An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses

Antony John Kunnan *The University of Michigan**

There has been relatively little research on analytical procedures for examining the dependability and validity of criterion-referenced tests especially when compared to similar investigations for norm-referenced ESL or EFL tests. This study used three analytical procedures, namely, G-theory, factor and cluster analyses, to investigate the dependability and validity of a criterion-referenced test developed at the University of California, Los Angeles in 1989.

Dependability estimates showed that test scores are not equally dependable for all placement groups and are rather undependable for two out of the four placement groups. Factor analysis of test scores for the placement groups showed that though two-factor solutions were the best solutions for the different groups, there were differences in the way the subtests loaded in the different groups, with progressively fewer subtests loading on the second factor as ability increased. This finding led to the extension study with cluster analysis which showed that a number of students might have been differently placed if subtest scores were used to place them.

I Introduction

While research on statistical procedures for examining the reliability and validity of norm-referenced ESL or EFL tests has been proliferating, relatively little research on similar procedures for criterion-referenced (CR) language tests (for definitions of CR tests, see Cartier, 1968; Glaser and Nitko, 1971; Hambelton, 1982; Nitko, 1984; and Popham, 1978) has been conducted (for exceptions in language testing, see Brown, 1989, 1990; Hudson, 1989; Hudson and Lynch, 1984; and for exceptions in educational measurement, for example, see Berk, 1980, 1984). This is partly because most language testing agencies (for example, Educational Testing Services, New Jersey, USA, and most north American universities) develop and use norm-referenced (NR) tests, and partly because statistical procedures and software for examining the dependability

* Formerly at the University of California, Los Angeles (UCLA).

and validity of a CR test have been inaccessible to language testing researchers.

This study investigated the dependability and validity of a CR test using three analytical procedures, namely, G-theory (with GENOVA [Crick and Brennan, 1982]), and factor and cluster analyses. The test used for this study was the new ESLPE (NESLPE), a criterion-referenced ESL placement test developed at the University of California, Los Angeles in 1989.[1]

II Background

1 The need for criterion-referenced dependability indices

Dependability in a CR approach (for CR tests), which is only a rough equivalent of reliability in an NR approach, is defined as the extent to which test scores can be considered consistent and dependable for decision-making. In an NR approach (for NR tests), reliability is estimated by investigating the consistency of test scores across multiple test administrations, between several forms of the test, or within a single test. This approach to reliability is derived from classical true score theory which has at least two basic restrictive assumptions: the notion of a normal distribution of scores and the notion of parallel tests, which means that under certain conditions observed scores on two test forms that are equivalent will be parallel. In addition, Popham and Husek (1969) argue that classical true score theory with NR reliability estimates are inappropriate for CR tests. This is because CR tests tend to be used in contexts which produce uniformly high scores with little variation and negatively skewed distributions which yield low NR reliability coefficients. Thus, it is clear that using NR reliability estimates with CR test scores would be inappropriate. In terms of score interpretations too, NR and CR approaches are different: the former approach is concerned with determining the relative standing of individuals while the latter is concerned with mastery or non-mastery of domains by individuals.

[1] A version of this paper was presented at the 13th Annual Language Testing Research Colloquium at the Educational Testing Service, New Jersey, in March 1991. A longer version was submitted to the Department of TESL and Applied Linguistics (UCLA) as a PhD qualifying paper (Kunnan, 1990). In the longer version, NESLPE development and factor solutions (with correlations, communalities, eigen values, scree plots and final factor matrices) are presented in full.

2 *Estimating criterion-referenced dependability indices*

Criterion-referenced estimates of dependability have been classified by Hambelton, Swaminathan, Algina, and Coulson (1978) into three different dependability concepts: (a) **agreement of mastery classification decisions** (placement classifications, for this study) (b) **agreement of decisions at cut scores** and (c) **dependability of domain scores**. The first concept is concerned with the consistency of placement decisions. The second concept is concerned with the deviations of student scores about the cut off scores and the consistency of these deviations across forms. And the third concept is concerned with the consistency of the individual's score. A brief discussion of each of these indices follows.

(a) *Agreement of placement classification decisions*: There are five different methods associated with the threshold-loss **agreement** approach for assessing the consistency of placement classification decisions: the Carver (1970), the Swaminathan, Hambelton and Algina (1973), the Huynh (1976), the Subkoviak (1975) and the Marshall–Haertel (1976) methods. Subkoviak (1984, 1988) recommends the Huynh method for standard data sets as it has small errors in estimates and can be used with one administration.

