

The Pennsylvania State University

The Graduate School

THE EFFECTS OF STIMULUS COMPLEXITY, TRAINING, AND GENDER
ON MENTAL ROTATION PERFORMANCE:
A MODEL-BASED APPROACH

A Thesis in

Psychology

by

Geoffrey F.W. Turner

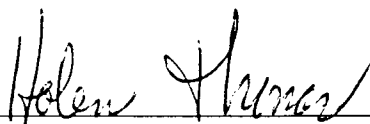
Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 1997

We approve the thesis of Geoffrey F.W. Turner

Date of Signature



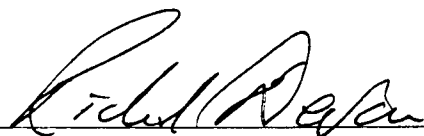
Hoben Thomas
Professor of Psychology
Thesis Advisor

30 July '97



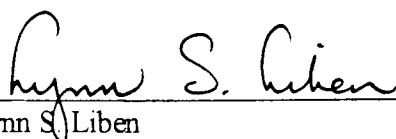
Kelly L. Madole
Assistant Professor of Psychology

30 July 1997



Richard Devon
Associate Professor of Engineering

30 July '97



Lynn S. Liben
Professor of Psychology
Head of the Department of Psychology

30 July '97

The Pennsylvania State University

The Graduate School

THE EFFECTS OF STIMULUS COMPLEXITY, TRAINING, AND GENDER
ON MENTAL ROTATION PERFORMANCE:
A MODEL-BASED APPROACH

A Thesis in

Psychology

by

Geoffrey F.W. Turner

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 1997

Abstract

To assess the impact of stimulus complexity, training, gender, and strategy use on mental rotation performance, a mathematical model was fit to data from first-year engineering students ($n = 556$). Subjects completed 24 Shepard-Metzler and 24 newly developed, more complex mental rotation items administered immediately before and immediately after one of two 12-week engineering graphics design courses. The training group (Penn State students) received extensive practice with mental rotations on a Computer-Aided-Design program, while the contrast group (Cooper-Union students) received a more traditional engineering graphics curriculum.

Studies assessing the effects of training, gender, and item complexity on spatial task performance have typically focused on conventional, normal-theory based analyses in trying to understand these individual differences. Yet there are no individual difference parameters available in these conventional models, and only variables chosen a priori by the researcher can be tested (i.e., age, sex, etc.). Empirical results also indicate that the normal-theory framework is less than ideally suited to evaluate individual differences in mental rotation performance. Instead, a mixture of binomials model is developed which expands on previous models of spatial task performance in that it was applied to longitudinal data of varying complexity to evaluate gender differences and performance change with different types of training. In addition, the current model formulation allows insights into the nature of subjects' strategy use.

The mixed binomial model posits that the general population is made up of more than one (in this case two or three) different "types" or "kinds" of individuals, each with its own level of performance on mental rotation items. Performance for each of these "types" or

"kinds" is represented by its own binomial distribution, and these component binomial distributions form the mixture. Parameters for one, two, three, and four component binomial mixture models were estimated and fit to the data. In almost every case, the two component model provided the best fitting, simplest model of performance, suggesting that the population consists of two performance groups or latent classes. Results indicated that the bivariate data (for example, data for each subject from time 1 and time 2) were well modeled by a bivariate mixture of binomials distribution which provides information concerning shifts in performance from one component group or latent class to another over time.

Results were consistent with previous studies which found a robust sex-difference favoring males. While males scored slightly higher "on average" than females, sex-differences have a very different interpretation than that usually presented in the literature. Within the current modeling perspective, sex-differences were shown to be the result of differential rates of membership in each of the two component groups. In effect, some members of both sexes appear to perform at near ceiling levels, while a segment of both the male and female populations appear to perform at the same lower level. However, a larger percentage of males was classified as having "come from" the better performing group, while a larger percentage of females was classified as having "come from" the worse performing group. Furthermore, while there are differences between "high" and "low" performers regardless of sex, there were few differences between males and females when they were from the same latent class. In addition, males and females showed approximately equal changes in the proportion of subjects classified in the higher performing group at time 1 and time 2, indicating that both sexes showed identical rates of improvement.

