Modelling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation

Antony John Kunnan

Language Testing 1994 11: 225
DOI: 10.1177/026553229401100301

Modelling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation

Antony John Kunnan *California State University, Los Angeles*

Construct validation has seen two proposals recently: the use of construct representation and nomothetic span variables through structural modelling and the concept of population generalizability. This study investigated the influence of two major test-taker characteristics (TTCs), social milieu or cultural background and exposure or previous instruction, on test performance (TP) in tests of English as a foreign language (EFL) for two native language groups, the non-Indo-European (NIE) and the Indo-European (IE). Data from the Cambridge-TOEFL comparability study (Bachman *et al.*, 1991; $N = 1448$) from eight sites in eight countries was used. The instruments were 1) a 45-item Likert scale background questionnaire which captured the test-taker characteristics; and 2) the First Certificate in English, administered by the University of Cambridge Local Examinations Syndicate, the TOEFL and the SPEAK, administered by the Educational Testing Service, Princeton, and the Test of English Writing. Modelling of the TTCs and the TP factors generally supported an equal influence factors model (where the factors have equal status) and an intervening factors type model (where the factors are not equal in status and one factor is an intervening factor) for both the NIE and IE groups.

I Modelling and generalizability in construct validation

Construct validation has seen new proposals in emphasis and orientation since Cronbach and Meehl's (1955) early proposal. Two of these proposals could rewrite the way construct validation research is conducted in the future. The first change is a conceptual one. In the late 1950s and 1960s, construct validation was strictly formulated as hypothesis testing and had a confirmationist bias or a convergent and discriminant validation emphasis (Campbell and Fiske, 1959). Responding to criticisms of this approach (see Bechtold, 1959; Loevinger, 1957), Bentler (1978) proposed a causal-modelling approach to construct validation. This proposal extended the Cronbach–Meehl framework by arguing for the construct validation of a substantive theory, focusing attention on the entire

nomological network of associations of a given construct to other constructs and manifest variables.

Building on this proposal, Embretson (1983) argued that the construct validation of a test depended on both *construct representation* (theoretical constructs that explain responses to items) and *nomothetic span* (the utility of the test for measuring individual differences). This new orientation to construct validation would include not only validation of test performance (which investigates construct representation) but also validation of structural relationships among test performance and individual difference variables, such as test-taker characteristics (which investigates the nomothetic span).

Messick (1989: 48–49) sums up the value of using this construct validation approach:

> ... if the constructs operative in test performance have been previously identified in the construct representation phase, quantitative models that permit a priori construct specification may be applied to the correlational data. For example, path-analytic or structural models (Bentler, 1980; James *et al.*, 1982; Joreskog & Sorbom, 1979) may be used to appraise the extent to which the component constructs can account for the test's external pattern of relationships.

From a philosophical perspective too, this construct validation approach fits well with the Kantian inquiry system which entails the formulation or identification of alternative perspectives on a theory or problem representation which explicitly recognizes the strong intertwining of theory and data (Churchman, 1971). In addition, following the recent methodological pluralism of Feyerabend (1975; 1981), construct validation studies have argued for exploration and alternative explanations rather than pure hypothesis testing (for example, Cooper and Regan, 1982; Kyllonen, Lohman and Woltz, 1984).

The second proposal in construct validation is one that deals with generalizability, especially as a result of the important theme of population heterogeneity within construct validation research that has recently emerged. Educational and language testing researchers have typically dealt with this issue as part of differential item functioning (DIF) studies or test bias studies, seeing the issue primarily as a problem of cultural fairness or bias in tests (for example, Briere and Brown, 1971; Briere, 1968; 1973; Chen and Henning, 1978; Berk, 1982; Zeidner, 1986; Oltman, Stricker and Barrows, 1988; Kunnan, 1990; Ryan and Bachman, 1992; Holland and Wainer, 1993). But, this issue needs to be considered more broadly. Laosa (1991: 6) argues that since '... population generalizability is a pivotal dimension of construct validity', researchers

need to '... establish *generalizability boundaries* that accurately demarcate the populations to which the accumulated empirical data permit generalizing'. Support for this position has also come from the psychometric perspective (see Bentler, 1986; Muthen, 1989). Muthen (1989: 558) argues that

> ... the homogeneity assumption is unrealistic when applied to a sample of students with varying instructional backgrounds. A good example is modelling of mathematics achievement for U.S. eighth grade students, where widely varying curricula or tracks are being followed.

Muthen provides additional examples from survey research in which he notes that '... the validity and reliability of certain items can be expected to vary across subgroups defined by race, gender, region, and issue salience' (p. 558) and psychiatric epidemiology where '... data come from a mixture of "normal" and "abnormal" subjects' (p. 558).

It seems imperative that construct validation research should be 1) designed so that construct validation models can capture relationships among constructs; and 2) designed so that multiple group models can be used to estimate fit to data.

Several psychometric models have been proposed for construct validation research that would accomplish the above two emphases. Overall, in order to establish the crucial relationship between construct representation and nomothetic span, Embretson (1985: 196) notes that '... correlational research in which measures of individual differences in the underlying cognitive variables, test scores, and external measures are obtained' is the way to proceed. Within this approach, there are several different proposed models (Embretson, 1985; Muthen, 1989).[1] Muthen (1989) discusses several models including regular multiple-group simultaneous structural modelling in the tradition of Joreskog (1978), the MIMIC model (multiple indicators, multiple causes) and multilevel analysis. Within regard to the most popular model, the regular multiple-group simultaneous structural modelling, Muthen (1989: 563) cautions researchers

> ... there are two important requirements of such an analysis that are not fulfilled. One is that sizeable samples are available for each group. We would want enough observations in each group to be able to compute stable correlation coefficients and covariances.

[1] Embretson (1985: 195) presents several component latent trait models (CLTM) which '... provide estimates of the cognitive demands in each item and specify the relationship of cognitive demands to the cognitive abilities that are reflected in item solving'. These models are not discussed here as latent trait modelling is outside the scope of this article.

1 Modelling in SLA and language testing

Two of the fields in applied linguistics that have used structural modelling techniques are second language acquisition (SLA) and language testing. Modelling in SLA, based largely on Gardner's (1985; 1988) socioeducational model of second language learning, has typically focused on individual differences in language learning, such as language aptitude, attitude and motivation, attrition, formal and informal language acquisition contexts, intelligence, cognitive style, monitoring and situational anxiety. Some of the studies that set out to investigate the causality of SLA include Gardner, Lalonde and Pierson, (1983), Nelson, Lomax and Perlman (1984), Clement and Krudenier (1985) and Gardner *et al.* (1987). Other studies that have used single latent variables have included language acquisition and age (Snow and Hoefnagle-Hohle, 1978), pronunciation and empathy (Guiora *et al.*, 1975), and achievement and classroom input (Hamayan and Tucker, 1980). While these research studies have enriched our understanding of the factors that influence SLA, Gardner's model and his studies (and the work of his associates) have been criticized for discrepancies in empirical findings (Oller, 1982), measures of language achievement (Bachman, 1988) and causal modelling issues (Au, 1988). More recent modelling studies (for example, Ecob, 1987; Hill, 1987; Turner, 1989; Sasaki, 1991) have improved on Gardner's model as well as included more complex modelling techniques.