The Huynh method involves the computing of coefficients $\hat{p}_o$ and $\hat{K}$ (Kappa). Coefficient $\hat{p}_o$ refers to the proportion of students consistently classified or placed correctly at each class level and coefficient $\hat{K}$ refers to the proportion of consistent placement classifications beyond chance. Coefficient $\hat{p}_o$ is sensitive to the selected cut score, test length, and score variability. The cut score, however, tends to have much more of an effect on the magnitude of this index than do the other two characteristics. Lower values are associated with cut scores near the mean and higher values at the tails of the score distribution. Coefficient $\hat{K}$, which is always lower than $\hat{p}_o$, is also sensitive to test length and score variability. Higher values of $\hat{K}$ are associated with cut scores near the mean and lower values at the tails. And the longer the test and the greater the score variance, the higher the index. Both coefficients have an upper limit of 1.00, while the $\hat{p}_o$ coefficient has a chance lower limit of .50 and the $\hat{K}$ coefficient has the lower limit bound of .00. Subkoviak's (1988) recent short cut procedures for estimating both these coefficients from a single test administration, and demonstrated by Brown (1990), have made these two indices easier to compute.

(b) *Agreement of decisions at cut scores*: This squared-error loss approach provides **agreement** indices based on squared deviations of

individual scores from the cut score. Unlike the threshold-loss agreement approach discussed above, this approach is sensitive to the degree of mastery or non-mastery. This approach is also concerned with the misclassification of individuals who are at extreme distances from the cut score rather than those who are close to the cut score. Brennan (1980), applying G-theory (Brennan, 1983; Cronbach, Gleser, Nanda and Rajaratnam, 1972), provides a $\Phi(\lambda)$ index, an index of dependability for CR tests, where λ equals the cut score. This index can be interpreted with assistance from the signal to noise ratio related to it. However, the sensitivity of this index to the cut score has been criticized by Shavelson, Block and Ravitch (1972) though this is a feature of the index that makes it suitable for assessing the absolute decisions of classification in CR testing. Hudson (1989) too found two problems with the interpretation of this index: when the item mean is near the value of the λ, the $\Phi(\lambda)$ estimate is lowered and the $\Phi(\lambda)$ may not be sensitive to false negative classifications. Both generalizability (G) and decision (D) studies within G-theory compute this index with cut scores.

(c) *Dependability of domain scores*: This estimate is a general purpose estimate of **dependability** of domain scores on CR tests without reference to a cut score. It refers to the consistency of an individual's scores on a domain and is provided by Brennan's (1980) Λ coefficient and the signal to noise ratio related to it. This index is the ratio of the universe score variance to observed score variance, where universe score variance is observed score variance minus error which includes variance in items. It is computed by G- and D-studies within G-theory, with the GENOVA software program, though its practical utility and interpretability have been questioned by Berk (1980) and Hudson (1989).

3 Validity of criterion-referenced tests

Though the CR testing literature is vast, only a few researchers have attempted to devine how to examine the validity of CR tests. Hambelton (1984) lists an assortment of methods that can be used to assess the construct validity of CR tests. These include content analysis, item-objective congruence analysis, Guttman scalogram analysis, exploratory and confirmatory factor analysis, experimental studies and the multitrait-multimethod approach. Each of these methods is best suited for specific purposes and since the construct validation process is so critical to test development and use, it is

necessary to make informed decisions about the choice(s) of methods necessary in any situation.

III Method

1 Subjects and instrument

The subjects were 390 non-native speakers of English who had been required to take the NESLPE as part of the admission and placement procedures at UCLA. All the subjects were given Form A of the NESLPE. The NESLPE followed Popham's (1978) principles of criterion-referenced test construction. The test had 100 multiple-choice and true/false items in all: 30 in the listening section, 40 in the reading and vocabulary section, 30 in the grammar section, and an additional composition writing section. The time allotted for the test was 90 minutes for the 100 items and an additional 30 minutes for the composition writing task. The order of sections was: composition, listening, reading and vocabulary, and grammar. However, this study will only consider 90 of the multiple-choice and true/false items: 30 from each section. This was necessary so that the GENOVA software program (Crick and Brennan, 1982), which requires a balanced design for all sections for computing generalizability statistics, could be used.

2 Procedures

Distributions, correlations, reliabilities, exploratory factor analysis and cluster analysis were done using SPSS-X on the 3090 mainframe computer at UCLA. The $\hat{p}_o$ and $\hat{K}$ estimates, were calculated manually, while all the other CR dependability estimates were computed by using GENOVA (Crick and Brennan, 1982) which implements G-theory. For GENOVA, a fixed effects design with I (items) nested within T (test sections) and the number of test sections set to three was used for the D-study.

Exploratory factor analysis (EFA) for all four ESL class groups was done to investigate the validity of the NESLPE. For the EFA, each of the three sections of the test was divided into two subtests in order to more carefully investigate the patterns of relationships among similar and different tests as a basis for examining validity. Thus, the following six subtests were created: Listening 1, Listening 2, Reading 1, Reading 2, Grammar 1 and Grammar 2. The new listening and reading divisions were created in such a way that the first lecture or passage and their items became sections Listening 1 or Reading 1. The second lecture or passage and their items became

Listening 2 or Reading 2. The first grammar section was divided by placing the first two paragraphs in Grammar 1 and the third and fourth paragraphs in Grammar 2.

