DOCUMENT RESUME

ED 137 320 TM 006 082

AUTHOR Doyle, Vincent
TITLE A Psychometric Analysis of the Mat-Sea-Cal Oral Proficiency Tests.
INSTITUTION Center for Applied Linguistics, Arlington, Va.
PUB DATE [Apr 77]
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Elementary Education; *English; *Factor Analysis; *Item Analysis; Language Proficiency; *Language Tests; *Spanish; *Test Reliability
IDENTIFIERS *Mat Sea Cal Oral Proficiency Tests

ABSTRACT

The Mat-Sea-Cal Oral Proficiency Tests are a series of comparable grammatical structure tests. They have been developed in six languages: English, Spanish, Cantonese, Mandarin, Ilokano and Tagalog. Their purpose is to identify linguistic skills and deficiencies of primary school children grades K through 4. This research reported on the psychometric qualities of the English and Spanish editions. Reliability was computed by the method of internal equivalence. Coefficients were .91 on the English test and .94 on the Spanish test. Point biserial coefficients were calculated as the discrimination index. Results varied by subtest (Listening Comprehension, Sentence Repetition, and Structured Response). Factor analysis, via principal factoring with varimax rotation, was employed to identify item pools. Results indicated that approximately 30 percent of all original items require revision. (These tests are labeled "Field Test Edition" by the authors.) The remaining items possessed good to excellent discrimination indices, and difficulty levels appropriate for criterion referenced measures. (Author)

A PSYCHOMETRIC ANALYSIS OF THE MAT-SEA-CAL ORAL PROFICIENCY TESTS

Vincent Doyle
(Center for Applied Linguistics, Arlington, Virginia)

Paper read at the American Educational Research Association Convention

April, 1977

session number: 8.06

The research reported herein has been conducted under the auspices of the Center for Applied Linguistics, Arlington, Virginia. The author is deeply indebted to that organization, and especially its director, Dr. Rudolph C. Troike, for continued support and encouragement in this endeavor.

Abstract

The Mat-Sea-Cal Oral Proficiency Tests are a series of comparable grammatical structure tests. They have been developed in six languages: English, Spanish, Cantonese, Mandarin, Ilokano and Tagalog. Their prupose is to identify linguistic skills and deficiencies of primary school children grades K through 4. This research reported on the psychometric qualities of the English and Spanish editions.

Reliability was computed by the method of internal equivalence. Coefficients were .91 on the English test and .94 on the Spanish test. Point biserial coefficients were calculated as the discrimination index. Results varied by subtest (Listening Comprehension, Sentence Repetition, and Structured Response). Factor analysis, via principal factoring with varimax rotation, was employed to identify item pools.

Results indicated that approximately 30 percent of all original items require revision. (These tests are labeled "Field Test Edition" by the authors.) The remaining items possessed good to excellent discrimination indicies, and difficulty levels appropriate for criterion referenced measures.

LANGUAGE ASSESSMENT AND THE MAT-SEA-CAL TESTS

Language assessment might be defined as the systematic attempt to ascertain one's ability to effectively receive and produce the elements of verbal expression. Linguistic appraisal measures the ability to synthesize proficiencies in phonology, morphology, syntax and lexicon into a meaningful Gestalt.

A major objective of primary education is facilitating the acquisition of this linguistic integrated configuration. Indeed, fostering a unified constellation communication proficiencies unlocks educational opportunities for the student. Most educational experiences require that by approximately fourth grade children are basically functional in communicative arts, of which speech is pivotal.[1] The adequacy of this assumption is challenged, yearly it seems, by U.S.O.E. statistics relative to the number of functionally illiterate adults in the population, the decline of achievement scores, the percentage of pupils failing basic skills' tests, and the like.