Modelling studies in language testing, on the other hand, have primarily focused on the nature of second language proficiency and construct validation of tests, despite Upshur's (1983:99) excellent argument for measuring individual differences (MID) '... in the search for explanations of natural language, its attainment, its use, and its effects'. Internal structure analyses of tests have been conducted to investigate Oller's (1983) claim of a 'unitary trait hypothesis' (see Oller and Hinofotis, 1980; Swinton and Powers, 1980; Bachman and Palmer, 1981; 1982; Upshur and Homburg, 1983; Vollmer, 1983; Vollmer and Sang, 1983; Sang *et al.*, 1986; Boldt, 1988). These studies as well as specific studies on oral communication (Hinofotis, 1983), pronunciation (Purcell, 1983) and the FSI oral interview (Bachman and Palmer, 1981) have confirmed that language proficiency is multicomponential and not unitary as proposed by Oller (1983). A few recent studies have investigated the relationships among test-taker characteristics or background variables, such as field independence and test performance (for example, Stansfield and Hansen, 1983; Hansen and Stansfield, 1984; Chapelle, 1988).

Most of the above-mentioned studies used exploratory factor analysis or the multitrait-multimethod approach with test performance data. Confirmatory factor analysis was used less commonly to confirm or reject causal models of language proficiency (examples include Bachman and Palmer, 1982; Hale, Rock and Jirele, 1989). None of the above studies, however, were designed to investigate the construct validation of tests' score use or interpretation by exploring structural relationships among latent variables that include both test-taker characteristics (TTCs) and test performance. In fact, it was only recently that Bachman (1990) effectively argued that TTCs are one of the three critical groups of factors that influence test performance and, therefore, affect validity and reliability of the tests in question – the other two being communicative language ability and test method. His framework (1990) provides an excellent starting point for empirical investigations of structural relationships among TTCs and test performance.

2 *Test-Taker characteristics that influence test performance*

Test-taker characteristics (TTCs) or background characteristics are one of the factors that influence test performance. These TTCs include personal characteristics or attributes such as age, native language and culture, ethnicity and gender; educational characteristics such as background knowledge, previous instruction and opportunity to learn the target language; as well as cognitive, psychological and social characteristics, such as learning strategies and styles, attitude and motivation, aptitude, intelligence, anxiety, personality, and field dependence-independence, extroversion and introversion.[2]

This study investigated the influence of the two major TTCs on EFL test performance: 1) social milieu or cultural background, used here broadly to include native language, culture and ethnicity of the test-takers; and 2) exposure or previous instruction as opportunity to learn the target language (English, in this case) separated into formal instruction and informal language-learning exposure contexts in the learners' home countries as well as in the English-speaking countries they visited, and Krashen's (1982) concept of self-monitoring by test-takers of their speaking and writing. These TTCs were designed as independent variables and EFL test performance (TP) factors as dependent variables.

[2] Skehan (1989) considers these characteristics as individual difference variables, while Spolsky (1989) characterizes them as conditions for second language learning in a preference model.

II Method

1 Purpose

The purpose of this study was to investigate the influences of the TTCs mentioned earlier on EFL test performance through structural modelling. Operationally, several models of the structure of TTCs and EFL test performance and the influences of the TTCs on EFL test performance were posited. These models were based on substantive theories in language testing (Bachman, 1990) and second language acquisition (Gardner, 1985), and on preliminary exploratory factor analyses. However, though the approach used was structural modelling, this study was exploratory in mode. It followed, in the words of Cronbach (1989), *not* a strong programme of construct validation which involves formal hypothesis testing but a weak programme of construct validation which involves widespread support for explanations from many perspectives. Cronbach (1989:16) notes that '... the investigation should aim to illuminate the test and the related construction so that persons making decisions see more clearly how to use the test, and those pursuing research know where the greatest perplexities lie.'

2 Research questions

The general research question investigated was: What influences do home-country formal instruction (HCF), home-country informal exposure or instruction (HCI), English-speaking country exposure or instruction (ESC) and monitoring (MON) have on the four EFL test performance factors? If there are influences, to what extent do they differ for the two native language groups?

3 Sample

Data from the Cambridge-TOEFL comparability study (Bachman *et al.*, 1991), which included 1448 subjects from eight sites in eight countries, were used: Bangkok, 169; Cairo, 89; Osaka, 189; Hong Kong, 196; Madrid, 196; Sao Paulo, 207; Toulouse, 197; and Zurich, 205. Descriptive information collected from all the subjects through the background questionnaire was as follows: the majority of the subjects were enrolled either as students, at the secondary school level (21.3%), or at the college level (full time, 27.6%; part time, 10.45) or in a language institute or other English course (17%), while 23.7% were not enrolled as students. The median age was 21, with the youngest test-taker 14 years of age and the eldest 58, and slightly over half (59.4%) were female.

4 Instruments

The instruments used were 1) a 45-item Likert scale background questionnaire which collected responses regarding some of the test-taker characteristics, such as previous instruction or exposure to English, and the use of monitoring; and 2) two EFL test batteries: the First Certificate in English (FCE), administered by the University of Cambridge Local Examinations Syndicate, the TOEFL and the SPEAK, administered by the Educational Testing Service, Princeton, and the Test of English Writing, a TWE-like test developed for the Cambridge-TOEFL comparability study (Bachman *et al.*, 1991). Responses to the questionnaire items and the EFL test scores were used to establish latent variables which could be used in modelling TTCs and test performance data.

5 Structural modelling approach

As discussed earlier, there are several psychometric models that can be used. Separate multiple-group structural modelling was used instead of the simultaneous multiple-group structural modelling (and the MIMIC and the multilevel models) because the chief interest in this study was to investigate the different structural relationships for the two population groups, not comparisons between the two populations groups for similar models which the simultaneous multiple-group modelling would have provided. Another reason for this choice was the fact that there was no previously established structural model for these variables and for these population groups to fall back on (and confirm).

Thus, two different population groups based on native language were examined. They were the non-Indo-European (NIE; = 380) native language group (Thai, Arabic, Japanese, Chinese) and the Indo-European (IE; = 605) native language group (Spanish, Portuguese, French, German). This was done in order to explore whether the sociocultural and educational differences between the two groups would result in different structural relationships between the TTCs and EFL test performance.

6 Refining models

Two procedures in EQS that helped refine the models by evaluating the parameters that were being estimated were the Lagrange multiplier test and the Wald test (see Bentler, 1989, for details). However, these tests were used only as recommendations and not followed blindly because often the recommendations did not have the support of substantive theory.