As an extension of this analysis, cluster analysis (Aldenderfer and Blashfield, 1984) of all the cases was performed. For this analysis, the scores on the original three sections, listening, reading and grammar, were used to obtain clusters from all 390 cases. The clustering method used was the complete linkage method or the 'furthest neighbour' technique in which the distance between two clusters is computed as the distance between their two furthest points. Four cluster groups, reflecting the four ESL classes (33A, 33B, 33C and 35 and exempt) were forced from the data. This analysis was done separately on the four scores: listening, reading, grammar and total. A dendrogram helped interpret the clusters of cases for each of the four clusters.

3 Research question

The research question of the study was: What is the dependability and validity of the New ESLPE? Dependability was examined with the help of Generalizability theory (Cronbach *et al.*, 1972) and validity by using exploratory factor analysis and cluster analysis.

IV Results and discussion

1 Descriptives

Table 1 provides general descriptive statistics for the total group ($N = 390$) for all the items ($k = 100$). Table 2 provides the same

Table 1 Descriptive statistics for all 100 items

Sections	*k*	Mean	SD	KR-20
List	30	21.80	5.78	.87
Read	40	29.66	6.47	.86
Gram	30	24.38	5.26	.88
Total	100	75.84	15.96	.95

Table 2 Descriptive statistics for 90 items

Sections	*k*	Mean	SD	KR-20
List	30	21.80	5.78	.87
Read	30	21.66	4.91	.82
Gram	30	24.38	5.26	.88
Total	90	67.84	14.44	.94

information for the abbreviated item test ($k = 90$).[2] While the means and standard deviations in these two tables are different, the reliability indices (KR-20) for the total test are quite similar: .946 and .940. Table 3 presents general descriptive statistics for the four ESL class groups separately.

Summarizing the tables, the lowest means for all sections and total are for the 33A group and the highest means for all sections and total are for the 35 and exempt group; the lowest means for a section within a group generally is listening, followed by reading-vocabulary (hereafter, reading) and, finally, grammar; KR-20 is quite low for all sections except for grammar and total test for the 33A group. Comparing Table 3 to Table 2, the variances for the groups are all smaller than those observed for the total group whereas the reliabilities vary considerably.

2 *Agreement of placement classification decisions*

Two threshold-loss agreement indices were computed to provide coefficients that estimate the agreement of placement classifications

Table 3 Descriptive statistics for groups ($k = 90$)

Sections	Mean	SD	KR-20
33A Group, $N = 71$			
List	**12.94**	3.27	**.41**
Read	**14.09**	3.96	**.60**
Gram	**16.24**	6.06	**.85**
Total	**43.27**	10.18	**.82**
33B Group, $N = 60$			
List	**18.27**	3.57	**.54**
Read	**19.28**	2.60	**.21**
Gram	22.90	3.04	**.55**
Total	60.45	3.13	**−.62**
33C Group, $N = 73$			
List	21.93	2.58	**.24**
Read	21.97	1.78	**−.40**
Gram	25.26	2.22	**.32**
Total	69.16	2.28	**−1.49**
35 and Exempt group, $N = 186$			
List	26.26	2.33	**.47**
Read	25.20	2.21	**.34**
Gram	27.62	1.51	.14
Total	79.09	4.11	.52

[2] All further tables presented in this study are for the abbreviated test ($k = 90$).

decisions. In the case of the abbreviated NESLPE, students who were exempted from any ESL class are those who secured total scores of 80% or above (the cut score was 72).[3] Students who scored less than 80% were placed in ESL classes that matched their overall ability in terms of the score.

Table 4 provides the ESL class group, the score range and the cut score for each level. In addition, it presents the dependability indices ($\hat{p}_o$ and $\hat{K}$) for each of these placement classification decisions for all groups: 33A, 33B, 33C and 35 and exempt groups. The $\hat{p}_o$ agreements are high for all groups, decreasing as the level goes higher. The dependability of placement classifications decisions as the cut scores is quite high; more agreeable at 33A and 33B in comparison to 33C and 35 and exempt groups. Still, they are high especially given the observation by Subkoviak (1980) that $\hat{p}_o$ agreement coefficients for one administration will be an underestimate of the values that would be obtained using two separate administrations. Coefficient $\hat{p}_o$ would, therefore, be best suited for the purpose of judging the dependability of placement classifications.

Coefficient $\hat{K}$ estimates are noticeably lower than the $\hat{p}_o$ coefficients because they are corrected for chance agreements (as though there were more than one administration). In addition, they are ordered differently from the $\hat{p}_o$ coefficients for the different class groups, with the agreement higher for the high ability group and lower for the lower ability group.