In the classroom the problem has been the identification of the children's initial language range. This knowledge is essential in order to determine the basis from which to commence instruction. Factors beyond the control of the school have influenced initial language development, among which are:

1. The nature of the child's pre-school linguistic environment.
2. Parental personality traits and attitudes.
3. Degree of association with adults.
4. Child rearing practices in the home.
5. Number of siblings and orderal rank among them.
6. Parental attitude toward their own speech community and toward second language group(s).[2]

The complexity of personal and environmental inputs presents the challenge to educators in rendering equal opportunity to pupils. "Merely providing the same educational opportunity to all students, does not satisfy the law when some students are effectively foreclosed from any meaningful education by a language barrier."[3] Ascertaining the extent and degrees of linguistic diversities, and then accomodating rich varieties of background into a series of organized learning activities constitute both the science and art of teaching.

Linguistic pluralism is common to the United States. Approximately 10 million predominantly Spanish speaking individuals reside in this country. The city of Seattle, as an example, is home to dozens of major dialects within its communities.

The humanist educator would assert that schooling should augment cognitive/affective growth regardless of language heritage. The non-native English speaking student offers the rich potential of first-hand sharing of culture and language. Even deficiencies of the native English speaker require special attention, so as to permit all children to be inundated with schooling's benefits.

Generically, the Mat-Sea-Cal Oral Proficiency Tests are a systematic, objective vehicle for determining aural-oral competencies. They are a series of six comparable grammatical structure tests (in Cantonese, English, Ilokano, Mandarin, Spanish, and Tagalog).

Authored by Drs. Betty and Joe Matluck, with the support of the Center for Applied Linguistics, these instruments were designed to

1. determine the child's ability to
   a. understand and produce the distinctive characteristics of the spoken language,
   b. express known cognitive concepts in the language, and
   c. handle learning tasks in the language;

2. provide placement and instructional recommendations with respect to alternate programs, such as English instruction and bilingual education.[4]

Subject matter encompasses eight concept groups: the skills of identifying, classifying, quantifying, interrogating, and negating, and of showing relationships such as spatial, case, and temporal. These concepts are assessed through three types of communication manifestations, or modes: listening comprehension, sentence repetition, and structured response. Eighty-one sentences comprise each language test. Administration is through a standardized, taped stimulus and a series of supporting visuals. Administration time averages 25-40 minutes, though no time restriction is placed on respondees. Testing and scoring instructions are fully described; samples of correct responses are provided with each item. In section one the examinee selects from one of three pictures. Sections two and three require verbal response, and both sections are completely taped. Item-types are multiple choice and short answer.

Initial field-testing was conducted in Seattle. Subsequent administrations have been completed in school districts in the states of California, Idaho, Texas and Washington.

The Mat-Sea-Cal Tests are in field test form, and are clearly marked as such. Their development has followed a standard, psychometric process outlined for instruments of the type.[5] This paper focuses on statistical qualities of the English and Spanish tests, as demonstrated through field-testing to date.

RELIABILITY

The first desirable characteristic of any instrument is the ability to demonstrate consistency in measurement over a series of administrations or individuals. Known as reliability, it depicts the degree of certainty to which one may base decisions

on gathered data. Without reasonably high reliability (.80 and up), the constructs under investigation have only tentatively been measured by the given item sequence.

Mat-Sea-Cal test items are dichotomously scored, i.e., correct or incorrect. Summation of item response-scores (of which there are 108 on the English test, and 118 on the Spanish counterpart) are converted to percentage scale, with normal score distribution: 0 to 100. Thus, an appropriate method for calculating reliability is that of internal equivalence, mathematically stated by the Kuder-Richardson formula.

This method was selected over other common procedures (such as alternate form or test-retest) for following reasons. Only one form per Mat-Sea-Cal test exists; thus, logistics precluded alternate form calculations. Being a power test, alternate form and interval equivalence coefficients would be nearly identical. Furthermore, psychometric theory does not endorse creation of second forms when only one is necessary for research or practical use.[6]

The test-retest approach also appeared less desirable for reliability computations. Retest coefficients require several months between administrations of the instrument. In this instance reliability would be affected to an unknown extent by a combination of schooling and maturational factors on oral proficiency development. By contrast, a short retest interval employed with a one-form instrument introduces a "memory" effect to examinees' performance on the readministration.

