EMPIRICAL VALIDATION STUDIES OF ALTERNATE RESPONSE MODES FOR WRITING ASSESSMENT

Frank J. Capell
Edys S. Quellmalz

CSE Report No. 145
August, 1980

Test Design Project

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U. S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official NIE endorsement should be inferred.

Table of Contents

Introduction

In the area of large scale assessment, including state assessment and minimum competency testing programs, there is increasing interest in the measurement of students' written performance. As the collection and scoring of written products is considerably more costly than testing in content areas suited to a selected response format (e.g., multiple choice), special attention is given to possible tradeoffs between the number and length of writing samples necessary for accurate assessment on the one hand, and efficient use of time and money on the other. In addition, proposals are entertained to the effect that some areas of writing competence can be assessed indirectly through the use of appropriately developed multiple choice tests. At issue, for example, is whether the task demands in writing assessment can be simplified to involve the production of paragraph-length writing samples and/or multiple choice testing, instead of eliciting one or more full-length essays from each examinee. Another issue, not considered here, are the relative costs and benefits of alternative rating systems for scoring written products (e.g., holistic vs. analytic scoring rubrics).

To compare the information yield of writing measures involving different response modes (i.e., essay, paragraph or multiple choice), data are needed that contrast the performance of a group of examinees across equivalently specified skill domains for each mode of measurement. This study considers data generated in three response modes, two written and one selected response, using a domain-referenced set of specifications

for writing assessment. For the two written conditions, essay and paragraph, examinees produce writing samples in response to written prompts delineating the stimulus attributes for the task, and these writing samples are then rated using an analytic scoring rubric. For the multiple choice response mode, paragraphs are generated from the above mentioned stimulus attributes, with accompanying questions constructed to reflect as nearly as possible three of the five dimensions on which the writing samples are rated. The five scores that result from application of the scoring rubric are General Impression, Focus, Organization, Support, and Mechanics; General Impression and Mechanics are excluded in the selected response condition.

The general issues of concern in this paper are the factorial equivalence of the scale scores derived in different response modes, and the comparative discriminant validity exhibited by the set of scores across response modes. These questions are approached from a multitrait-multimethod perspective, using the model for the analysis of covariance structures developed by Joreskog (1973, 1977; Joreskog & Sorbom, 1978). Analyses treat the five content scales as "traits," and the three response modes as "methods," and address the following specific research questions:

1. Do the five content scales display empirical distinctiveness and homogeneity across measurement methods, such that each scale relates to its underlying "trait" in an invariant manner?

2. Taking as a criterion for scale content, measures derived from full length writing samples, how do paragraph and multiple choice measures compare with respect to:

   a. The discriminant validity of the information they provide?

   b. Their degree of relationship to the underlying trait they purport to measure?

3. Are there particular content scale-response mode combinations that are of especially good or poor quality from a measurement standpoint?

Method

Sample. Complete data were available for a sample of 148 eleventh and twelfth grade students judged by their teachers to be average or above average. Students were drawn from three high schools in the Los Angles area, in a school district where socioeconomic status ranged from upper-lower to upper-middle class. Available test scores on a portion of the sample confirmed their approximately average status on standard verbal ability.

Data Collection and Variables Definition. Students were administered four instruments: two essay writing tasks, one paragraph writing task, and a set of multiple items. Both narrative and expository writing samples were elicited, with essays being on the topics of drugs and violence, and paragraphs on the topic of alcohol use. The complete set of tasks generated 18 scores: the five content scale ratings in the three written conditions (15), and three number-correct scores in the multiple choice condition. Prior to entering the scores into analyses, all variables were standardized within genre (narrative vs. expository) and topic, and then restandardized, to produce the set of variables shown in Figure A.

Analysis Methods. All analyses are based on correlation matrices among the variables described above. The LISREL computer program for the analysis of covariance structures was used to estimate the parameters of all models. LISREL provides standard errors for all parameter estimates (factor loadings and factor intercorrelations), as well as a Chi square

goodness of fit test of overall model adequacy.