Table 4 Dependability of placement classification decisions

Group	Range	Cut score		Coefficients	
		raw	proportion	$\hat{p}_o$	$\hat{K}$
*	0–46	–	–	–	–
33A	47–55	47	**.52**	.98	.58
33B	56–64	56	**.62**	.95	.63
33C	65–71	65	.72	.90	.68
35 & Ex.	72–90	72	.80	.86	.71

* low for regular UCLA placement

3 *Agreement of decisions at cut scores*

While the $\hat{p}_0$ and $\hat{K}$ coefficients estimate the agreement of mastery non-mastery decisions, treating these as categories, squared-error loss agreement indices do this with sensitivity along

[3]These cut scores for the abbreviated test ($k = 90$) are modified from the placement scores for the full version of the NESLPE ($k = 100$).

the score continuum. This approach, in other words, takes into consideration differences of students' scores from the cut score, that is, degrees of mastery or non-mastery, rather than the simple categorization.

Table 5 presents information from the D-studies for the whole group based on the fixed model design. This design sets the size of the object of measurement, *P*, at infinite, the test section facet (or T) to three and the items facet (or I) to sizes of 20, 30 and 40. Reading Table 5 from left to right, it is apparent that there is a steady drop in Φ (λ) agreement coefficients from low to high cut scores: they are higher for lower scores and lower for the higher scores at the right end. For example, the first row shows a drop from .97 to .91. Comparing the Φ (λ), for differing numbers of items, it is clear that the coefficients increase as the number of items increases.

Table 5 Agreement of decisions at cut scores (from the D Studies) Fixed model design: P = Indefinite, $T = 3$, $I = 20, 30, 40$

Group	33A	33B	33C	35 and Ex.
Raw cut score	47	56	65	72
Items = 20 Φ (λ)	.97	.95	.91	.91
Items = 30 Φ (λ)	.98	.96	.94	.94
Items = 40 Φ (λ)	.99	.97	.95	.95

4 *Dependability of domain scores*

Table 6 presents Brennan's Φ coefficient (1980), the index of domain score dependability, for the different groups with facet T fixed at three and facet I fixed at 30. These conditions reflect the present format of the NESLPE. Since the Φ coefficient can be interpreted as a general purpose estimate of the dependability of a domain score of a CR test and the total group provides the highest coefficient, the NESLPE can be said to be best dependable for the

Table 6 Dependability of domain scores (from the D Studies), Fixed model design P = Infinite, $T = 3$, $I = 30$

Group	Coefficients Φ
33A	.80
33B	.55
33C	.30
35 and Ex.	.48
Total	.93

total group. As for the other groups, the 33A group has the next highest coefficient followed by the 33B group. The 33C group has a low coefficient indicating the low dependability of the domain score for this group.

5 *Summary of indices*

Table 7 presents a summary of agreement dependability and (NR) reliability coefficients for all groups. All the agreement indices (except the $\hat{K}$) are highest for the lowest cut score and lower for the higher cut scores. Comparing the $\hat{p}_o$ and the Φ (λ) coefficients may be the most useful. In this case because both indices are quite close and perform in the same manner, it is not difficult to interpret them. According to both indices, the agreement of placement classification decisions and the agreement of decisions at these cut scores are generally within acceptable limits.

Brennan's Φ coefficient and the KR-20 coefficient have similar patterns across the groups. Both the coefficients are high for the 33A group and for the total group but sag in the middle with the 33C group getting the lowest coefficients. Besides, as noted by Brennan (1984), Φ will be less than KR-20. In this case for example, Φ for the 33A group is .80 which is less than KR-20 (.82). Similarly, the Φ for total group is .86 while the KR-20 is .94. And, the dramatic drop in the Φ (.30) for the 33C group is accompanied by an equally dramatic drop in the KR-20 (−1.49).

From all this information, two observations can be supported: one, that the NESLPE is not dependable to the same extent for all groups. A possible reason for the first observation could be that the NESLPE is not able to assess the ability of students at all levels accurately because there is not sufficient item to specification congruence at all levels. In addition, if the specifications were more carefully laddered, they would have a better chance of assessing student ability at the different levels.

A second observation is that both reliability and dependability

Table 7 Summary of agreement, dependability and reliability coefficients

Group	$\hat{p}_o$	$\hat{K}$	Coefficients* Φ (λ) (Items = 30)	Φ	KR-20
33A	.98	.58	.98	.80	.82
33B	.95	.63	.96	.55	−.62
33C	.90	.68	.94	.30	−1.49
35 and Ex.	.86	.71	.94	.48	.52
Total	–	–	–	.93	.94

*$\hat{p}_o$, $\hat{K}$, Φ (λ) and Φ are CR coefficients; KR-20 is an NR coefficient

coefficients for the total group can be deceptive and falsely encouraging. KR-20 and Φ for the total group are high suggesting that the NESLPE is a dependable test while for the different placement groups, these indices are lower, particularly for the 33A, 33B, and 33C groups suggesting that the NESLPE score is not a dependable indicator of proficiency for these groups.

Finally, the agreement indices are much higher than the dependability coefficients. This is a good sign, since this indicates reasonably consistent placement decisions. Dependability of domain score estimates are of much less importance, since no other use than placement is made of these scores.