Subjects receiving the enhanced curriculum (Penn State students) showed greater improvement than those receiving the traditional curriculum (Cooper-Union students). These performance increases were found to be characterized by dramatic shifts in the number of items correctly solved as indicated by subjects changing component group membership rather than subjects showing smaller, incremental increases in performance. This evidence supports theories which predict abrupt rather than gradual performance change. While the rate of improvement for females was non-significantly different than that for males, the greatest performance increases were seen in Penn State females, suggesting a sex by treatment interaction.

In addition to sex and curriculum, item complexity was found to affect performance. The complex items were more difficult than the Shepard-Metzler items in that fewer subjects were classified in the higher-level performance group on the set of newer items. In addition, some subjects showed a "same" response bias for the Shepard-Metzler items while others showed a "different" response bias for the more complex items, suggesting that even though problems which require "same" and "different" responses are of approximately equal difficulty, they are not always solved in the same fashion. The response biases for "same" and "different" items are likely the result of piecemeal rotation strategies, while consistent ceiling performance is likely the result of a holistic rotation strategy.

In sum, sex-differences, improvement over time, and complexity affects can all be accounted for by a mixture of binomials model. These findings provide support for the view that latent classes of performance define strategy groups. Moreover, because performance change is abrupt, performance changes over time appear to be the result of changes in strategies. Furthermore, the model supports a growing body of literature which suggests that

spatial abilities are fundamentally unique and distinct from other abilities with respect to their latent class structure.

TABLE OF CONTENTS

LIST OF TABLES	xi
LIST OF FIGURES	xiii
ACKNOWLEDGEMENTS	xviii
CHAPTER I 1
Summary 1
Introduction 4
The Mental Rotation Construct. 5
General Findings 5
Psychometric Approach 7
Information Processing Approach 9
Imagery theories 9
Process theories11
Propositional Theories11
Developmental theory13
Process theories13
Propositional theories13
Piaget's theory15
Practice and Training17
A Model Based Approach20
Motivation20
Benefits of Modeling24
Description of the Current Model25
A Model of Task Performance - The Univariate Case26
Parameter estimation28
Tests and confidence intervals29
Model Selection and Assessment29
Chi-squared30
Variance Accounted For31
Classification32
Model assumptions32
Constant θ33
Independence34
Data analysis strategy35
A Model of Performance Change - The Bivariate Case.36
Parameter estimation37
Model assessment37
Intertask correlations38
Summary of Model Approach39
Connections to Other Models39

Statement of Purpose40
Purpose40
Research Questions41
Design Overview42
Hypotheses42
CHAPTER II51
Method51
Participants51
Tasks52
Procedure54
Intervention54
Design55
Assessment56
Response Measure: Reaction Time Versus Accuracy56
CHAPTER III59
Data Description59
CHAPTER IV63
Univariate Binomial Mixture Analyses.63
Basic-Level Analyses63
Probability of Success Estimates ($\hat{\theta}$)66
Proportion Estimates ($\hat{\pi}$)72
General Model Findings78
Summary-Level Analyses78
Partitioning Based on Component Membership82
Covariant Analysis83
Achievement Tests83
Rotation Angle.87
Summary91
Assumptions of the Binomial Mixture92
Constant θ92
Independence Across Trials94

CHAPTER V96
Bivariate Binomial Mixture Analyses96
Parameter Estimation and Interpretation99
Basic-Level Analyses99
Response Bias	102
Improvement Over Time	104
Item Complexity	106
Curriculum	106
Summary-level Analyses	107
Correlation	112
Model Fit	115
Summary	117
CHAPTER VI	118
Discussion	118
Performance Change Over Time	121
Stimulus Complexity	124
Strategy Use	125
Strategy Use and Rotation Angle	128
Strategy and Complexity	131
Strategy Change Over Time	134
Sex Differences	135
Directions for Future Research	139
Summary	142
CHAPTER VII. TABLES AND FIGURES	144
REFERENCES	247
APPENDIX A	259