7 Assessment of fit and the evaluation of models

Assessment or 'test of fit' of a model, as Cuttance (1987: 256) states, '... refers to parametric statistical tests ... those based on a particular statistical distribution', and *evaluation of the model* '... refers to measures of the methodological validity of a model'. The parametric tests used for the assessment of fit of models were 1) the X^2 statistic for the specified model against the unconstrained or null model; 2) the X^2 ratio which was suggested by Wheaton *et al.* (1977) as a way of dealing with the effect of large sample size on the X^2 statistic; 3) the Bentler–Bonett normed fit index (BBNFI) and the Bentler–Bonett non-normed fit index (BBNNFI); 4) the comparative fit index (CFI)[3]; and 5) the Satorra and Bentler scaled test statistic (SBX^2) developed by Satorra and Bentler (1988a; 1988b) that is computed as part of the robust statistics. Cuttance's (1987) methods for assessing the methodological validity of models were also followed by inspecting the parameter estimates, construct loadings, *t*-test values and the standardized residual matrix.

8 Statistical software and estimation method

SPSSX was used for exploratory factor analysis and EQS 3.0 version (Bentler, 1989) for structural modelling. The specific estimation method used was maximum likelihood (ML) and ML with ROBUST. ML is used when the 'normal theory' assumption that variables are multivariate normally distributed is met. Chou, Bentler and Satorra (1989) have shown that robust statistics are more trustworthy than ordinary statistics.

III Results

1 Distributions and reliabilities

Table 1 presents the names, labels and descriptions of variables and constructs. Means, standard deviations and internal consistency reliabilities are presented in Table 2. Coefficients of skewness and kurtosis for all the 25 variables were in the range −0.6 to 0.6 for the NIE group and the −0.8 to 0.7 for the IE group, indicating that the distribution was close to normal and suggesting that the maximum likelihood with robust estimation procedure would be appropriate. Internal consistency reliability estimates for the TTCs ranged from 0.68 to 0.75 and for the EFL tests from 0.79 to 0.97. Correlations for the variables for both groups are presented in the Appendix.

[3] This index developed by Bentler (1990) avoids the underestimation of fit sometimes noted for the BBNFI in small samples.

a Modelling test-taker characteristics (TTCs): A four-factor structural model for 12 observed variables was designed for multiple-group data – that is, for both the NIE and the IE groups.[4] These constructs represented the four TTCs with paths from each of them to each of the three observed variables associated with the constructs. The model included correlations among HCF and ESC, HCF and MON, ESC and MON. Table 3 presents the results of the

Table 1 Names, labels and descriptions of variables and constructs

Variable/ factor	Label	Description
V1	BQ07	Item 7: Home country formal instruction
V2	BQ09	Item 9: Home country formal instruction
V3	BQ10	Item 10: Home country formal instruction
V4	BQ14	Item 14: Home country informal instruction/exposure
V5	BQ15	Item 15: Home country informal instruction/exposure
V6	BQ17	Item 17: Home country informal instruction/exposure
V7	BQ20	Item 20: English-speaking country instruction/exposure
V8	BQ23	Item 23: English-speaking country instruction/exposure
V9	BQ28	Item 28: English-speaking country instruction/exposure
V10	BQ39	Item 39: Monitoring
V11	BQ40	Item 40: Monitoring
V12	BQ41	Item 41: Monitoring
V13	FCE1	FCE paper 1 – reading comprehension
V14	FCE2	FCE paper 2 – composition
V15	FCE3	FCE paper 3 – use of English
V16	TFL2	TOEFL section 2 – structure and written expression
V17	TFL3	TOEFL section 3 – vocabulary and reading comprehension
V18	TEW	Test of English Writing
V19	FCE4	FCE paper 4 – listening comprehension
V20	FCE5	FCE paper 5 – face-to-face oral interview
V21	TFL1	TOEFL section 1 – listening comprehension
V22	SPCO	SPEAK comprehensibility
V23	SPGR	SPEAK grammar
V24	SPPR	SPEAK pronunciation
V25	SPFL	SPEAK fluency
F1	HCF	Home country formal instruction
F2	HCI	Home country informal instruction/exposure
F3	ESC	English-speaking country instruction/exposure
F4	MON	Monitoring
F5	RW1	Reading-writing factor 1 – FCE papers
F6	RW2	Reading-writing factor 2 – TOEFL papers
F7	LS1	Listening-speaking factor 1 – interactional
F8	LS2	Listening-speaking factor 2 – noninteractional
F9	G	General factor

Note also:
E1 to E25 are associated with V1 to V25; and D4 or D5 to D8 (and D9) are associated with F4 or F5 to F8 (and F9).

[4] Only the 12 best observed variables from the group of 45 were chosen for use in these analyses. The others were dropped because of very low internal consistency reliability estimates.

modelling for the two groups. Judging from the X^2 statistic, the model barely fits for the NIE group ($p < 0.57$) and does not fit for the IE group ($p < .001$). But, since the other fit indices indicated a good fit for both groups and this supplemented the EFA results for the single group, it was decided to use this model in the modelling of TTCs and TPs later.

b Modelling test performances (TP): A correlated four-factor model was attempted for the 13 observed variables for the two groups. The motivation for this model came from the EFA results from Bachman *et al.* (1991) which suggested a higher-order general factor and four first-order factors. Moreover, modelling in language testing research (Bachman and Palmer, 1981) has shown that when a higher-order factor with first-order factors is the best explanation, a correlated first-order factor (without the higher-order factor) solution is both mathematically equivalent and practically not an

Table 2 Descriptives and internal consistency reliabilities for all variables

	NIE group			IE group			Both groups
Variable	Mean	SD	Alpha	Mean	SD	Alpha	
BQ07	4.35	1.09		2.99*	1.00		
BQ09	2.25	.97		1.63*	.63		
BQ10	3.15	.78	0.53	2.04*	.58	0.53	0.70
BQ14	1.43	.96		1.38	.83		
BQ15	.92	1.76		1.41*	1.96		
BQ17	1.27	.55	0.78	1.31	.54	0.62	0.73
BQ20	.35	.56		.54*	.56		
BQ23	.45	1.33		1.31*	1.99		
BQ28	1.23	.56	0.78	1.48*	.61	0.73	0.75
BQ39	2.36	.70		2.70*	.71		
BQ40	2.25	.73		2.49*	.70		
BQ41	2.76	.83	0.61	2.94*	.71	0.64	0.68
FCE1	24.68	4.92		27.48*	4.06		0.79
FCE2	23.20	5.55		26.02*	4.83		NA
FCE3	23.37	5.75		26.76*	4.27		0.85
FCE4	12.79	3.28		14.69*	2.41		0.62
FCE5	25.99	5.72		28.55*	4.90		NA
TFL1	49.11	6.23		51.32	5.98		0.89
TFL2	50.42	6.56		52.81	6.21		0.83
TFL3	49.89	6.41		53.96*	4.97		0.87
TEW	3.65	.89		4.10*	.77		0.90
SPCO	194.40	38.79		208.96	36.62		0.97
SPGR	1.85	.44		2.00	.42		
SPPR	2.00	.33		2.25	.30		
SPFL	1.85	.42		2.04	.39		

Note:
*Significant differences between the two groups at $p < .001$.

