaluation of field test measures relative to the stage of test development at which they are at.

In summary, the Mat-Sea-Cal's rating on administrative usability may be challenged. Given its present status in the test development process, it is deserving of at least a rating of "fair".

SYNOPSIS

This paper critiqued the Northwest Lab's review of the Mat-Sea-Cal Tests, which was presented in the Lab's publication, Oral Language Tests for Bilingual Students (1976). Several points of difference between the Lab's reviewers and this author were noted.

Specific differences regarding the Mat-Sea-Cal's overall rating on the four evaluation criteria were as follows:

Criterion	N.W.R.E.L. Rating	Critique's Rating
1. Measurement Validity	poor	fair
2. Examinee Appropriateness	good	good
3. Technical Excellence	poor	good
4. Administrative Usability	poor	fair-good

Questions were also raised concerning N.W.R.E.L.'s repeated evaluation of certain items (i.e., alternate forms, scoring, administration, instructions, etc.). Further, the point value range within evaluation categories varied considerably: zero-four, zero-two; zero-one. The reviewers never presented a justification or a discussion of these range differences.

On the technical side, concern was expressed for the reviewers' collective knowledge of the concept of reliability. Also, the reviewers did not extend a concerted effort to obtain informative source documents.

In conclusion, two recommendations must be made on the basis of this critique. First, contrary to the U.S.O.E. endorsement (pp. 5 - 6), the N.W.R.E.L. publication is not recommended for educators' use in

selecting oral proficiency measures. The booklet contains too many omissions and misinterpretations.

Second, an attempt is needed to inform educators of the shortcomings contained in the N.W.R.E.L. publication. This is necessary so that the development of the Mat-Sea-Cal Tests, with cooperation from school districts, will not be hindered by the statements made in the N.W.R.E.L. review.