The overall English test reliability coefficient was computed to be .91, on the Spanish test .94. Calculations on subsamples (divided by categories within ethnicity, sex, geography, and educational attainment) ranged from .82 to .96. Sample size for the overall coefficients was 3000.

Three specific conclusions may be drawn from these findings. First, the English and Spanish versions of the test are sufficiently reliable to permit further development and refinement on the present sample of items. In other words, the state of measurement consistency is such that a complete rewriting of test items is unnecessary. Second, sufficient confidence may be placed in data generated by the Mat-Sea-Cal that other avenues of linguistic research may be supported by its information base. Third, the Mat-Sea-Cal Tests appear to be relatively homogeneous in nature. Reliability coefficients, for whole tests, in excess of .90 usually are an indication of homogeneity.

ITEM DIFFICULTY INDEX

Item difficulty is a descriptive statistic measuring the ease (or difficulty) examinees' experience in correctly responding to the individual items. The acceptable level for item difficulties is hinged upon the basic purposes for testing, as specified in the test's blueprint.

For the Mat-Sea-Cal information on respondents was desired in an area approaching "minimum oral proficiency" (operationally defined as 70 percent performance on the instrument[8]). Further, the tests were to be criterion referenced, used as a basis for diagnosis and remediation. Thus, items with difficulty indicies between 50 and 90 percent appeared to be most appropriate. This would permit respondents to exhibit both the strengths (i.e. through the easier items) and weaknesses (i.e., through the more difficult questions) in their language patterns. By concentrating the given number of items within the restricted range, a more reliable portrait of aural-oral abilities is obtained.

Tables one and two present, by communication mode, the difficulty indicies for indicies for items of the English and Spanish Tests, respectively.

(Note: the following symbols are used: LC - Listening Comprehension; SRep - Sentence Repetition; and SR - Structured Response. Numerals followed by an "a" or "b" indicate a response pair, one item measuring two language manifestations.) Perusal of the tables would indicate the following items need to be scrutinized critically:

English test:

Listening comprehension: 1,2,3,5a,5b,6,7,12,13a, 13b,14,18a thru 23b,25a,25b

Sentence repetition: 10a,10b,11,14a,14b,17a,17b,19a, 20,22a,22b,23a,25

Structured response: 7,10,11,13,19

Spanish test:

Listening comprehension: 1,2,3,6,8,12,16

Sentence repetition: 4,5a,5b,7,14a,16b,21a,22b,25,26a

Structured response: 10,15,21,23,26

Table #1

English Test

Item #		% responding correctly, P(i), (over all grade levels)	Point-biserial correlation, R(i) (over all grade levels)
LC	1	96.	.29
	2	46.	.22
	3	98.	.12
	4	84.	.45
	5a	96.	.38
	5b	96.	.37
	6	99.	.27
	7	97.	.35
	8	74.	.28
	9a	87.	.40
	9b	87.	.41
	10	83.	.35
	11	89.	.43
	12	99.	.21
	13a	96.	.42
	13b	97.	.39
	14	96.	.25
	15	88.	.31

English Test:

Item #	% responding correctly, P(i), (over all grade levels)	Point-biserial correlation, R(i) (over all grade levels)
LC 16	94.	.30
17	95.	.56
18a	98.	.12
18b	98.	.11
19a	99.	.16
19b	99.	.17
20a	99.	.17
20b	99.	.18
21a	99.	.25
21b	99.	.25
22a	99.	.19
22b	99.	.19
23a	96.	.18
23b	96.	.18
24a	94.	.22
24b	94.	21
25a	97.	.21
25b	97.	.21
26a	94.	.21
26b	94.	.20
27a	82.	.28
27b	83.	.25
3Rep 1	82.	.49
2a	82.	.49
2b	83.	.52
3	93.	.54
4	95.	.51
5	91.	.54
6	93.	.55
7a	84.	.49
7b	84.	.52
8	90.	.53
8	95.	.58
9b	91.	.51
9c	80.	.49
10a	96.	.51
10b	96.	.46
11	97.	.49
12	87.	.29
13a	93.	.50
13b	93.	.56
14a	96.	.53
14b	97.	.45
15a	93.	.44
15b	90.	.39