6 *Exploratory factor analysis*

EFA were performed on the Pearson product-moment correlations for the six sections. All correlation matrices were examined for appropriateness of the common factor model. They satisfied Bartlett's test of sphericity; all group matrices had high values for this statistic and the associated significance level was small. Thus, it was possible to reject the hypothesis that the population correlation matrix is an identity. Another test used to examine the correlation matrices was the Kaiser-Meyer-Olkin measure of sampling adequacy. Values ranged from .90 for the total group to .44 for the 33C group. All group matrices with values above .50 are said to have adequate sampling (Kaiser, 1974); only the 33B and 33C groups had marginally less adequate sampling.

Several extraction methods were used for all the groups: the principal axes factoring, alpha factoring and unweighted least squares. After initial factor matrices and rotated matrices of all extractions were examined, it was decided to use the alpha factoring method. This was because no computation problems were encountered with the alpha factoring extractions and because the principal axes factoring (PAF) extraction terminated due to communalities of variables exceeding 1.0 and the unweighted least squares extraction had problems with the degrees of freedom not being positive. In cases where solutions from two or all three extractions were available, the differences in solutions produced by the different extractions were minimal.

In addition, alpha factoring was preferred because it is based on principles similar to G-theory. Kim and Mueller (1978) state that in alpha factoring '. . . variables included in the factor analysis are considered a sample from the universe of variables, while assuming that these variables are observed over a given population of individuals' (1978: 26). Computationally, too, the alpha factoring

method offers a good choice: Kaiser and Derflinger (1990) state that it is a psychometric method rather than a statistical method, and it 'treats the number-of-factors question more sensibly, . . . (and) is numerically better behaved' (1990: 32).

In actual computation, the communality estimates given by the square multiple correlations are used first, followed by an adjustment of the matrix following the assumption that the observed variables are only a sample from the universe of variables. The variables are rescaled according to the communality and the iteration process continues until the communalities converge.

After it was decided to use the alpha factoring method, the problem of number-of-factors to be extracted arose. The initial decision about the appropriate number of factors to be extracted was made after scrutinizing the eigen values obtained from the initial extraction using the criteria of substantive importance and the scree-test. Several numbers of factors were then extracted, and oblique rotated factor structures were examined to determine if factors were correlated. For those solutions in which interfactor correlations were small, orthogonal rotations were performed. The final determination regarding the number of factors and the best solution was made on the basis of two criteria, simplicity and interpretability. Simplicity was evaluated by examining the factor loadings for salient loadings and interpretability by evaluating the extent to which salient factor loadings corresponded to the sections of the test. Only the final interpretable factor solutions and related statistics are presented here.[4]

Tables 8 to 12 present factor solutions for the different groups.

Table 8 Exploratory factor analysis: 33A group, factor structure matrix (oblique rotation)

Variable	Factor 1	Factor 2
LIST1	.25980	.49482
LIST2	.09690	.30255
READ1	.58568	.64847
READ2	.55808	.13256
GRAM1	.73990	.47885
GRAM2	.68967	.39590
Factor correlations		
Factor 1	1.00000	
Factor 2	.42369	1.00000

[4]EFA of placement subgroups was done primarily in order to describe the factor structure for each of the groups. Muthen (1989) argues that a new approach which uses the Pearson-Lawley formulas avoids the problems associated with subpopulation factor analysis.

Table 9 Exploratory factor analysis: 33B group, rotated factor matrix (orthogonal rotation)

Variable	Factor 1	Factor 2	h^2
LIST1	.77076	.18898	.62978
LIST2	.36224	.73359	.66937
READ1	.37421	.04865	.14240
READ2	.48885	.21479	.28511
GRAM1	.08631	.52180	.27973
GRAM2	.00340	.40191	.16154
Eigenvalues	1.11175	1.05617	2.16792
Tot. Var. %	18.53	17.60	36.13

Table 10 Exploratory factor analysis: 33C group, rotated factor matrix (orthogonal rotation)

Variable	Factor 1	Factor 2	h^2
LIST1	.43501	.02395	.18980
LIST2	.23047	.70385	.54852
READ1	.23244	.42450	.23423
READ2	.41490	.01420	.17234
GRAM1	.73613	.14077	.56171
GRAM2	.54578	.36859	.43374
Eigenvalues	1.34413	.76921	2.11334
Tot. Var. %	22.40	13.30	35.70

Table 11 Exploratory factor analysis: 35 and exempt, rotated factor matrix (orthogonal rotation)

Variable	Factor 1	Factor 2	h^2
LIST1	.41221	.05195	.17262
LIST2	.37695	.13608	.16061
READ1	.30888	.12164	.11021
READ2	.47526	.14452	.24676
GRAM1	.17204	.09349	.03834
GRAM2	.09666	.76960	.60163
Eigenvalues	.67223	.65792	1.33015
Tot. Var. %	11.20	10.97	22.17

Table 12 Exploratory factor analysis: total group, rotated factor structure (orthogonal rotation)

Variable	Factor 1	Factor 2	h^2
LIST1	.44694	.70394	.69528
LIST2	.42250	.71617	.69141
READ1	.59488	.51449	.61859
READ2	.65133	.49321	.66775
GRAM1	.78984	.45686	.83258
GRAM2	.70711	.40695	.66561
Eigenvalues	3.99873	.18249	4.18122
Tot. Var. %	66.50	3.00	69.50

For each of the four class groups and the total group, two-factor solutions were the most parsimonious and interpretable. All solutions except that for the 33A group were orthogonal.