Results

The MTMM analyses begin by considering the data for the "essay 1" and "essay 2" methods only, examining the ten scores defined for these two conditions: two measures each of General Impression (gie_1 and gie_2), Focus (fe_1 and fe_2), Organization (oe_1 and oe_2), Support (se_1 and se_2), and Mechanics (me_1 and me_2). The model specified for these variables includes five "trait" factors (one for each subscale) and two "method" factors (one for each essay). Figure 1 illustrates Model I, and the LISREL estimates of the free and constrained model parameters (along with their standard errors in parentheses) are contained in Table 1. The figure shows Model I allowing the trait or subscale factors to be freely intercorrelated, while the method factors are specified to be uncorrelated with each other and with the subscale factors.

Leaving the trait intercorrelations free to be estimated reflects our expectation that the five components of writing ability tapped by the subscales are not independent of one another. The restrictions on the method factor correlations, on the other hand, reflect the hypothesis that they act as independent additive components in the explanation of the observed scores. In addition, the matrix of factor loadings (hereafter referred to as lambda) in the table reveals that we have constrained the loading of each pair of subscale measures on their corresponding trait factor to equal one another. These constraints are equivalent to a test of the hypothesis that subscale scores from different essays will exhibit the same degree of relationship to the trait factor they measure. The model

as a whole cannot be rejected; the chi-square goodness of fit test yields a probability of .138 (ns), suggesting that the model provides an adequate account for the observed data.

Loadings of the essay variables on their corresponding trait factors are all large in magnitude and highly significant, ranging from a low of .521 for Organization to a high of .77 for Mechanics. Except for General Impression and Organization, the loadings of subscale scores on method factors are moderate. One interpretation for the relatively high concentration of method variance in both gie_1 and gie_2 and oe_1 and oe_2 is that residual trait variance that is method-specific is shared by these two subscales. This could be the case if raters depend on their impression of the organization of a given writing product more than on other characteristics of it, in formulating their general impression rating.

Turning to the matrix of factor intercorrelations (hereafter called psi),* we see that the estimates of the relations among the trait factors are all quite high, ranging from a low of .661 for the correlation between Mechanics and Support to a high of .916 between General Impression and Organization. The Mechanics factor appears to be the most independent of the set.

Model II adds paragraph as a method and expands to fifteen the number of variables included in the analysis, by adding the five subscale scores defined for the paragraph response mode. The five trait factors specified in Model I will, under Model II, each have an additional measure of the corresponding trait loading on them (no constraints are placed on these

*Since method factors are restricted to be uncorrelated with one another and with trait factors, the table omits the corresponding portions of the psi matrix which contains only fixed parameters.

loadings); and there will be a new method factor, Paragraph, to absorb irrelevant covariation specific to this mode of responding. Table 2 presents the results of the LISREL estimation of Model II.

Model II provides an adequate overall fit to the observed intercorrelations (chi-square with 70 df = 79.173, p = .212). This result provisionally supports the hypothesis that the scores generated by application of the scoring rubric to paragraph-length writing samples can be interpreted as measuring the same underlying content as the scores derived from full length essays. Inspection of the lambda matrix shows that the loadings for paragraph subscale scores on their associated trait factors are of substantial magnitude in each case, and that the loadings on the paragraph factor follow the same general pattern as for the two essay method factors. With one exception, the paragraph variables appear to relate to trait factors less strongly than do the essay scores. The exception is an interesting one: "sp" provides a clearer definition of the Support factor (i.e., the loading on "S" is higher for sp than for se_1 and se_2). This would seem to suggest that the task of judging the use of support is carried out more accurately in the context of a single paragraph than it is in longer writing samples; a test of this hypothesis, however, would require multiple measures of the sp variable.

As in Model I, the trait intercorrelations in Psi are all quite large, indicating considerable interdependence among the subscales. Again, Mechanics exhibits lower levels of relationship to the other subscales.

Comparison of Models I and II reveals two main differences. First, there is some instability in the size of the essay variables' loadings

on the associated trait factors as we move from the first to the second model. This leads to the interpretation that the factors composed of both essay and paragraph variables do not measure precisely the same content as factors composed of essay variables only. Second, the estimates of trait intercorrelations in Model II are slightly greater in magnitude than the corresponding Model I estimates. Thus, although the inclusion of paragraph scores may have broadened the content of the factors, it seems also to have diminished their distinctiveness. Depending on one's a priori notions about the comparative validity of essay and paragraph data, Model II may be moving us closer to or further away from the true state of affairs. While the differences between the models are relatively small, we will examine this issue in more detail in the context of Model IV.