English Test:

Item I		% responding correctly, P(i), (over all grade levels)	Point-biserial correlation, R(i) (over all grade levels)
SRep	16	89.	.61
	17a	95.	.56
	17b	97.	.49
	18	94.	.49
	19a	95.	.60
	19b	94.	.60
	20	95.	.51
	21	91.	.56
	22a	96.	.50
	22b	96.	.58
	23a	97.	.52
	23b	94.	.54
	24	75.	.47
	25	97.	.49
	26	87.	.58
SR 1	1	92.	.15
	2	98.	.50
	3	91.	.37
	4	92.	.40
	5	95.	.38
	6	90.	.43
	7	97.	.44
	8	94.	.45
	9	90.	.35
	10	45.	.25
	11	96.	.47
	12	78.	.49
	13	39.	.32
	14	61.	.43
	15	73.	.27
	16	62.	.38
	17	90.	.37
	18	55.	.32
	19	35.	.28
	20	83.	.33
	21	91.	.53
	22	82.	.42
	23	79.	.25
	24	59.	.26
	25	59.	.41
	26	61.	.41
	27a	91.	.24
	27b	91.	.24
	28a	88.	.30
	28b	88.	.29

Spanish Test

	Item #	% responding correctly, P(i), (over all grade levels)	Point-biserial correlation, R(i) (over all grade levels)
LC	1	93.	.26
	2	43.	.08
	3	95.	.42
	4	71.	.39
	5a	89.	.32
	5b	89.	.34
	6	95.	.34
	7	87.	.45
	8	41.	.15
	9a	73.	.16
	9b	62.	.17
	10	68.	.32
	11a	83.	.29
	11b	84.	.32
	12	98.	.33
	13a	89.	.50
	13b	89.	.48
	14	87.	.24
	15	86.	.40
	16	38.	.12
	17	85.	.38
	18a	92.	.38
	18b	92.	.38
	19a	92.	.57
	19b	92.	.57
	20a	88.	.48
	20b	88.	.47
	21a	92.	.54
	21b	92.	.55
	22a	91.	.55
	22b	91.	.55
	23a	81.	.37
	23b	82.	.37
	24a	91.	.60
	24b	91.	.60
	25a	90.	.58
	25b	90.	.59
	26a	82.	.59
	26b	82.	.59
	27a	75.	.32
	27b	75.	.31
SRep	1a	71.	.60
	1b	67.	.61

Spanish Test:

Item #	% responding correctly, P(i), (over all grade levels)	Point-biserial correlation, R(i) (over all grade levels)
SRep 2a	54.	.54
2b	63.	.52
3a	68.	.63
3b	61.	.58
4	26.	.39
5a	42.	.51
5b	41.	.53
6	75.	.63
7	32.	.43
8a	88.	.64
8b	88.	.64
9a	73.	.58
9b	51.	.44
10a	79.	.64
10b	81.	.72
11a	86.	.59
11b	84.	.61
12	85.	.66
13a	68.	.61
13b	73.	.61
14a	45.	.45
14b	70.	.54
15a	80.	.70
15b	78.	.65
15c	71.	.72
16a	53.	.49
16b	36.	.44
17a	76.	.67
17b	72.	.66
18a	88.	.67
18b	57.	.52
19a	82.	.62
19b	76.	.55
20a	75.	.63
20b	73.	.52
21a	48.	.54
21b	62.	.56
22a	59.	.61
22b	48.	.53
23a	83.	.67
23b	66.	.57
24	85.	.67
25	49.	.57
26a	29.	.40
26b	59.	.55

Spanish Test:

Item #		% responding correctly, P(i), (over all grade levels)	Point-biserial correlation, R(i) (over all grade levels)
SR	1	77.	.68
	2	82.	.74
	3	73.	.66
	4	61.	.52
	5	87.	.72
	6	68.	.60
	7	73.	.63
	8	81.	.69
	9	81.	.72
	10	33.	.30
	11	87.	.61
	12	53.	.42
	13	59.	.46
	14	62.	.54
	15	38.	.35
	16	74.	.61
	17	79.	.64
	18	75.	.68
	19	79.	.56
	20	68.	.53
	21	39.	.38
	22	64.	.49
	23	43.	.48
	24	66.	.41
	25	56.	.45
	26	44.	.43
	27a	80.	.53
	27b	81.	.53
	28a	55.	.48
	28b	56.	.48

Approximately 40 percent of the English and 20 percent of the Spanish items require careful investigation as a result of their difficulty indicies. Such figures are not abnormally high for instruments in the development stage, as is the Mat-Sea-Cal. Further, one expects a fair proportion of items to be modified between field test and commercial forms. However, before items are discarded or rewritten, other statistics, especially the item discrimination, are examined.

ITEM DISCRIMINATION INDEX

Most instruments are designed to make distinctions between respondees, based on some criterion. In statistical lexicon this is referred to as the item discrimination index. It identifies non-discriminating questions on the basis of correlational analysis between each item and a criterion score. The criterion measure most often employed is the total score on the instrument itself.[9]

The total percentage score on the respective English and Spanish Tests was employed for analyzing the two language-item pools. As stated previously, all responses were scored as either correct or incorrect. Thus, a point biserial coefficient was computed as the discrimination index. Standards proposed by Guilford and Fruchter, and Ebel[10] were invoked for interpretation of the resulting item-total coefficients. Specificially, items with correlations below .30 were recommended for revision, or exclusion from the instrument. Items exhibiting correlations between .30 and .40 were subject to further investigation (in the form of factor analysis). Items with indicies above .40 were regarded as sufficient in discriminating power for retention in the revised forms. (Discrimination indicies are listed in Tables one and two.)

Compared to these standards, most items of the English and the Spanish Mat-Sea-Cal Tests were psychometrically acceptable. Concern may be raised with the discrimination power of the following items:

English test:

Listening comprehension: 1,2,3,6,8,10,12,14,15,16, 18a thru 27b

Sentence repetition: 12

Structured response: 1,9,10,13,15,18,19,23,24,27a thru 28b

Spanish test:

Listening comprehension: 1,2,5a,5b,6,9a,9b,10,11a,11b, 12,14,16,27a,27b

Sentence repetition: none

Structured response: 10,15

From these data it appeared that a large pool of items on both English and Spanish Mat-Sea-Cal Tests delineated between the orally proficient and those lacking in structure/concept skills. Those items in the grey area of discrimination power, i.e., with indicies between .30 and .40, were subject to further analysis to determine their congruence to test purposes.

FACTOR ANALYSIS

All items with discrimination indicies of .30 and greater were included in the variable pool for factor analysis. For both English and Spanish Tests (separately) on analysis was conducted within each communication mode/item sequence (Listening Comprehension, Sentence Repetition, and Structured Response).

Factoring was intended to explore mathematical relationships among the item-variables that were not, at the time, known. Attention was focused upon latent phenomena of the constructs under consideration, as exhibited by data generated from the item sequences. The end product was descriptive typologies which reflected a substantive sharing of common variation among groupings of item-variables.

In addition, the procedure was commenced for its data reduction potential. A large series of variables can be "rearranged" or "reduced" to a smaller set of source items which account for significant inter-relations in the data. This possibility was also investigated, as a shorter Mat-Sea-Cal was desired in certain situations.

For calculations the principal factoring method was used. Diagonal elements of the correlation matrix were initially replaced by the squared multiple correlation coefficients. Eigenvalues, representing total variance accounted for by a factor, were computed. The number of factors extracted for rotation generally corresponded to Guttman's 1.0 criterion.[11] Varimax rotation was employed, and subsequent commurality estimates were the respective eigenvalues for each extracted factor. Iteration proceeded until convergence occurred, that is, the difference between successive eigenvalues was .01, or less.

All factors had to contain at least three "pure" item variables (i.e., a variable which loaded on one and only one factor). Factor loadings of .35 or greater were considered significant (i.e., the minimum correlation for an item to load on a factor).