Table 13 summarizes all solutions for all groups, and shows the differences in factor structures of the NESLPE for the different groups. For the 33A group, the two-factor oblique solution shows that though the listening and grammar subtests loaded on the same factors, the two reading subtests loaded on separate factors. The interfactor correlation was moderately high (.424). The two-factor orthogonal solution for the 33B group produced another pattern, in which the listening subtests loaded on separate factors. The two-factor orthogonal solution for the 33C group show that one subtest of each of the listening and reading loaded on the second factor and for the 35 and exempt group only one subtest of the grammar loaded on the second factor. At the bottom of the table, for the total group, the two subtests of listening loaded on one factor while the reading and grammar subtests loaded on the other factor.

From the point of view of the skills, the two listening subtests loaded on the same factor for three groups (33A, 35 and exempt, and the total groups) but on separate factors for the 33B and the 33C groups. The reading subtests loaded on the same factor for all groups except for the 33A and 33C groups and the grammar subtests loaded on the same factor for all groups except for the 35 and

Table 13 Summary of factor solutions for different groups

Group	Solution	Factors	
		1	2
33A	2 factors Obliq.	Read2 Gram1 Gram2	List1 List2 Read1
33B	2 factors Ortho.	List1 Read1 Read2	List2 Gram1 Gram2
33C	2 factors Ortho.	List1 Read2 Gram1 Gram2	List2 Read1
35 and Ex.	2 factors Ortho.	List1 List2 Read1 Read2 Gram1	Gram2
Total	2 factors Ortho.	Read1 Read2 Gram1 Gram2	List1 List2

exempt group. One way to interpret this is that test sections which have higher variance have the subtests loaded on separate factors.

Going back to Table 13, two differences between the factor structures in the groups can be observed: first, only the 33A group has an oblique solution and second, the variables that loaded on the factors for each of the groups is different. The first difference might indicate that students with lower level ability (33A group) have inseparable skill ability as against students at higher levels of ability who have distinct skill abilities. The second differences seems to indicate that the NESLPE does not measure the same abilities across all groups.

These findings also seem to indicate that the NESLPE is not unidimensional. This lack of unidimensionality raises a critical question: if the test is not unidimensional, then should placement decisions be based on a single composite score which is supposed to represent a single indicator of language ability? Or, if placement decisions are based on single composite scores, would there be any misclassification of students?

Previous factor analytic studies on similar though different data sets have shown a single-factor solution based on total group analyses (example, Davidson, 1988) and this has legitimatized somewhat placement decisions based on the single composite score. These single composite scores reflected the so-called unidimensional factor structure of the test. So, administrators could add up all scores for subtests to make the single composite score. At UCLA, too, this single composite score, based on all the skills tested, was generally used to place students into ESL classes or to exempt them. In this type of procedure, a very low score would place a student into a low level ESL class, like ESL 33A at UCLA, which could focus on listening and speaking skills. A higher score would place a student into a higher level ESL class, like ESL 35 at UCLA, which could focus on reading and writing skills. Thus, at UCLA, the single composite score determined not only the level, but also the kind of class the student would place into. This can clearly be a disservice to many students. For example, a student with low listening section scores, but with very high reading scores would be placed into a higher level ESL class which could focus on reading and writing skills and generally neglect listening and speaking skills.

7 *Cluster analysis*

In order to investigate whether there was any misclassification of students based on the composite score, a cluster analysis was performed. The test score data for the total group ($N = 390$,

$k = 100$) was first clustered on the single composite score and then on each of the section scores, listening, reading and grammar, separately. The results show that the cluster solutions for each cluster have different group memberships. The cluster solutions also show that many students would have placed differently into ESL classes if section scores rather than the single composite score was used.

Table 14 presents UCLA ESL class placement (in ascending case numbers) based on the total score and the cluster membership based on the total score.[5] The sample sizes for the two groupings show big differences between actual placement and cluster membership for the 33A and 33C groups. However, more critical than sample size is the range of individuals in each group: the UCLA grouping seems

Table 14 UCLA placement compared to cluster membership* ($N = 390$, $k = 100$)

Groups	UCLA placement	#	Cluster membership	#
33A	1– 71	71	1– 37	37
33B	72–131	60	38– 98	61
33C	132–204	73	99–236	138
35 and ex.	205–390	186	237–390	154

* based on ID numbers

broader at the 33A and 35 and exempt groups when compared to the cluster grouping which is broad at the 33C group.