Table 3 Goodness-of-fit indices for modelling TTCs for both NIE and IE groups

Index	NIE	IE
X^2	67.90	190.82
df	51	51
$p<$	0.057	0.001
X^2/df	1.33	3.74
SB X^2	67.85	181.18
BBNFI	0.95	0.89
BBNNFI	0.98	0.90
CFI	0.99	0.92

Notes:
SB X^2 = Sartorra–Bentler scaled X^2; BBNFI = Bentler–Bonett normed fit index; BBNNFI = Bentler–Bonett non-normed fit index; CFI = comparative fit index.

unsatisfactory explanation either. Modelling test performance using this approach was successful. Table 4 presents the results of the modelling for the two groups. Judging only from the X^2 and the X^2 statistics, the model was not a very good one. However, the other fit indices, especially the CFI, indicated that the model fits quite well for both the NIE and the IE groups.

c Modelling relationships among TTCs and TP factors: After the above analyses of the TTCs and the TP factors were done, it was decided to model the data with two structural models with 12 TTCs (with four constructs or factors) and 13 TP variables (with five constructs or factors). The two models were model 1, in which the four test-taker characteristic factors (HCF, HCI, ESC and MON) were designed to have equal influences on the four TP factors (RW1, RW2, LS1 and LS2); and model 2 was designed to have the exposure factors (HCF, HCI and ESC) influence MON (an intervening factor) which in turn influenced the four TP factors.

Table 4 Goodness-of-fit indices for modelling test performance for both NIE and IE groups

Index	NIE	IE
X^2	221.70	205.66
df	59	59
$p<$	0.001	0.001
X^2/df	3.76	3.49
SB X^2	220.78	202.45
BBNFI	0.94	0.95
BBNNFI	0.94	0.95
CFI	0.95	0.97

Table 5 Goodness-of-fit indices for models 1 for both groups

Index	NIE	IE
X^2	577.66	767.12
df	258	258
$p<$	0.001	0.001
X^2/df	2.24	2.97
SB X^2	568.07	754.38
BBNFI	0.89	0.88
BBNNFI	0.93	0.90
CFI	0.94	0.92

2 *Model 1*

This model had 25 variables grouped into four independent factors, HCF, HCI, ESC and MON, and four dependent factors, RW1, RW2, LS1 and LS2. Table 5 presents the goodness-of-fit indices for the best fitting models for both groups. While the X^2 ratio for both groups was relatively high, the NIE group was better (2.24). This indicated that the model for the NIE group fits the data for that group better than did the model for the IE group.

Examination of standardized path coefficients for paths between independent factors and dependent factors presented in Table 6 and Figures 1 and 2 showed differences between the two groups: two path estimates for each of the two groups did not have counterparts in the other group as they were dropped based on the Wald test. In the NIE group, the influence of MON on RW1 is .223 but the corresponding path for the IE group was not significant. Similarly, for the IE group, the influence of ESC on RW1 was .217 while the corresponding path for the NIE group was not significant. Other

Table 6 Model 1: standardized path coefficients for paths between factors for both NIE and IE groups

	HCF	HCI	ESC	MON
RW1				**.223****
	−.163**		.217**	
RW2		**−.128****		**−.243****
		−.127**		
LS1	**.135****	**.082***	**.231****	
	−.348**	.130**	.226**	
LS2	**−.105***	**.126****	**.176****	
	−.191**		.227**	

Notes:
Blank space indicates influence was not estimated or significant; NIE group estimates are in the first line (in **bold**) and IE group estimates are in the second line; estimates with two asterisks are significant at $p<01$ or $t>2.58$ and estimates with one asterisk are significant at $p<.05$ or $t>1.96$; the disturbances of dependent factors were correlated but are not shown here.

noticeable differences between the two groups included the following: 1) the influence of HCF on LS1 (.135 for the NIE group and −.348 for the IE group); and 2) the influence of MON on RW2 (−.243 for the NIE group and .047 for the IE group). However, there was one influence that was of comparable strength for both groups: the influence of ESC on LS1 (.231 for the NIE group and .226 for the IE group).

From the results presented above for model 1, it was apparent that the models did not produce either a clear overall statistical fit or lack of fit for both groups. Thus it was decided to evaluate model 2, which was theoretically a more interesting model as it followed Gardner's (1985) socioeducational model.

3 Model 2

In this model, there were 25 variables grouped into three kinds of constructs: HCF, HCI and ESC (previous exposure to English); MON (monitoring), and RW1, RW2, LS1 and LS2 (test performance factors). In this model, HCF, HCI and ESC were independent factors, MON was a dependent and an intervening factor, and RW1, RW2, LS1 and LS2 were dependent factors. The paths, therefore, from the independent factors to the dependent factors could be