In all forty-five computerized, factor-analytic runs were made using the Statistical Package for the Social Sciences.[12] A summary of the findings is presented in Table 3. In general, three types of items were discerned: those that should be retained in their present form, those which are relatively easy though acceptable in discrimination power, and those which require revision.

It should be noted that factor analyzing of "coupled items", those with an "a" and a "b" part, proved difficult. Both parts typically correlated highly. As a result, "a" and "b" items pairs had to be analyzed separately, on different factor runs.

On the English test the Listening Comprehension section needs the greatest amount

Table 3

FACTOR PATTERNING

English Test

Listening Comprehension:

factor 1: P = (83 - 89)*, R = (35 - 44)**
factor 2: P = (96 - 97), R = (35 - 41)

conclusions: (on individual items)

1. retain in present form: #4,9a,9b,10,11, and 17
2. easy items (high P_i, acceptable R_i): #5a,5b,7,13a, and 13b
3. revise: #1-3,6,8,12,14-16, and 18a - 27b.

Sentence Repetition:

factor 1: P = (43 - 96), R = (48 - 62), concept: "number"
factor 2: P = (95 - 96), R = (46 - 57)

conclusions: (on individual items)

1. retain in present form: #1-3,5-9c,13a,13b,15a-16,18-21, 22b,23b,24, and 26
2. easy items (high P_i, acceptable R_i): #4,10a-11,14a,14b, 17a,17b,22a,23a, and 25
3. revise: #12

Structured Response:

factor 1: P = (55 - 78), R = (33 - 49), concept: "temporality"
factor 2: P = (90 - 98), R = (39 - 54), concept: "identification"
complex items: P = (93 - 90), R = (38 - 43)
non-loading items: P = (35 - 39), R = (30 - 33)

conclusions: (on individual items)

1. retain in present form: #3,4,6,8,9,11,12,14,16-18,20-22,25, and 26
2. easy items (high P_i, acceptable R_i): #2,5, and 7
3. revise: #1,10,13,15,19,23,24,and 27a-28b

---

*P = (xx - yy): is the range of the difficulty index of variables included in this factor.

**R = (xx - yy): is the range of the discrimination index (point-biserial coefficients) of variables included in this factor.

Spanish Test

Listening Comprehension:

factor 1: P = (71 - 92), R = (33 - 61)
factor 2: P = (92 - 98), R = (32 - 43)

conclusions: (on individual items)

1. retain in present form: #4,5b,7,13a,13b,15,17, and 19a-26b
2. easy items (high P_i, acceptable R_i): #3,6,12,18a, and 18b
3. revise: #1,2,5a,8-11b,14,16,27a, and 27b

Sentence Repetition:

factor 1: P = (73 - 86), R = (55 - 68)
factor 2: P = (24 - 59), R = (38 - 55)

conclusions: (classification of items by difficulty index)

1. high group: [P = (73 - 86), R = (55 - 68)]: #8a-9a,11a-12,18a, 19a,19b, 23a, and 24

2. low group: [P = (24 - 59), R = (38 - 55)]: #1a,1b,3a,10a,13a, 13b,15a, 15c,17a,17b, 18b,21b, and 23b

3. complex group: [P = (57 - 81), R = (52 - 73)]: #2a,4-5b,7,9b,14a, 16a,16b,21,22b,25-26b

---

4. complex/high*: [P = (70 - 81), R = (52 - 73)]: #6,10b,14b,15b,20a and 20b

5. complex/low**: [P = (59 - 63), R = (52 - 61)]: #2b,3b, and 22a

Structured Response

factor 1: P = (53 - 87), R = (42 - 74), concepts: "number", "identification", and "case relationship"
factor 2: P = (38 - 56), R = (35 - 49)

conclusions: (on individual items)

1. retain in present form: #1-9,11-14,16-20,22,24, and 26-27b
2. difficult items (low P_i, acceptable R_i): #15,21,23,25,28a, and 28b
3. revise: #10

---

*These variables are "complex", though on some runs load as "pure" on the high P_i factor.