PSU MRT TEST	259
APPENDIX B	262
Model Development and Estimation in Bivariate Binomial Mixtures .	262
APPENDIX C	265
Model Chi-Square Goodness-of-Fit Statistics	265
APPENDIX D	266
One, Two, Three, and Four Component Model Solutions for All Basic-level Groups.	266
APPENDIX E	274
Z-scores for Penn State and Cooper-Union Males' and Females' Theta Estimates	274
APPENDIX F	275
Z-scores for Penn State and Cooper-Union Males' and Females' Theta Estimates within Item-type by Item-status by Time Grouping	275
APPENDIX G	276
Two Component Restricted Model Estimates and Fit Statistics.	276
APPENDIX H	284
Z-scores for Penn State and Cooper-Union Males' and Females' Pi Estimates	284
APPENDIX I	285
One, Two, Three and Four component Model Parameter Estimates and Fit Statistics for the Summary-level Groups .	285
APPENDIX J	292
Parameter Estimates and Fit Statistics For the Three Component Restricted Models	292

LIST OF TABLES

Table 1.	Sample Sizes by Sex and Treatment Condition	145
Table 2.	SAT Comparison Between Penn State and Cooper-Union Students	146
Table 3.	Variable Notation	147
Table 4.	Descriptive Statistics	148
Table 5.	Penn State Female Model Estimates and Fit Indices	149
Table 6.	Penn State Male Model Estimates and Fit Indices	150
Table 7.	Cooper-Union Female Model Estimates and Fit Indices	151
Table 8.	Cooper-Union Male Model Estimates and Fit Indices	152
Table 9.	Individual Cell Contributions to Overall Chi-Squared Values	153
Table 10.	Male Vs. Female Achievement Test Comparisons	154
Table 11.	Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities of Old-Same Items at Time 1	155
Table 12.	Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for Old-Different Items at Time 2	156
Table 13.	Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for New-Different Items at Time 1	157
Table 14.	Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities of New-Same Items at Time 2	158
Table 15.	Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for Old Items at Time 1	159
Table 16.	Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for New Items at Time 2	160
Table 17.	Comparison of High- and Low-Performing Components'	

	Achievement Test Scores Based on Posterior Probabilities for Old-Different Items at Time 1	161
Table 18.	Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for New-Same Items at Time 1	162
Table 19.	Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for Old-Same Items at Time 2	163
Table 20.	Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for New-Different Items at Time 2	164
Table 21.	Joint Two Component Frequency Tables and Model Estimates for All Subjects on Same and Different, Old and New Items at Times 1 and 2	165
Table 22.	Joint Two Component Frequency Tables and Model Estimates for Males on Same and Different, Old and New Items at Times 1 and 2	168
Table 23.	Joint Two Component Frequency Tables and Model Estimates for Females on Same and Different, Old and New Items at Times 1 and 2	171
Table 24.	Joint Two Component Frequency Tables and Model Estimates for Penn State Subjects on Same and Different, Old and New Items at Times 1 and 2	174
Table 25.	Joint Two Component Frequency Tables and Model Estimates for Cooper-Union Subjects on Same and Different, Old and New Items at Times 1 and 2	177
Table 26.	Patterns of Joint Proportion Estimates	180
Table 27.	Joint Two Component and Univariate Three Component Model Comparison	181
Table 28.	Joint Three Component Frequency Tables and Model Estimates for All Subjects on Old and New Items at Times 1 and 2 .	182
Table 29.	Pattern of Membership in the Non-Zero Cells in the Three Component Bivariate Mixture	184

LIST OF FIGURES

Figure 1.	State Change Parameters Under the Hypothesis That No Learning Occurs	185
Figure 2.	Hypothetical Growth Function Over Two Time Points of the Kind Predicted by Piaget and Inhelder (1956)	186
Figure 3.	Hypothetical Change in Probabilities of Success Without Component Membership Change	187
Figure 4.	Penn State Females' Performance on New-Same Items at Time 1, with Normal, 2- and 3- Component Binomial Mixture Estimates	188
Figure 5.	Penn State Females' Performance on New-Same Items at Time 2, with Normal, 2- and 3- Component Binomial Mixture Estimates	189
Figure 6.	Penn State Females' Performance on New-Different Items at Time 1, with Normal, 2- and 3- Component Binomial Mixture Estimates	190
Figure 7.	Penn State Females' Performance on New-Different Items at Time 2, with Normal, 2- and 3- Component Binomial Mixture Estimates	191
Figure 8.	Additive Shift Model of Sex-Differences	192
Figure 9.	Additive Shift Failure in Mental Rotation Performance	193
Figure 10.	Probability of Success Estimates for Each of the 32 Basic-Level Groups	194
Figure 11.	Proportion estimates for Each of the 32 Basic-Level Groups	195
Figure 12.	Expected and Observed Bimodal Distributions for Theta 1 and Theta 2 for the 32 Basic-Level Groups	196
Figure 13.	Old-Different Item "High" Component Proportion Estimates and Standard Errors for Penn State and Cooper-Union Males and Females at Time1 and Time 2	197
Figure 14.	New-Different Item "High" Component Proportion Estimates and Standard Errors for Penn State and Cooper-Union Males and Females at Time1 and Time 2	198
Figure 15.	Old-Same "High" Component Proportion Estimates	