The third MTMM analysis builds on the previous two by adding the three scores derived from the multiple choice items administered to study subjects. Recall that only items analogous to the Focus, Organization and Support subscales were included in the multiple choice test. Model III differs from Model II, then, by the specification of trait loadings for these three subscores and the addition of a multiple choice method factor. Figure 1 displays the path diagram for Model III (and Model IV); and Table 3 the LISREL estimates of the model parameters.

As in the first two analyses, Model III provides a reasonably good fit to the data (chi-square with 112 df = 125.163, p = .186), implying that the same 5-trait structure is not violated by the inclusion of the multiple choice scores. The sizes of the trait loadings for the three

response modes, as well as the increases in the trait intercorrelations suggest that the trait factors have drifted closer together as a result of adding the multiple choice variables. Thus, while the multiple choice scores apparently share some content with the writing variables to which they are purportedly analogous, they seem also to possess a higher degree of "latent collinearity" (Yates, 1979) in the trait factor space. Whether this situation arises because the multiple choice variables are related to writing ability in some non-specific fashion or because all of the variables, but especially the multiple choice scores, share a common dependence on general ability, would require additional analysis which include test scores based on individual ability. In any event, we are more confident here than for Model II in interpreting the increased interdependence among trait factors as an indication that the multiple choice scores possess generally lower validity as distinctive components of writing ability than do measures derived from actual writing samples.

Model IV examines the relationship of the paragraph and multiple choice variables to the set of trait factors defined solely on the basis of the essay variables. The Model IV rests on the assumption that, at least in the case of the multiple choice scores, the essay-only factors presented a clearer picture of the underlying content of the CSE writing scores; Model IV treats those factors as "unmeasured" criterion variables against which to compare scores from the other two response modes. This can be accomplished in LISREL by modifying the specification for Model III in two places. First, instead of estimating trait loadings for the essay variables, new specifications fix their values to equal those estimated in

Model I. Second, fixing their values at those obtained in the essay-only solution for Model I places a similar constraint on the trait intercorrelations in psi. These two sets of restriction will ensure that the trait factors found in Model I will reappear in Model IV. The LISREL estimates of the free parameters in Model IV are contained in Table 4.

The only parameter estimates of direct interest in Table 4 are the trait factor loadings for the paragraph and multiple choice variables. The data indicate near-uniform reduction in their magnitude in comparison to the estimates obtained from Model III. This shift does not reduce overall model fit (chi-square with 127 df=136.919, p=.258). In all but one instance, paragraph and multiple choice trait factor loadings are lower than the corresponding loadings for essay variables. The one exception is a recurrence of the finding from Model II that the measure of Support derived from a paragraph-length writing sample outperforms the Support measures based on full-length essays. On the other hand, Support as measured by multiple choice items seems to reflect relatively little of what is measured in actual writing samples. The remaining two multiple choice scores, mcf and mco, seem to convey a roughly comparable amount of information about subscale content to that contained in a single paragraph.

Discussion

The MTMM analyses suggest, first, that repeated applications of the method of analytic scoring used in this study in fact produce measures that tap the same underlying content. Second, it was found that the factors reflecting the content of the five subscales are highly intercorrelated, and this interdependence appears to be present no matter what response mode subjects are assessed in.

The MTMM analyses also produce information on the extent to which the various subscale-response mode combinations contain "method variance" not related to their substantive content. The specific models tested suggest that scores on the General Impression and Organization subscales contain large method components when the measures are taken from constructed responses. A plausible explanation for this finding is that the method factor loadings for these variables are inflated by within-occasion (e.g., a given essay) residual linkages between GI and O brought about by raters' tendency to depend more on Organization than on other specific features in formulating their General Impression rating. The remaining three subscales all were found to contain proportionately larger amounts of content-related variance than method-related variance, with Mechanics appearing to be the purest of the three. The patterning of method variance saturation in the five subscales was the same for the three writing samples available for each subject.