**These variables are "complex", though on some runs load as "pure" on the low P_i factor.

of work. The second half of the Listening mode (items 18a through 27b) requires revision, as do six other items. Questions 5a, 5b, 7, 13a and 13b are classified as relatively easy, though possessing reasonable discrimination indicies.

The Sentence Repetition mode supported two-three factors. Items typically grouped according to their difficulty index, thus the two factor solution appeared more appropriate. The difficulty index of the easier item group was comparable to that of similar factor in the Listening section. Also, the communication concept "number" completely loaded on the factor with lower difficulty indicies.

In Structured Response nine items had low discrimination indicies. These need revision, and subsequently were not factor analyzed. The remaining items gravitated into four factors. However, the third factor repeatedly contained only two pure variables, and the fourth was completely composed of complex loadings. As a result, a two factor solution was specified, and the analysis re-performed. The findings were similar to the other two modes, items aggregating by difficulty index. The factor composed of relatively easy items also contained most of the items in the communication concept of identification. The "temporality" variables accrued, en masse, to the other factor.

On the Spanish test in the Listening Comprehension mode, two factors emerged. Again labelling of factors went according to the difficulty index of the respective items. Of note, also, the entire second half of this section is composed of item pairs, (#18a through 27b) and thus had to be analyzed separately. Furthermore, investigating any sizable portion of either "a" or "b" pair-set with the non-paired items yielded a special two-factor solution. One factor contained only members from the paired grouped, the other included all non-paired variables. Analysis of

"a" items (18a,19a . . .27a), then "b" items (18b, 19b . . .27b), separately, resulted in single factor solutions. These item groups are obviously highly correlated. Thus, a few items from the "a" and the "b" groups may be omitted without detriment to assessment purposes. This would result in a shorter Listening Comprehension mode on the Spanish test.

The Sentence Repetition section supported three variable groups, but only two factors. The factors were identifiable by difficulty index. Items with a low percentage of correct responses formed one group. Items with a distinctly higher difficulty indicies loaded on the second factor. The third group of variables had "complex" loadings, that is, they aligned with both factors. Also, a few items vacillated between variable groups on different analytic runs (and are identified in Table 3 as "complex/high" and "complex/low").

In Structured Response the discernable pattern of difficulty levels emerged. The two factor solution facilitated easy classification of all but three items. Further, communication concept items of number, identification, and case relationship loaded on the factor with whose variables possessed higher difficulty indicies.

SUMMARY AND RECOMMENDATIONS

The Mat-Sea-Cal Oral Proficiency Tests, in English and in Spanish (Field Test Edition), have been proposed as a means of assessing children's linguistic skills. The research reported herein examined certain statistical qualities of these instruments.

Both English and Spanish Tests satisfied psychometric standards for reliability. This permitted further meaningful investigation into other characteristics of the instruments, as data derived from them were judged as consistent. Also, the magnitude of the reliability coefficients suggested that these tests were relatively

homogeneous in content. The method of internal equivalence was employed for reliability calculations.

In addition, both language tests contained a large number of items which possessed a desirable difficulty index, specified as the 50 to 90 percent range. These were intended as criterion referenced instruments designed to assess fluency near the minimum oral proficiency level (defined as 70 percent performance). The Spanish Test demonstrated a broad sampling of the target difficulty index. Most of its items were deemed acceptable. The English Test proved more homogeneous with a large concentration in the higher percentages of the index. This was particularly true in the Listening Comprehension mode. As a result, the expenditure of additional effort will be required, particularly in this one section.

Point-biserial coefficients were computed for an item discrimination indicies. By and large most items met accepted psychometric criteria on discrimination power. The Sentence Repetition mode appeared the strongest in this matter, the Listening Comprehension the weakest. An absolute minimum of .30 was invoked for proceeding with further analysis. Items in the .30 - .40 range were rendered extra attention, as such figures suggest the need for additional refinement.