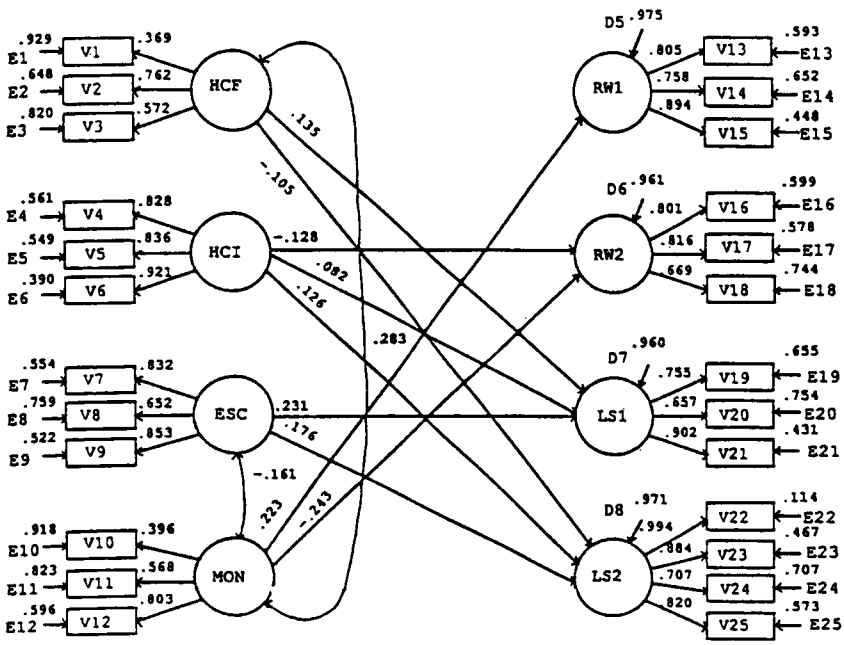

Figure 1 Model 1: standardized estimates for paths for the NIE group

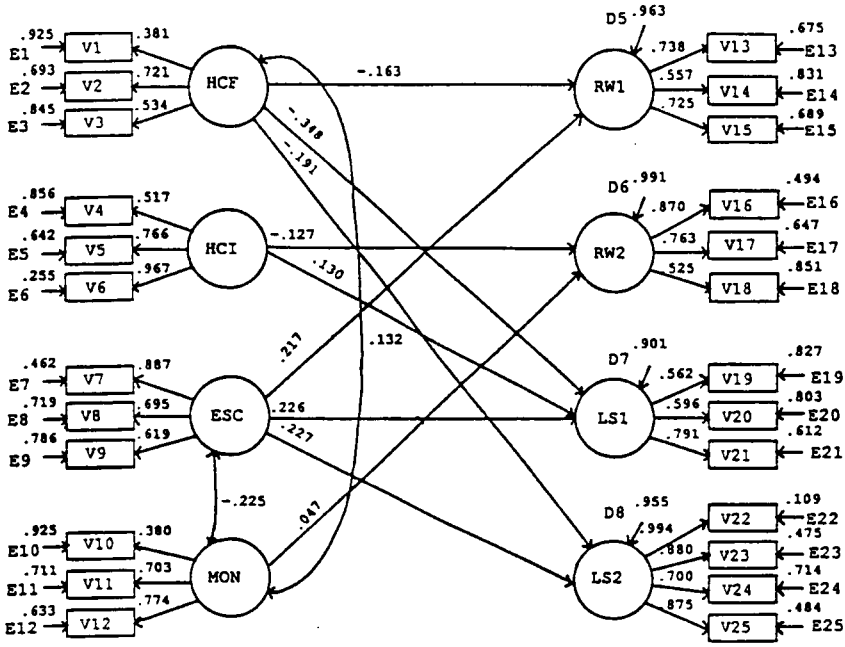

Figure 2 Model 1: standardized estimates for paths for the NIE group

drawn directly to them or through the intervening factor providing for both direct and/or indirect influences.

Table 7 presents the goodness-of-fit indices for the best fitting models for both groups. Though for both groups the X^2 ratio was not clearly acceptable, the other fit indices were almost the same for the groups as those for model 1.

Direct (D), indirect (I) and total (T) influences of the exposure factors (HCF, HCI and ESC) on the intervening factor (MON), and the four exposure factors directly on the dependent factors (RW1, RW2, LS1 and LS2) for both groups, are presented in Figures 3 and 4 for the NIE and the IE groups respectively.

Table 7 Model 2: goodness-of-fit indices for both groups

Index	NIE	IE
X^2	556.38	779.64
df	257	259
$p<$	0.001	0.001
X^2/df	2.17	3.01
SB X^2	398.77	558.78
BBNFI	0.90	0.88
BBNNFI	**0.93**	**0.90**
CFI	**0.94**	**0.92**

As presented in Table 8 and Figure 3, for the NIE group, each of the exposure factors had direct (D) influences on at least two of the test performance factors. The exposure factor with the strongest direct influence was HCF with direct influence on both of the reading-writing test factors: RW1 (.226) and RW2 (.289); on one of the two listening-speaking test factors: LS1 (.319); and on MON (.266). HCI had moderate direct influence on both the listening-speaking test factors: LS1 (.105) and LS2 (.141), while ESC had direct influence on both the listening-speaking factors: LS1 (.246) and LS2 (.164) but negative influence on MON (−.142). MON had direct influence on both the reading-writing factors: RW1 (.186) and RW2 (.183). Indirect (I) influences which were the products (or sums of products) of direct influences were few in number and generally weak, as shown in Table 8: HCF on RW1 (.050) and RW2 (.049) and ESC on RW1 (−.026) and RW2 (.026). In terms of total (T) influence, there were at least five noteworthy positive influences: HCF on RW2 (.338), on LS1 (.319), on RW1 (.276) and on MON (.266); and ESC on LS1 (.246). Other notable influences were MON on RW1 (.186) and on RW2 (.183).

Table 8 Model 2: direct (D), indirect (I) and total (T) influences on test performances for the NIE group

	HCF	HCI	ESC	MON
RW1				
D	.226**			.186**
I	.050*		−.026	
T	.276**		−.026	.186**
RW2				
D	.289**			.183**
I	.049		−.026	
T	.338**		−.026	.183**
LS1				
D	.319**	.105	.246**	
I				
T	.319**	.105	.246**	
LS2				
D		.141	.164**	
I				
T		.141	.164**	
MON				
D	.266**		−.142*	
I				
T	.266**		−.142*	

Notes:
Blank space indicates influence was not estimated or significant; estimates with two asterisks are significant at $p < 01$ or $t > 2.58$ and estimates with one asterisk are significant at $p < .05$ or $t > 1.96$; all other influences are not significant; the disturbances of dependent factors were correlated but are not shown here

As presented in Table 9 and Figure 4, ESC had direct (D) influence on the three test performance factors: RW1 (.221), LS1 (.216) and LS2 (.212). HCF and HCI, on the other hand, had direct negative influence on three test performance factors: HCF on LS1 (−.205), HCI on RW1 (−.144) and LS2 (−.068). MON was directly influenced differently by HCF and ESC: HCF had a positive direct influence (.160) while ESC had a direct negative one (−.228). MON had weak direct influence on two test performance factors: RW1 (.106) and RW2 (.108). Indirect (I) influences were once again few in number and generally weak: HCF on RW1 (.017), and ESC on RW1 (−.024) and RW2 (−.025). In terms of total (T) influence, there were five notable influences: ESC on MON (−.228), on LS1 (.216), on LS2 (.212); HCF on LS1 (−.205) and ESC on RW1 (.197).

Once again, from the results presented above for model 2, it was apparent that the models did not produce either a clear overall statistical fit or lack of fit for both groups. There was some improvement in model fit from model 1 for the NIE group and a worse fit from model 1 for the IE group. Again, however the comparative fit indices (0.94 for the NIE group and 0.92 for the IE

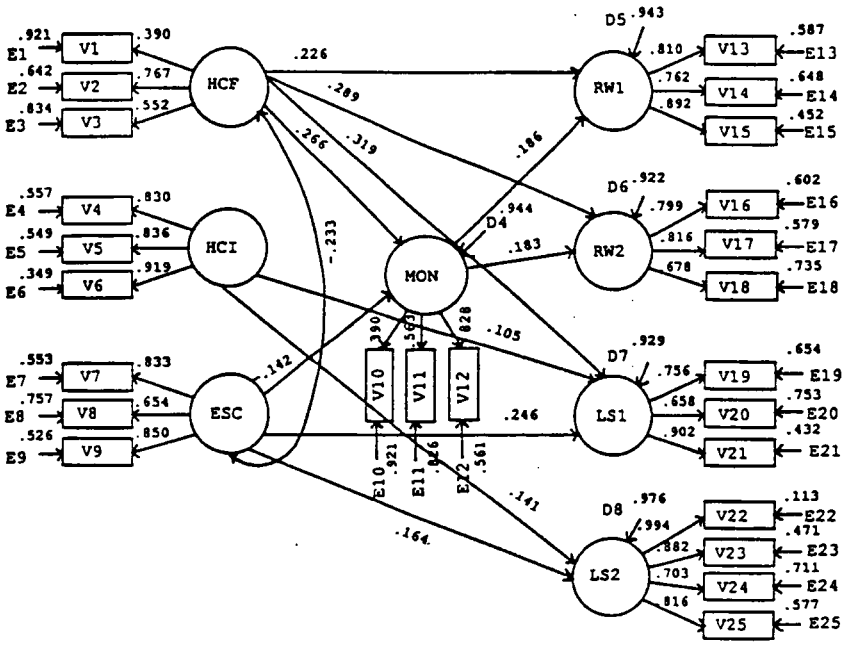

Figure 3 Model 2: influences of test-taker characteristics on test performance factors for the NIE group

group) showed that the models were quite good.