An interesting picture of the effects of varying response mode emerged from the analyses. While models can be fitted to the data from all three response modes that confirm the subscale content, the degree of independence

of the resulting subscale factors appears to be affected by which response modes are included in the analysis. The most differentiated subscale factor structure is obtained by including only essay variables in the analysis; interdependence among the subscale factors increases with the addition of both paragraph and multiple choice measures. Thus, the effect of shortening the assessment task for the examinee through examination of just paragraph or multiple choice tasks does not simply increase the measurement error. The savings in testing time are obtained also at the cost of clarity and distinctiveness in the information about each of the subscales. When the subscale content factors are located in the variable space so as to maximize their relationship to scores derived from the essay response mode, all other subscale-response mode combinations except one provide weaker substantive information. The one exception is the measure of Support based on paragraph-length writing samples which seems to be superior to the corresponding essay variables in its ability to capture subscale content. It may be that the use of support is less equivocally evaluated in the context of a single paragraph than in an essay containing multiple paragraphs, each of which may suggest a different view of the examinee's ability to provide supporting detail.

Figure A. Description of Variables for MTMM Analyses

WRITING VARIABLES

	"Essay 1"[1]	"Essay 2"[1]	Paragraph[2]	Multiple Choice[2]
General Impression	gie_1	gie_2	gip	---
Focus	fe_1	fe_2	fp	fmc
Organization	oe_1	oe_2	op	smc
Support	se_1	se_2	sp	smc
Mechanics	me_1	me_2	mp	---

[1] Standardized within genre-treatment conditions, then restandardized after counterbalancing for topic, genre, and serial position.

[2] Standardized within genre, then restandardized.

TABLE 1

LISREL Estimates for MTMM Model Involving

5 Traits and 2 Methods

LAMBDA	GI	F	O	S	M	E_1	E_2
gie_1	.550 (.077)	0	0	0	0	.776 (.079)	0
gie_2	.550 (.077)	0	0	0	0	0	.657 (.087)
fe_1	0	.641 (.069)	0	0	0	.209 (.090)	0
fe_2	0	.641 (.069)	0	0	0	0	.428 (.088)
oe_1	0	0	.521 (.081)	0	0	.726 (.085)	0
oe_2	0	0	.521 (.081)	0	0	0	.769 (.084)
se_1	0	0	0	.557 (.077)	0	.498 (.087)	0
se_2	0	0	0	.557 (.077)	0	0	.408 (.094)
me_1	0	0	0	0	.770 (.062)	.237 (.081)	0
me_2	0	0	0	0	.770 (.062)	0	.182 (.080)

PSI	GI	F	O	S	M
GI	1.0				
F	.721 (.115)	1.0			
O	.916 (.061)	.849 (.113)	1.0		
S	.907 (.100)	.791 (.113)	.792 (.119)	1.0	
M	.772 (.101)	.723 (.087)	.684 (.116)	.661 (.110)	1.0

CHI Square with 20 df = 26.915, p = .138.

TABLE 2

LISREL Estimates for Model II

LAMBDA	GI	F	O	S	M	E_1	E_2	P
gie_1	.560 (.067)	0	0	0	0	.792 (.069)	0	0
gie_2	.506 (.067)	0	0	0	0	0	.653 (.080)	0
gip	.531 (.089)	0	0	0	0	0	0	.728 (.073)
fe_1	0	.618 (.066)	0	0	0	.212 (.080)	0	0
fe_2	0	.618 (.066)	0	0	0	0	.415 (.080)	0
fp	0	.488 (.087)	0	0	0	0	0	.456 (.079)
oe_1	0	0	.524 (.070)	0	0	.699 (.076)	0	0
oe_2	0	0	.524 (.070)	0	0	0	.756 (.077)	0
op	0	0	.436 (.093)	0	0	0	0	.812 (.074)
se_1	0	0	0	.543 (.067)	0	.511 (.077)	0	0
se_2	0	0	0	.543 (.067)	0	0	.403 (.085)	0
sp	0	0	0	.648 (.089)	0	0	0	.415 (.078)
me_1	0	0	0	0	.778 (.061)	.205 (.068)	0	0
me_2	0	0	0	0	.778 (.061)	0	.138 (.071)	0
mp	0	0	0	0	.728 (.077)	0	0	.214 (.070)