	and Standard Errors for Penn State and Cooper-Union Males and Females at Time1 and Time 2	199
Figure 16.	New-Same Item "High" Component Proportion Estimates and Standard Errors for Penn State and Cooper-Union Males and Females at Time 1 and Time 2	200
Figure 17.	"High" Component Mixing Proportions and Standard Errors for All Four Item-Type by Item-Status Sets for Time 1 and Time 2	201
Figure 18.	Female "High" Component Proportion Estimates and Standard Errors for Old and New Items	202
Figure 19.	Male "High" Component Proportion Estimates and Standard Errors for Old and New Items	203
Figure 20.	"High" Component Proportion Estimates for Penn State and Cooper-Union Subjects' Performance on All Items at Time 1 and Time 2	204
Figure 21.	"High" Component Proportion Estimates for Male and Female Subjects' Performance on All Items at Time 1 and Time 2	205
Figure 22.	"High" Component Proportion Estimate Differences for Penn State and Cooper-Union Males and Females on All Items	206
Figure 23.	"Low," "Intermediate," and "High" Component Proportion Estimates and Standard Errors for All Subjects on Old and New Items	207
Figure 24.	"Low," "Intermediate," and "High" Component Proportion Estimates and Standard Errors for Penn State and Cooper-Union Subjects' Performance on All Items	208
Figure 25.	"Low," "Intermediate," and "High" Component Proportion Estimates and Standard Errors for Penn State and Cooper-Union Subjects' Performance on All Items	209
Figure 26.	"Low," "Intermediate," and "High" Component Proportion Estimates and Standard Errors for All Subjects' Performance on "Same" and "Different" Items	210
Figure 27.	"Low," "Intermediate," and "High" Component Proportion Estimates and Standard Errors for All Subjects' Performance on All Items at Time 1 and Time 2	211
Figure 28.	"High" Component Mixing Proportion Item-Type by Item-Status Interactions at Time 1 and Time 2	212

Figure 29.	X, Y, and Z axes of Mental Rotation Objects	213
Figure 30.	Non-Commutative Property of Sequential Rotations About Two Axes	214
Figure 31.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Summed Angle of Rotation for "High" and "Low" Performers	215
Figure 32.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Object-Defined Minimum Angle of Rotation for "High" and "Low" Performers	216
Figure 33.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation on the X-Axis for "High" and "Low" Performers	217
Figure 34.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation on the Y-Axis for "High" and "Low" Performers	218
Figure 35.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Angle of Rotation about the Z-Axis for "High" and "Low" Performers	219
Figure 36.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Summed Angle of Rotation for Male and Female "High" and "Low" Performers	220
Figure 37.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Object-Defined Minimum Angle of Rotation for Male and Female "High" and "Low" Performers	221
Figure 38.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation on the X-Axis for Male and Female "High" and "Low" Performers	222
Figure 39.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation About the Y-Axis for Male and Female "High" and "Low" Performers	223
Figure 40.	Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation About the Z-Axis for Male and Female "High" and "Low" Performers	224
Figure 41.	"Low" Component Success Rates on Old-Same Items at Time 2	225