IV Discussion

1 Home-country formal instruction (HCF)

The influence of HCF instruction on the TP factors in model 1 for both groups was not substantial, except for a fairly strong negative influence on LS2. In model 2 for the NIE group, however, its influence was substantial on three of the test performance factors: RW1, RW2 and LS1. These three factors included the FCE, the TEW and the TOEFL, but did not include the SPEAK, which makes up the LS2 factor. This shows that HCF instruction was an important factor influencing the performance of NIE test-takers on the FCE, the TEW and the TOEFL but not on the SPEAK. A possible explanation could be that since the SPEAK is a non-interactional tape-mediated, speeded, oral test, performance on this test may not benefit very much from formal instruction. In addition, NIE test-takers seem to have benefited from HCF instruction, the major source of instruction for this group.

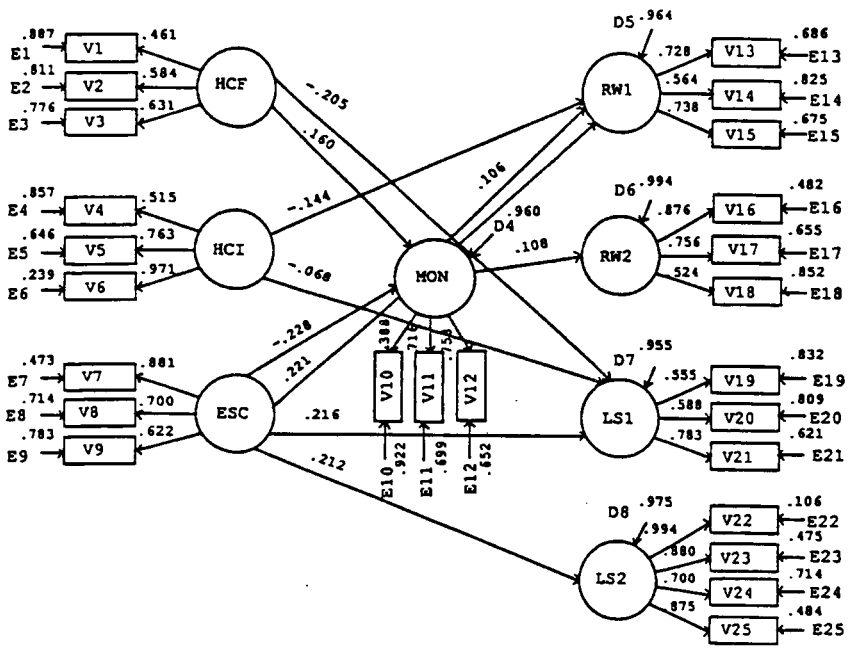

Figure 4 Model 2: influences of test-taker characteristics on test performance factors for the IE group

2 Home-country informal exposure (HCI)

HCI exposure did not have as much influence on test performance as HCF instruction for either group across all models. Mode 1 influences were weak; in model 2, for the NIE group, HCI exposure showed a significant though moderate influence on both the listening-speaking factors. For the IE group, HCI exposure showed negative though moderate influences on two TP factors, RW1 and LS2. HCI exposure influence on both the listening-speaking factors for the NIE group is an interesting result. It indicates that informal exposure to English was influential in the performance of NIE test-takers on these two groups of tests: the listening and speaking test parts of the TOEFL and the FCE (which are interactional) and the SPEAK test (which is noninteractional).

3 English-speaking country instruction or exposure (ESC)

Although HCF instruction and HCI exposure did have substantial and moderate influences on the TP factors, neither of them seemed to be as substantive as English – speaking country exposure

Table 9 Model 2: direct (D), indirect (I) and total (T) influences on test performance factors for the IE group

	HCF	HCI	ESC	MON
RW1				
D		−.144**	.221**	.106*
I			−.024	
T		−.144**	.197**	.106*
RW2				
D				.108*
I			−.025*	
T			−.025*	.108*
LS1				
D	−.205**		.216**	
I				
T	−.205**		.216**	
LS2				
D		−.068*	.212**	
I				
T		−.068*	.212**	
MON				
D	.160*		−.228**	
I				
T	.160*		−.228**	

Notes:
Blank space indicates influence was not estimated or significant; estimates with two asterisks are significant at $p<01$ or $t>2.58$ and estimates with one asterisk are significant at $p<.05$ or $t>1.96$; all other influences are not significant; the disturbances of dependent factors were correlated but are not shown here.

influence. ESC instruction or exposure showed substantial positive influences in model 1 and model 2 for both groups on LS1 and LS2. In model 2, for both the NIE and the IE group, there were substantial positive influences on LS1 and LS2 and a not too surprising negative influence on monitoring. For the IE group, there was a moderate influence on RW1 as well.

4 Influence of and on monitoring (MON)

MON had a substantial influence on TP factors for both groups. In model 1 for both groups, monitoring had a strong positive influence on RW1 and a negative influence on RW2. This might have been due to the differences in test methods used by the two RW factors. But in model 2, which presented monitoring as an intervening factor, HCF instruction had a strong influence on monitoring and monitoring had a moderate influence on RW1 and RW2 for both groups (though more for the NIE group). This result indicates that monitoring was related to HCF instruction, suggesting that learners who have formal instruction, such as the NIE group, could be the ones who strive more for correctness and, therefore, monitor more than those who are more likely to have had more informal learning, and are less concerned with correctness, like the IE group. In addition, it was also interesting to note that monitoring moderately influenced both the reading and writing tests, since both these tests provide enough time to respond.

5 EFL test performance

The four EFL TP factors were modelled as correlated dependent factors in a skills components model of language proficiency. This structure did not collapse in any of the modelling. The first two factors, the reading-writing factors, could be distinguished by the fact that the variables that made up RW1 were FCE papers 1, 2 and 3 (an FCE written mode); and the variables that made up RW2 were TOEFL sections 2 and 3, and the TEW (an ETS written mode). The listening-speaking factors could be distinguished too: the variables that made up LS1 were FCE papers 4 and 5, and TOEFL section 1 (an interactional mode); and the variables that made up LS2 were the SPEAK scores for pronunciation, fluency, grammar and comprehensibility (a noninteractional mode). The robustness of this four – factor structure across models and groups provides evidence that there was a significant difference between the FCE and the ETS reading-writing sections and between the two listening-speaking tests: the FCE papers (interactional) and SPEAK (noninterac-

tional). A reasonable explanation for the distinctiveness of RW1 and RW2 and LS1 and LS2 could be that differences in test methods as well as differences in language skills measured were responsible for the differences.

6 Model comparisons

While model 1 represents the view that previous instruction and exposure to English and self-report of monitoring are equal in status as influences on test performances, model 2 represents Gardner's (1985) intervening factors view, though he does not include monitoring in any of his studies. These two models are conceptually different but the results did not show up this difference. For example, in terms of the X^2/df ratios, for the NIE group, model 1 was 2.24 and model 2 was 2.17, and for the IE group, model 1 was 2.97 and model 2 was 3.01. In terms of the CFI, one of the most robust goodness-of-fit indices (Bentler, 1990), for the NIE group, for both models 1 and 2, the value was 0.94, and for the IE group, for both models 1 and 2, the value was 0.92. The only statistical difference between the two groups was that while the fit indices showed that model 2 was slightly better than model 1 for the NIE group, it was the other way around for the IE group. But since the improvement of fit for the NIE group from model 1 to 2 and the degradation of fit for the IE group from model 1 to 2 was so small, it was unclear whether the difference in models was significant.