PSI	GI	F	O	S	M
GI	1.0				
F	.746 (.085)	1.0			
O	.938 (.039)	.876 (.081)	1.0		
S	.866 (.064)	.889 (.073)	.871 (.073)	1.0	
M	.802 (.070)	.787 (.071)	.767 (.085)	.697 (.078)	1.0

CHI-Square with 70 df = 79.173. p = .212.

FIGURE 1:

Path Diagrams for LISREL MTMM Models

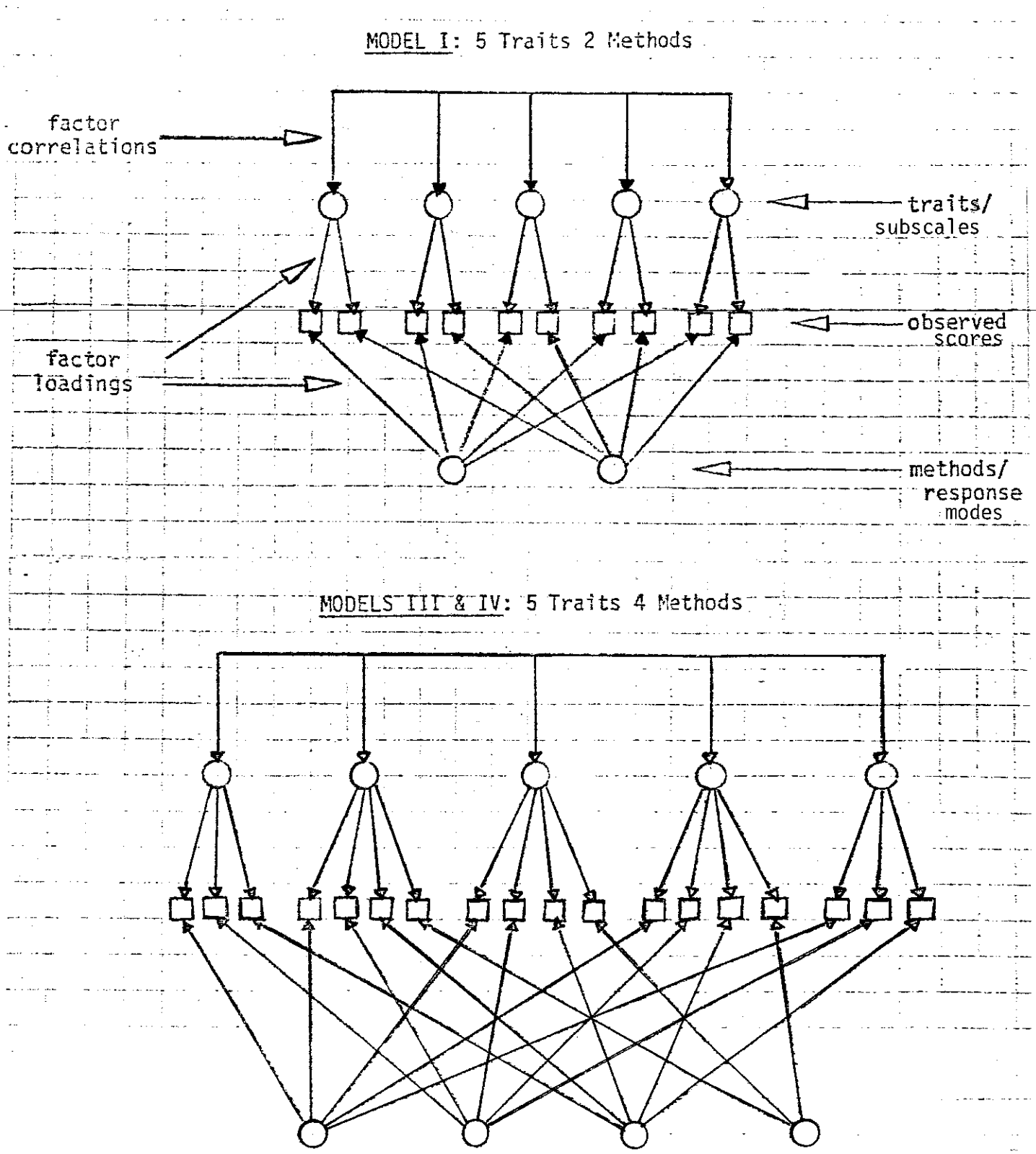

Table 3: LISREL Estimates for Model III

LAMBDA

	GI	F	O	S	M	E_1	E_2	P	MC
gie_1	.533 (.065)	0	0	0	0	.813 (.067)	0	0	0
gie_2	.533 (.065)	0	0	0	0	0	.677 (.077)	0	0
gip	.586 (.087)	0	0	0	0	0	0	.702 (.073)	0
fc_1	0	.604 (.065)	0	0	0	.239 (.078)	0	0	0
fe_2	0	.604 (.065)	0	0	0	0	.399 (.076)	0	0
fp	0	.535 (.087)	0	0	0	0	0	.423 (.080)	0
fme	0	.512 (.087)	0	0	0	0	0	0	.442 (.124)
oe_1	0	0	.495 (.067)	0	0	.734 (.074)	0	0	0
oe_2	0	0	.495 (.067)	0	0	0	.748 (.075)	0	0
op	0	0	.490 (.090)	0	0	0	0	.780 (.075)	0
omc	0	0	.520 (.090)	0	0	0	0	0	.463 (.126)
se_1	0	0	0	.487 (.066)	0	.531 (.076)	0	0	0
se_2	0	0	0	.487 (.066)	0	0	.420 (.083)	0	0
sp	0	0	0	.636 (.087)	0	0	0	.392 (.079)	0
smc	0	0	0	.458 (.091)	0	0	0	0	.411 (.121)
me_1	0	0	0	0	.772 (.061)	.220 (.067)	0	0	0