Figure 42.	"High" Component Success Rates on Old-Same Items at Time 2	226
Figure 43.	"Low" Component Success Rates on Old Items at Time 2	227
Figure 44.	"Intermediate" Component Success Rates on Old Items at Time 2	228
Figure 45.	"High" Component Success Rates on Old Items at Time 2.	229
Figure 46.	Inter-Item Correlations on Old-Different Items at Time 2 for the "Low" Component	230
Figure 47.	Inter-Item Correlations on Old-Different Items at Time 2 for the "High" Component	231
Figure 48.	Inter-Item Correlations on Old Items at Time 2 for the "Low" Component	232
Figure 49.	Inter-Item Correlations on Old Items at Time 2 for the "Intermediate" Component	233
Figure 50.	Inter-Item Correlations on Old Items at Time 2 for the "High" Component	234
Figure 51.	Observed and Expected Frequencies for Old-Same and Old-Different Items at Time 1	235
Figure 52.	Observed and Expected Frequencies for Old-Different and New-Different Items at Time 1.	236
Figure 53.	Observed and Expected Frequencies for Old Items at Time 1 and Time 2	237
Figure 54.	Within- and Across-Component Transition Frequencies within Item-Type and Time	238
Figure 55.	Within- and Across-Component Transition Frequencies within Item-Status and Time	239
Figure 56.	Within- and Across-Component Transition Frequencies within Item-Type and Item-Status.	240
Figure 57.	Within- and Across-Component Transition Frequencies within Item-Status and Time	241
Figure 58.	Within- and Across-Component Transition Frequencies within Item-Type and Time	242
Figure 59.	Within- and Across-Component Transition Frequencies within	

	Item-Type and Item-Status.	243
Figure 60.	Funnel Graph of Within-Cell Correlations from the Summary-Level Item-Sets	244
Figure 61.	Models of a Sex by Training Interaction	245
Figure 62.	Male and Female "High" Component Proportion Estimates over Time with a Logarithmic Regression Function	246

ACKNOWLEDGEMENTS

As with most works of this kind, this thesis directly benefited from others' generosity of time, effort, and spirit. First, Dr. Richard Devon's generosity with both data and direction served as the impetus that set this study in motion and facilitated its completion. In addition, my past and present committee members Dr. Julian Thayer, Dr. Clifford Clogg, Dr. David Palermo, Dr. Kelly Madole, and Dr. Richard Devon, could not have been more accommodating or helpful. Special thanks are due to Dr. Lynn Liben, whose guidance, support, and contagious scientific curiosity from my first days at Penn State were instrumental in justifying my fascination with child development.

Others contributed less directly, but no less importantly. I will always be grateful to my parents for sharing the value of education with me and then providing their unwavering support for those pursuits. I need also thank Dr. Ann McGillicuddy-DeLisi, without whom I would never have found a home in psychology or at Penn State. For her unflagging emotional support during all aspects of this thesis - even as she wrote her own - I thank my wife, Laura.

I am most grateful to my mentor Dr. Hoben Thomas, whose patience, wisdom, and unfailing professional support permanently changed the way I see the world.

THE EFFECTS OF STIMULUS COMPLEXITY, TRAINING, AND GENDER
ON MENTAL ROTATION PERFORMANCE: A MODEL-BASED APPROACH

CHAPTER I

Summary

Mental rotation ability has long been recognized as an important component of general spatial ability, intelligence, and vocational achievement. Consistent empirical findings, including a sizable sex-difference favoring males and equivocal gains in performance with training, have been explained by a variety of theories.

Each of the different theories used to explain spatial performance promotes different kinds of variables to account for empirical findings. The psychometric approach has focused on task similarities based on factor analytic techniques. Within this view, differences in intertask correlations between males and females, for example, describe the nature of sex-differences. Information processing theories have focused on the mental processes which affect performance; for example, the differences in the ability to encode a particular set of stimuli. Developmental theories, such as Piaget's theory, describe performance change in terms of general structural changes.

Distinguishing between these theories' predictions concerning spatial abilities has been hindered by fundamentally flawed assumptions concerning the nature of the data used to evaluate their positions. Normal-theory models of performance (e.g. factor analyses, ANOVA's, and Pearson correlation coefficients) have been used almost exclusively to analyze mental rotations performance, yet these models are poorly suited

to evaluate questions of individual differences. The structural model assumptions that underlie these types of analyses define individual differences in terms of experimenter chosen variables such as age or gender. Important variables excluded in these analyses by the investigator become error variance. Furthermore, differences between these variables in, for example an analysis of variance context, can only be viewed as differences between means. For this type of interpretation to be useful, each group must be adequately characterized by a measure of central tendency. Any other form of data are a priori inappropriate. It is argued here that the data describing mental rotations performance and change over time can not be well characterized by a single measure of central tendency and that individual differences become obscured through the use of these kinds of analytic techniques.