Table 15 presents the number of students who would have been placed differently if section scores were used. Specifically, if students were clustered on the listening score, three students from the lowest cluster group (33A) would have been placed in 'level 3'.[6] Similarly, two students from the second cluster group would have been placed into 'level 4' had they been placed according to their listening scores. A more critical difference is the four students in the 33C group who were placed at the higher level though based on their listening section score would have been placed into the lowest level.

The cluster groupings based on reading show more difference: 28 students from the 33A group would have been placed higher (27

[5]UCLA placements are only best approximates: students could be placed lower if their performance on the writing task (which is rated and used for placement) is rated lower than the level they were initially placed into. In practice, about 10% to 15% of the students are replaced according to this procedure. This information could be valuable for assessing the dependability of placement decisions.

[6]The term 'level' with numbers 1, 2, 3 and 4 will be used rather than 'beginning' listening, 'intermediate' listening, 'advanced' listening, etc. This will help keep the present course numberings and their curricula free of any association with the cluster groups based on section scores which may imply beginning, intermediate and advanced courses.

into level 3 and 1 into level 4) and 13 students from 33B (all would have been placed into level 4). Clustering on grammar scores show the most difference: 36 students from the 33A group and 22 from the 33B group would have been placed higher.

When these figures are added up, it may seem that as many as 108 placement differences have occurred from all the three section score analyses but these 108 do not refer to 108 students as there is overlap. Besides, these placement differences can be taken as serious violations only if the section scores were based on a strict objectives-based test, which is laddered, and a matching level-based class for each skill is organized. Another way of looking at this is: if the placements based on **total** score are correct, then the numbers in Table 15 indicate how many students would have been misplaced if the section scores had been used.

Table 15 Number of students who would have been placed differently if section scores were used

Group	Level	Listening	Reading	Grammar
33A	1	3*	27*	29*
			1**	7**
33B	2	2**	13**	22**
33C	3	4@	—	—
35 and ex.	4	—	—	—

* would have been placed into level 3;
** would have been placed into level 4;
@ would have been placed into level 1

What these figures here give us in an indication of how placements can differ if they are based on section scores rather than on the total score. Thus, the use of section scores might be a more accurate procedure to consider for placement especially because the factor structure of the placement groups is different for each group, thus, making a composite score less reliable for placement.

V Conclusion

The dependability estimates of the NESLPE showed that dependability for the total group was different from the estimates for the four ESL class groups. Furthermore, it showed that the dependability of domain scores was the lowest for the 33C and 35 and exempt groups. Thus, this analysis showed that test scores are not equally dependable for all groups and are very undependable for two out of the four groups. These low dependability estimates for the two groups could be due to less accurate item to specification congruence for those groups. Agreement indices also differed across cut scores. But while dependability indices for some groups were

unacceptably low, agreement indices for all cut scores were generally above acceptable levels.

Validity of the NESLPE was investigated with EFA which showed that though two-factor solutions were the best solutions for different groups, there were differences in the way the subtests loaded in the different groups, with progressively fewer subtests loading on the second factor as ability increased. This is consistent with the findings of Oltman and Stricker (1988) who found a greater test dimensionality of ability at lower levels than at higher levels. This finding led to the extension study with cluster analysis which showed the number of students who might have been differently placed if section scores were used to place them. Overall, this study benefited from group level analyses, following Upshur and Homburg (1983), as it revealed results that are normally hidden in total group analysis. In addition, it confirmed an earlier study (Kunnan, 1986) that testing for placement and diagnosis should be criterion-referenced and scores used for placement should be skill profiles rather than total scores.

To conclude, this study showed that the dependability and validity of a CR test like the NESLPE could be investigated relatively easily with G-theory, factor and cluster analytical procedures. Continuous monitoring of the NESLPE, however, is essential to improve the dependability and the validity of the test, in addition to research in determining appropriate test length and in setting test performance standards.

Acknowledgements

I would like to thank Lyle F. Bachman, Robert Boldt, Fred Davidson, Brian Lynch, Donna Brinton, Jack Upshur and two anonymous reviewers for their helpful comments on an earlier version. Any deficiencies, of course, are mine.

VI References

Aldenderfer, M.S. and **Blashfield, R.K.** 1984: *Cluster analysis*. Newbury Park, CA: Sage Publications.

Berk, R.A., editor, 1980: *Criterion-referenced measurement: the state of the art*. Baltimore: The Johns Hopkins University Press.

—— 1984: *A guide to criterion-referenced measurement*. Baltimore: The Johns Hopkins University Press.

Brennan, R.L. 1980: Applications of generalizability theory. In Berk, R.A., editor, *Criterion-referenced measurement: the state of the art*. Baltimore: The Johns Hopkins University Press, 186–232.

—— 1983: *Elements of generalizability theory*. Iowa City: American College Testing Program.