me_2	0	0	0	0	.772 (.061)	0 (.069)	.138	0	0
mp	0	0	0	0	.746 (.077)	0	0	.186 (.071)	0

PSI

	GI	F	O	S	M
GI					
F	.789 (.072)	1.0			
O	.933 (.037)	.915 (.062)	1.0		
S	.919 (.057)	.943 (.062)	.953 (.058)	1.0	
M	.816 (.064)	.785 (.065)	.783 (.073)	.766 (.072)	1.0

CHI-SQUARE W/112 df = 125.163, p = .186

Table 4: LISREL Estimate for Model IV

LAMBDA

	GI	F	O	S	M	E_1	E_2	P	MC
gie_1	.550	0	0	0	0	.788 (.068)	0	0	0
gie_2	.550	0	0	0	0	0	.656 (.078)	0	0
gip	.520 (.079)	0	0	0	0	0	0	.704 (.020)	0
fe_1	0	.641	0	0	0	.260 (.079)	0	0	0
fe_2	0	.641	0	0	0	0	.381 (.076)	0	0
fp	0	.485 (.083)	0	0	0	0	0	.448 (.078)	0
fmc	0	.477 (.085)	0	0	0	0	0	0	.451 (.117)
oe_1	0	0	.521	0	0	.746 (.074)	0	0	0
oe_2	0	0	.521	0	0	0	.738 (.075)	0	0
op	0	0	.442 (.084)	0	0	0	0	.795 (.073)	0
omc	0	0	.486 (.092)	0	0	0	0	0	.457 (.118)
se_1	0	0	0	.557	0	.515 (.075)	0	0	0
se_2	0	0	0	.557	0	0	.407 (.084)	0	0
sp	0	0	0	.623 (.083)	0	0	0	.406 (.076)	0
smc	0	0	0	.387 (.091)	0	0	0	0	.465 (.122)
me_1	0	0	0	0	.770	.231 (.068)	0	0	0
me_2	0	0	0	0	.770	0	.136 (.069)	0	0
mp	0	0	0	0	.720 (.071)	0	0	.197 (.070)	0

CHI-SQUARE W/127 df = 136.919, p = .258