However, there are more informative methods of evaluating individual differences. There is evidence to indicate that responding on accuracy trials is a stochastic process affected by the different component processes used to make mental rotation decisions. Further, evidence suggests the existence of at least two "types" or "kinds" of subjects whose performance is being measured. The current study attempts to frame questions of mental rotations performance in terms of a strong mathematical model designed to illuminate the nature of individual differences in mental rotations performance and performance change which takes these findings into account.

A mixture of binomial distributions model that has been successfully used to account for individual differences on other spatial tasks has been adapted to the current task. Individual differences in performance over time are viewed in terms of a small number of discrete performance groups, each with a different performance level. In

effect, different "types" or "kinds" of performers are viewed as having been sampled from the general, mixed population. The proposed model allows individuals to be categorized into their component groups based on their performance, so that, for example, performance shifts can be tracked over time.

Mathematical models, such as the one proposed here, have the advantage of precisely describing and evaluating variables of importance. Once individual difference variables are successfully accounted for, theoretical questions can be addressed more adequately. Furthermore, the proposed model provides a general framework with which to view performance within the broader context of mathematical psychology.

An intervention using a Computer-Aided-Design (CAD) package was provided to all Penn State first-year engineering students in the fall of 1993 in an attempt to improve their ability to perform vocationally-related mental rotations (See Table 1 for a summary of the subject distribution and Chapter II, page 48 for a more detailed account of the study conditions). Their performance on a mental rotations task (which used items from the Vandenberg and Kuse (1978) mental rotations task and similar, but more complex items, see Appendix A) was compared to a group of Cooper-Union College first-year engineering students who received a traditional engineering graphics curriculum, before and after training (See Table 1). The traditional engineering graphics curriculum used a crude wire-frame CAD program that does not allow objects to be manipulated as naturally or as easily as the intervention CAD software.

A comparison of these two groups of subjects was used to reveal differences in training techniques and provide insight into performance as it relates to performance change over time, the effects of stimulus complexity on performance, and whether

different strategies are used to solve these types of mental rotation problems. Specific hypotheses are presented in detail at the end of this chapter. These questions, once answered in a more satisfying way, can then be used to address theoretical issues concerning how performance changes over time. To provide a brief example, consider the issue of transfer of mental rotations skill (this issue is operationalized in more detail in the hypotheses section). Olson and Bialystok (1983) argue that the cognitive operations required for mental rotations (and spatial cognition in general) have their basis in the ability to construct propositional descriptions. The only way to acquire suitable labels for object parts is through interaction with them. As a result, training with one set of objects should not improve performance on unfamiliar objects because the labels are unique to each object and can not be re-used. Olson and Bialystok (1983) found evidence to support this view. In contrast, Piagetian theory would predict that the operations involved in an internalized abstract coordinate system could then be applied to a host of problems which require those operations. In the current study, the relative advantage of the treatment group would be denied by Olson and Bialystok, but supported by Piaget. Other theoretical differences will be examined below.

Introduction

Correlational and mean-based analyses of manifest or observable variables have been used to the exclusion of almost all other methods in describing spatial abilities (Linn & Petersen, 1985). However, evidence of latent-classes of performers (i.e., clusters of subjects defined by their similarity on some latent or unobservable variable) has been found on several spatial tasks suggesting that these normal-theory based approaches are inappropriate (Thomas & Lohaus, 1993; Thomas & Turner, 1990; Turner, 1991). In

addition, factor analyses, based on inter-task correlational matrices, depend on strategy similarity across subjects, yet this assumption may not be valid (Cooper & Mumaw, 1985; Lohman & Kyllonen, 1983; Lohman, 1988). Theoretical positions either tested with or based on seriously false models are bound to be substantially inaccurate. As Lohman and Kyllonen (1983) point out, recognition of this possibility entails a fundamental change in the way test scores are conceptualized. This kind of reconceptualization is exactly what the present study proposes. The goal of the current research is to interpret performance and performance change on a mental rotations task from a modeling perspective, and then apply these findings to current theory. This effort should be viewed in part as exploratory because there is no widely accepted model of performance or performance change except those implicitly based on normal-theory. After performance and performance change have been properly modeled, explanations for individual differences in performance based on this model will be developed.