—— 1984: Estimating the dependability of the scores. In Berk, R.A., editor, *A guide to criterion-referenced test construction*. Baltimore: The Johns Hopkins University Press, 292–334.

Brown, J.D. 1989: Improving ESL placement tests using two perspectives. *TESOL Quarterly* 23, 65–83.

—— 1990: Short-cut estimators of criterion-referenced test consistency. *Language Testing* 7, 77–97.

Cartier, F.A. 1968: Criterion-referenced testing of language skills. *TESOL Quarterly* 1, 27–32.

Carver, R.P. 1970: Special problems in measuring change with psychometric devices. In *Evaluative research: strategies and methods*. Pittsburg: American Institute for Research, 48–63.

Crick, J.E. and **Brennan, R.L.** 1982: *GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual)*. Dorchester, Mass.: Computer Facilities, University of Massachusetts at Boston.

Cronbach, L.J., Gleser, G.C., Nanda, H. and **Rajaratnam, N.** 1972: *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley.

Davidson, F.G. 1988: An exploratory modeling survey of the trait structures of some existing language test data sets. Unpublished doctoral dissertation. Los Angeles: UCLA.

Glaser, R. and **Nitko, A.J.** 1971: Measurement in learning and instruction. In Thorndike, R., editor, *Educational measurement*. Washington, DC: American Council of Education, 625–70.

Hambelton, R.K. 1982: Advances in criterion-referenced testing technology. In Reynolds, C.R. and Gutkin, T.B., editor, *The handbook of school*

Hambelton, R.K. 1982: Advances in criterion-referenced testing technology. In Reynolds, C.R. and Gutkin, T.B., editor, *The handbook of school psychology*. New York: John Wiley & Sons.

Hambelton, R.K., Swaminathan, H., Algina, J. and **Coulson, D.B.** 1978: Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research* 48, 1–47.

Hudson, T. 1989: Measurement approaches in ability level functional language tests. Unpublished doctoral dissertation. Los Angeles: UCLA.

Hudson, T. and **Lynch, B.** 1984: A criterion-referenced approach to ESL achievement testing. *Language Testing* 1, 171–201.

Huynh, H. 1976: On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement* 13, 253–64.

Kaiser, H.F. 1974: An index of factorial simplicity. *Psychometrika* 39, 31–36.

Kaiser, H.F. and **Derflinger, G.** 1990: Some contrasts between maximum likelihood factor analysis and alpha factor analysis. *Applied Psychological Measurement* 14, 29–32.

Kim, J.-O. and **Mueller, C.W.** 1978: *Factor analysis: statistical methods and practical issues*. Newbury Park, CA: Sage.

Kunnan, A.J. 1986: Making classroom testing useful to teachers and learners. *Indian Journal of Applied Linguistics* 12, 101–17.

—— 1990: An investigation of the dependability and validity of the New ESLPE. PhD qualifying paper. Department of TESL and applied linguistics. Los Angeles: UCLA.

Marshall, J.L. and **Haertel, E.H.** 1976: The mean split-half coefficient of agreement: a single administration index for reliability of mastery tests. Unpublished manuscript. University of Wisconsin.

Muthen, B.O. 1989: Factor structure in groups selected on observed scores. *British Journal of Psychological Society* 42, 81–90.

Nitko, A.J. 1984: Defining 'criterion-referenced test'. In Berk, R.A., editor, *A guide to criterion-referenced test construction*. Baltimore: The Johns Hopkins University Press, 8–28.

Oltman, P.K., Stricker, L.J. and **Barrows, T.** 1988: *Native language, English proficiency, and the structure of the TOEFL*. (TOEFL research report no. 27). Princeton, NJ: Educational Testing Service.

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

Popham, W.J. and **Husek, T.R.** 1969: Implications of criterion-referenced measurement. *Journal of Educational Measurement* 6, 1–9.

Shavelson, R.J., Block, J.H. and **Ravitch, M.M.** 1972: Criterion-referenced testing: comments on reliability. *Journal of Educational Measurement* 9, 133–37.

Subkoviak, M.J. 1975: Estimating reliability from a single administration of a mastery test. (Occasional Paper no. 15). Madison: Laboratory of Experimental Design, University of Wisconsin.

—— 1980: Decision-consistency approaches. In Berk, R.A., editor, *Criterion-referenced measurement: the state of the art*. Baltimore: The Johns Hopkins University Press, 129–85.

—— 1984: Estimating the reliability of mastery-nonmastery classifications. In Berk, R.A., editor, 1984. *A guide to criterion-referenced measurement*. Baltimore: The Johns Hopkins University Press, 267–91.

—— 1988: A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement* 25, 47–55.

Swaminathan, H., Hambelton, R.K. and **Algina, J.** 1973: Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Research* 11, 263–67.

Upshur, J.A. and **Homburg, T.J.** 1983: Some relations among tests at successive ability levels. In Oller, J., editor, *Issues in language testing research*. Rowley: Newbury House, 188–202.