V Implications and conclusions

This construct validation study through structural modelling provided a unique opportunity to explore the dynamic and complex network of structural relationships among some TTCs and EFL test performance for two primary reasons. First, though only two major TTC factors were used in these analyses, it was evident that this approach uncovered more information about the relationships of those factors to EFL test performance than would have been possible if these factors were treated individually, or if any other procedure was used. Secondly, the two native language groups were modelled separately so that the native languages and cultures, and the opportunity to learn English in those two contexts (NIE and IE), could be examined separately. Research with additional factors, such as gender, age, attitude and motivation, learning strategies and styles, to name a few, could provide fuller descriptions. But since not all of these and other TTCs (personal attributes, educational, cognitive, psychological and social characteristics) will have sig-

nificant influence on test performance, the challenge for language testing researchers is to identify the TTCs that influence test performance, and then to model those TTCs with test performance in a network fashion to arrive at a model that could explain the major influences on test performance. A theory of construct validation that includes both content representation and nomothetic span could then emerge.

To conclude, Upshur (1982: 119) notes that '. . . measurement of individual differences has potential for more direct contributions to theory development in the language sciences'. He provides three different aspects for researchers considering the measurement of individual differences and explanation in the language sciences: 'establishing a research agenda, elaborating variables, and evaluating theoretical models' (p. 119). Three challenges within these aspects should be noted by researchers attempting structural modelling: incompleteness of structural models (difficulty in knowing whether a model is complete or not), undecidability of best model from available models (difficulty in deciding which and when a model is superior), and inaccuracy in measurement of variables (difficulty in accurately measuring variables, though they may be precisely measured). Finally, since data from the human and language sciences (including language learning and testing) tends to have a great deal of complexity and uncertainty (West and Salk, 1987), structural models like the ones discussed here may only be scratching at the surface of the complexity. Perhaps a complex systems analysis, following the example of Pena-Taveras and Cambel (1989), may be required for theory development in the language sciences.

VI References

Au, S.Y. 1988: A critical appraisal of Gardner's social-psychological theory of second-language (L2) learning. *Language Learning* 38, 75–100.

Bachman, L.F. 1988: Language testing–SLA research interfaces. *Annual Review of applied Linguistics* 9, 193–209.

——1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F., Davidson, F., Ryan, K. and **Choi, I.-C.** 1991: An investigation into the comparability of two tests of English as a foreign language: the Cambridge-TOEFL comparability study. Final report. Author.

Bachman, L.F. and **Palmer, A.S.** 1981: The construct validation of the FSI oral interview. *Language Learning* 31, 67–86.

——1982: The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–65.

Bechtold, H.P. 1959: Construct validity: a critique. *American Psychologist* 14, 619–29.

Bentler, P.M. 1978: The interdependence of theory, methodology, and empirical data: causal modeling as an approach to construct validation. In Kandel, D.B., editor, *Longitudinal research on drug use.* Washington, DC: Hemisphere Publishing, 267–302.

——1986: Structural modelling and *Psychometrika*: an historical perspective on growth and achievements. *Psychometrika* 51, 35–51.

——1989: *EQS: structural equations program manual.* Los Angeles, CA: BMDP Statistical Software.

——1990: Comparative fit indexes in structural models. *Psychological Bulletin* 107, 238–46.

Berk, R., editor, 1982: *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

Boldt, R.F. 1988: *Latent structure analysis of the TOEFL. TOEFL Research Report 28.* Princeton, NJ: Educational Testing Service.

Boomsma, A. 1987: The robustness of maximum likelihood estimation in structural equation models. In Cuttance, P. and Ecob, R., editors, *Structural modelling by example.* Cambridge: Cambridge University Press, 160–88.

Briere, E.J. 1968: Testing ESL among Navajo children. In Upshur, J. and Fata, J., editors, Problems in foreign language testing. *Language Learning* 3, 11–21.

——1973: Cross cultural bias in language testing. In Oller, J. and Richards, J. editors, *Focus on the learner: pragmatic perspectives for the language teacher.* Rowley, MA: Newbury House, 214–27.

Briere, E.J. and **Brown, R.H.** 1971: Norming tests of ESL among Amerindian children. *TESOL Quarterly* 5, 327–4.

Campbell, D.T. and **Fiske, D.W.** 1959: Convergent and discriminant validity in the multitrait-multimethod matrix. *Psychological Bulletin* 56, 81–105.

Carroll, J.B. 1983: Psychometric theory and language testing. In Oller, J., editor, *Issues in language testing research.* Rowley, MA: Newbury House, 80–107.

Chapelle, C. 1988: Field independence: a source of language test variance? *Language Testing* 5, 62–68.

Chen, Z. and **Henning, G.** 1985: Linguistic and cultural bias in language proficiency tests. *Language Testing* 2, 155–63.

Chou, C.-P., Bentler, P.M. and **Satorra, A.** 1989: Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. Paper presented at the AERA, San Francisco, CA.

Churchman, C.W. 1971: *The design of inquiring systems: basic concepts of systems and organization.* New York: Basic Books.

Clement, R. and **Krudenier, B.G.** 1985: Aptitude, attitude and motivation in second language proficiency: a test of Clement's model. *Journal of Language and Social Psychology* 4, 21–37.

Cooper, L.A. and **Regan, D.T.** 1982: Intelligence, attention, and percep-

tion. In Sternberg, R.J., editor, *Handbook of human intelligence.* Cambridge: Cambridge University Press, 123–69.

Cronbach, L.J. 1989: Construct validation after thirty years. In Linn, R.L., editor, *Intelligence: measurement, theory, and public policy.* Urbana, IL: University of Illinois Press, 147–71.

Cronbach, L.J. and **Meehl, P.E.** 1955: Construct validity in psychological tests. *Psychological Bulletin* 52, 281–302.

Cuttance, P. 1987: Issues and problems in the applications of structural equation models. In Cuttance, P. and Ecob, R., editors, *Structural modelling by example.* Cambridge: Cambridge University Press, 241–79.

Ecob, R. 1987: Applications of structural equation modelling to longitudinal educational data. In Cuttance, P. and Ecob, R., editors, *Structural modelling by example.* Cambridge: Cambridge University Press, 138–59.

Ecob, R. and **Cuttance, P.** 1987: An overview of structural equation modelling. In Cuttance P. and Ecob, R., editors, *Structural modelling by example.* Cambridge: Cambridge University Press, 9–23.

Embretson, S.E. 1983: Construct validity: construct representation versus nomothetic span. *Psychological Bulletin* 93, 179–97.

——1985: Multicomponent latent trait models for test design. In Embretson, S. E., editor, *Test design: developments in psychology and psychometrics.* Orlando, FL: Academic Press, 195–218.