An in-depth factor analysis constituted the final phase of the pursuit. Items with discrimination indicies above .30 were included. The analysis for each test was conducted within the three communication modes. For the analysis principal factoring was applied. Squared multiple correlation coefficients were inserted as initial estimates of commurality, thereafter eigenvalues. Extracted factors were required to have at least three "pure" loadings (of .35 or greater). Generally, two or three item pools were discovered with each communication mode. The items

separated themselves most often according to difficulty index. Complete groups of items representing certain communication concepts did, at times, load on one factor.

The findings suggest that a reduction in the number of items per test is possible. Selection of items would follow the test design, that is, the performance region in which assessment was desired.

Also, it was noted, that combining two language manifestations into one question failed to be a discriminating technique. Item pairs were so highly correlated that the magnitude of their relationship outweighed either item's intercorrelation with all other variables, combined. Thus, each item needs to be a separate entity in future revisions of the instruments.

Next, a small series of relatively easy, but discriminating, items exist on each test. This raises an interesting possibility. Such questions may be separated and used as a mini pre-test for children suspected of having little oral fluency in the given language. As the items possess discrimination power, they offer a reasonably accurate and objective measure. As they are relatively easy, only students with the largest of language deficiencies could be expected to do poorly on them. However, for such students, an exhaustive, in-depth assessment is unnecessary; a brief, but accurate appraisal is what is required.

Finally, a thorough linguistic examiniation of the data is in order. The content of questions missed frequently, and items rarely missed, begs scrutiny. Perhaps certain parts of these tests are too easy or too difficult. Or, perhaps an order of language skill acquisition exists, alone, or in combination with maturational and/or environmental effects. Potential findings from such investigations might provide new directions for classroom instruction in language development, an interesting thought, indeed.

References

[1]National Institute of Education, Linguistic Communication: Prospects for Research: Report of the Study Group on Linguistic Communication, [by George A. Miller, ed.] ([Washington]: National Institute of Education, 1973), introduction.

[2]Muriel R. Saville and Rudolph C. Troike, A Handbook of Bilingual Education (2d ed., Washington: Teachers of English to Speakers of Other Languages, 1973), p.18.

[3]Lau et al. v Nichols, 414 U.S. 563, 39 L.Ed. 2d 1, 94 S.Ct. 786 (1974).

[4]Joseph H. Matluck and Betty M. Matluck, Mat-Sea-Cal Oral Proficiency Tests: English, Spanish, Cantonese, Mandarin, Ilokano, and Tagalog, 6 Vols. (Field Test edition; Arlington, Virginia: Center for Applied Linguistics, 1974), p. 2.

[5]Vincent Doyle, "A Model for the Development of Language Assessment Instruments which Insures Psychometric Quality" (paper read at the Pacific Northwest Council on Foreign Languages Convention, April, 1977, Spokane, Washington).

[6]Julian C. Stanley, "Reliability", Educational Measurement, 2d ed., ed. Robert L. Thorndike (Washington: American Council on Education, 1971), pp. 405-08.

[7]Ibid., p. 434 - 35.

[8]Joseph H. Matluck and Betty M. Matluck, "The Mat-Sea-Cal Instruments for Assessing Language Proficiency" (paper read at the American Educational Research Association Convention, April, 1976, San Francisco, California).

[9]Sten Henrysson, "Gathering, Analyzing, and Using Data on Test Items", Educational Measurement, 2d ed., Robert L. Thorndike (Washington: American Council on Education. 1971), p. 135.

[10]J. P. Guilford and Benjamin Fruchter, Fundamental Statistics in Psychology and Education, (5th ed., New York: McGraw-Hill Book Co., 1973), p. 456; see also Robert Ebel, Measuring Educational Achievement (Englewood Cliffs, N.J.: Prentice-Hall, 1965), p. 364.

[11]Louis Guttman, "Some Necessary Conditions for Common-Factor Analysis," Psychometricka, XIX, (1954), 149-161.

[12]~~Joe-On Kim~~ Tae-On Kim, "Factor Analysis", Statistical Package for the Social Sciences, 2d ed., chap. 24, ed. Norman H. Nie, and others (New York: McGraw-Hill Book Company, 1975), pp. 468-514.