The remainder of this chapter is divided into three sections. The first is concerned with a summary of the major research findings on mental rotation ability and a subsequent discussion of relevant theory including the role of strategies. The influences of practice and training on performance are described next, followed by a mathematical approach to psychological investigations and the specific model formulations involved in the present study. Finally, a description of the study's design and a list of hypotheses are presented.

The Mental Rotation Construct

General Findings

Spatial abilities have long been viewed as important components of general cognitive ability (Thurstone, 1938). The ability to mentally manipulate two- or three-dimensional images rapidly and accurately, and then act on the resulting mental representation has been considered a central component of general spatial ability. Skill in mental rotation facilitates mastery of many substantive areas, including mathematics, chemistry, engineering, architecture, and aviation (Brinkman, 1966; Gordon & Leighty, 1988; Kyllonen, Lohman, & Snow, 1984; Seddon, Eniaiyaju, & Jusoh, 1984; Shubbar, 1990), and has some predictive validity in personnel selection (Ghiselli, 1973).

Reaction time studies have found that the time required to mentally rotate a stimulus monotonically increases with the amount of rotation required (Cooper & Shepard, 1973; Shepard & Cooper, 1982; Shepard & Metzler, 1971). This finding is consistent across tasks which use widely differing stimuli including letter-like characters and abstract geometric figures (e.g., Kail, Carter & Pellegrino, 1979; Shepard & Cooper, 1982).

When accuracy, as opposed to reaction time, is measured, object complexity appears to have a large impact. When two-dimensional letter-like objects are rotated, adults have near ceiling accuracy scores. Complex, three-dimensional stimuli, on the other hand, substantially reduce accuracy rates (e.g., Folk & Luce, 1987; Yuille & Steiger, 1982).

By about age 9, males show a distinct superiority in both accuracy of judgments and speed of rotation (Johnson & Meade, 1985; Linn & Petersen, 1985). This sex

difference has been observed across a wide variety of rotation tasks, though it is more pronounced in the Shepard and Metzler (1971) version. In their meta-analysis, Linn and Petersen (1985) estimated that males perform on the order of one standard deviation better than females. These results give the impression that women are less spatially able than men. However, Tapley and Bryden (1977) indicate that there is a substantial minority of women who outperform men on psychometric measures of spatial ability. Eliot (1987) points out that sex-differences are characterized by trends, but that the majority of data variance is more readily accounted for by individual differences than sex-differences.

However, as Thomas and Lohaus (1993) make clear, the normal-theory based approach most often used to analyze mental rotation data is not amenable to studying individual differences. Lohman and Kyllonen (1983) describe the two most common methods of viewing individual differences as an attempt to find either cognitive correlates (the psychometric tradition) or cognitive components (the information processing approach). These two methods are considered in turn. Because performance change over time is of interest, developmental and learning perspectives will also be discussed.

Psychometric Approach

Historically, psychometricians have considered spatial abilities from a factor analytic perspective. The carving of spatial abilities into a small set of component abilities and establishing their relationships has been accomplished by administering large batteries of spatial tests and then examining their covariance structures. These models work under the assumption that tasks load on a particular factor as a result of their shared underlying processes. It is these presumed mental processes that govern

individual differences in performance. Lohman's (1979, cited in Eliot, 1987) re-analysis of data from early factor analytic studies led to him to postulate that there are three basic spatial factors: speeded rotation, spatial orientation, and visualization. In a review of the psychometric literature, McGee (1979) came to similar conclusions. While the process of mental rotation is common to many tests, proficiency is due to both speed-related and visualization-related elements. The visualization factor appears to best describe the kind of relatively unspeeded, complex rotation problems characterized by the Vandenberg and Kuse (1978) mental rotation task under examination in this investigation.

While psychometric theory has provided valuable information regarding mental rotation ability, there are inherent limitations to this approach. Factor analyses rest on the assumption that tasks loading on the same factor share mental process requirements, and consequently that solution strategies on each task are consistent for each subject (Lohman, 1979, cited in Eliot, 1987). There is enough empirical evidence to doubt the validity of this assumption (Carpenter & Just, 1986). Lohman and Kyllonen (1983) found that when subjects could be sorted according to natural strategy groups, different factor structures emerged for each strategy group, suggesting that the combined factor structure was misleading. Furthermore, Lohman (1988) argues that spatial tasks load highly on a g-factor because this factor characterizes subjects' ability (although not necessarily inclination) to use multiple strategies in solving similar problems. Factor analysis also assumes multivariate normality, yet the suggestion of different strategy groups and strategy shifting cast doubt on this assumption.