Gardner, R. 1985: *Social psychology and second language learning: the role of attitudes and motivation.* London: Edward Arnold.

——1988: The socio-educational model of second language learning: assumptions, findings and issues. *Language Learning* 38, 101–26.

Gardner, R., Lalonde, R.N., Moorcraft, R. and **Evers, F.T.** 1987: Second language attrition: the role of motivation and use. *Journal of Language and Social Psychology* 6, 1–47.

Gardner, R., Lalonde, R.N. and **Pierson, R.** 1983: The socioeducational model of second language acquisition: an investigation using LISREL causal modelling. *Journal of Language and Social Psychology*, 2, 1–15.

Guiora, A.Z., Paluszny, M., Beit-Hallahmi, B., Catford, J.C. and **Cooley, R.E.** 1975: Language and person: studies in language behaviour. *Language Learning* 25, 43–61.

Hale, G.A., Rock, D.A. and **Jirele, T.** 1989: *Confirmatory factor analysis of the Test of English as a Foreign Language. TOEFL Research Report 32.* Princeton, NJ: Educational Testing Service.

Hamayan, E.V. and **Tucker, G.R.** 1980: Language input in the bilingual classroom and its relationship to second language achievement. *TESOL Quarterly* 14, 453–68.

Hansen, L. and **Stansfield, C.** 1984: Field dependence-independence and language testing: evidence from six Pacific-Island cultures. *TESOL Quarterly* 18, 311–24.

Hill, P.W. 1987: Modelling the hierarchical structure of learning. In —— Cuttance, P. and Ecob, R., editors, *Structural modelling by example.*

Cambridge: Cambridge University Press, 65–85.

Hinofotis, F. 1983: The structure of oral communication in an educational environment: a comparison of factor analytical rotation procedures. In Oller, J., editor, *Issues in language testing research.* Rowley, MA: Newbury House, 170–87.

Holland, P.W. and **Wainer, H.**, editors, 1993: *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Joreskog, K.G. 1978: Structural analysis of covariance and correlational matrices. *Psychometrika* 43, 443–77.

Krashen, S.D. 1982: *Principles and practice in second language acquisition.* Oxford: Pergamon Press.

Kunnan, A.J. 1990: Differential item functioning and native language and gender groups: the case of an ESL placement examination. *TESOL Quarterly* 24, 741–46.

Kyllonen, P.C., Lohman, D.F. and **Woltz, D.** 1984: Componential modelling of alternative strategies for performing spatial tasks. *Journal of Educational Psychology* 76, 1325–45.

Laosa, L.M. 1991: *The cultural context of construct validity and the ethics of generalizability. Research Report 91–51.* Princeton, NJ: Educational Testing Service.

Loevinger, J. 1957: Objective tests as instruments of psychological theory. *Psychological Reports* 3, 635–94.

Messick, S. 1989: Validity. In Linn, R.L., editor, *Educational measurement* (3rd edn. New York: American Council on Education, 13–103.

Muthen, B. 1988: Some uses of structural equation modeling in validity studies: extending IRT to external variables. In Wainer, H. and Braun, H., editors, *Test validity*. Hillside, NJ: Lawrence Erlbaum Associates, 213–38.

——1989: Latent variable modelling in heterogeneous populations. *Psychometrika* 54, 557–85.

Nelson, F.H., Lomax, R.G. and **Perlman, R.** 1984: A structural equation model of second language acquisition of adult learners. *Journal of Experimental Education* 53, 29–39.

Oller, J.W. 1982: Gardner on affect: a reply to Gardner. *Language Learning* 32, 183–89.

——1983: Evidence for a general language proficiency factor: an expectancy grammar. In Oller, J., editor, *Issues in language testing research.* Rowley, MA: Newbury House, 3–10.

Oller, J.W. and **Hinofotis, F.** 1980: Two mutually exclusive hypotheses about second language ability: indivisible or partially divisible competence. In Oller, J. and Perkins, K., editors, *Research in language testing*. Rowley, MA: Newbury House, 13–23.

Oltman, P.K., Stricker, L.J. and **Barrows, T.** 1988: *Native language, English proficiency, and the structure of the TOEFL. TOEFL Research Report 27.* Princeton, NJ: Educational Testing Service.

Pena-Taveras, M.S. and **Cambel, A.B.** 1989: Nonlinear, stochastic model for energy investment in manufacturing. *Energy – The International Journal* 14, 421–33.

Purcell, E.T. 1983: Models of pronunciation accuracy. In Oller, J., editor, *Issues in language testing research*. Rowley, MA: Newbury House, 133–53.

Ryan, K. and **Bachman, L.F.** 1992: Differential item functioning on two tests of EFL proficiency. *Language Testing* 9, 12–29.

Sang, F., Schmitz, B., Vollmer, H.J., Baumert, J. and **Roeder, P.M.** 1986: Models of second language competence: a structural equation approach. *Language Testing* 3, 54–79.

Sasaki, M. 1991: Relationships among second language proficiency, foreign language aptitude, and intelligence: a structural equation modeling approach. Unpublished PhD dissertation, University of California, Los Angeles.

Satorra, A. and **Bentler, P.M.** 1988a: *Scaling corrections for statistics in covariance structure analysis. UCLA Statistics Series 2*. Los Angeles, CA: University of California.

——1988b: Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association* 308–13.

Skehan, P. 1989: *Individual differences in second language learning*. London: Edward Arnold.

Snow, C.E. and **Hoefnagle-Hohle, M.** 1978: Age differences in second language acquisition. In Hatch, E., editor, *Second language acquisition: a book of readings*. Rowley, MA: Newbury House, 333–44.

Spolsky, B. 1989: *Conditions for second language learning*. Oxford: Oxford University Press.

Stansfield, C. and **Hansen, J.** 1983: Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly* 17, 29–38.

Swinton, S.S. and **Powers, D.E.** 1980: *Factor analysis of the TOEFL-. TOEFL Research Report 6*. Princeton, NJ: Educational Testing Service.

Turner, C.E. 1989: The underlying factor structure of L2 cloze test performance in Francophone, university-level students: causal modelling as an approach to construct validation. *Language Testing* 6, 172–97.

Upshur, J. 1983: Measurement of individual differences and explanation in the language sciences. *Language Learning* 33, 99–140.

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

Vollmer, H.J. 1983: The structure of foreign language proficiency. In Hughes A., and Porter, D., editors, *Current developments in language testing*. London: Academic Press, 3–29.

Vollmer, H.J. and **Sang, F.** 1983: Competing hypotheses about second language ability: a plea for caution. In Oller, J., editor, *Issues in language testing research*. Rowley, MA: Newbury House.

West, B.J. and **Salk, J.** 1987: Complexity, organizations, and uncertainty. *European Journal of Operational Research* 30, 117–28.

Wheaton, B., Muthen, B., Alwin, D.F. and **Summers, G.F.** 1977: Assessing reliability and stability in panel models. In Heise, D.R., editor, *Sociological methodology 1977*. San Francisco, CA: Jossey Bass, 84–136.

Zeidner, M. 1986: Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3, 80–98.