Factor analysis is further hindered by an abundance of rotational techniques which often yield different results. Currently, there is little agreement as to which

rotational method is best (Cooper & Mumaw, 1985; Eliot, 1987; Poltrock & Brown, 1984). Linn and Petersen (1985) point out that factor analytic techniques are sensitive to the tasks involved, and as such solutions for different tasks may never lead to a general factorial representation of spatial abilities.

Information Processing Approach

Information processing accounts of mental rotation ability have been divided into three areas: imagery (or visualization) theories, process-based theories, and propositional theories.

Imagery theories. The imagery account is based on an assumed relationship between visual imagery and spatial abilities. Paivio (1971) describes imagery as a symbolic process that anticipates alternative responses to cognitive problems. While visualization is usually defined as a general ability encompassing more than spatial abilities, visualization seems to play an important role in how space is conceptualized (Lohman, 1986; Poltrock & Brown, 1984).

Kosslyn (1980) has advanced a general theory of imagery which has implications for the study of spatial representations. He argues that people make use of analog mental representations that preserve the qualities found in perceptual representations, like distance and position relationships. Kosslyn has developed a computer model that mimics the processes unique to mental images. These processes (e.g., scanning, zooming, rotating and refreshing images) are quite similar to those that would be found on a Computer-Aided-Design (CAD) program. As evidence for his theory, he has tried to demonstrate that subjects report on or use information from their mental images as though they had perception-like qualities. For example, subjects scan images to find

pieces in the same way that they would scan a photograph or an actual scene. The time required to search for an object in a mental image seems to correspond to its distance from an initial viewpoint. He has also found that subjects' images have a limited resolution and spatial extent in the same way that perceptual images do. Wallace and Hofelich (1992) have found evidence to support this view, observing that the components of imagery influential in image rotation are predicted by Kosslyn's theory.

Shepard reasoned that if an analog process of mental rotation of the kind suggested by Kosslyn is correct, then the time necessary to mentally rotate an image should reflect the amount of rotation necessary. To test this hypothesis, Shepard and Metzler (1971) presented subjects with pairs of two-dimensional line drawings of three-dimensional figures. Subjects were asked to determine as quickly and accurately as possible whether the two images were the same or different. Results indicated that the time required to make judgments for identical objects was linearly related to the angle of rotation necessary to bring the objects into congruence. He and his colleagues (Shepard & Metzler, 1971; Shepard & Cooper, 1982) contend that this pattern of observed reaction times could only be the result of analog mental rotation. Because this view relies on mental imagery to explain mental processes, the quality of subjects' mental images should have a profound impact on their ability (Carpenter & Just, 1986; Juhel, 1991; Pellegrino & Kail, 1982; Poltrock & Brown, 1984). Lohman (1986) and Poltrock and Brown (1984) have gone so far as to argue that the quality of internal visual representations characterizes spatial ability.

Olson, Eliot and Hardy (1988), Ozer (1987) and Tapley and Bryden (1977), for example, found that visualization skill was related to spatial performance for men, but not

women. These findings provide partial support for the interpretation that the mental images of high spatial individuals are more highly organized, and thus more useful in problem solving, than those of low spatial individuals (Eliot, 1987). Further support for this view comes from Shepard and Metzler's (1971) original work. They argued their analog mental rotation theory based on the performance of subjects pre-selected for their high spatial ability. It would appear that at least high spatial subjects have the requisite visual imagery skills to mentally rotate stimuli.

Process theories. Kail and his colleagues (e.g., Kail et al., 1979; Mumaw, Pellegrino, Kail & Carter, 1984) acknowledge that imagery plays a role in spatial task performance, but focus on the component processes required to make spatial judgments. The total time required to respond "same" or "different" to images rotated out of alignment is assumed to be the additive sum of four individual processes: encoding, rotating, comparing, and responding. The rotation process varies as a function of the angular separation between the stimuli, but all four processes are thought to be influenced by variables such as the familiarity and complexity of the stimuli. Pellegrino and Kail (1982), for example, predict that subjects in general will be less likely to respond correctly to more complex stimuli because there are more processing operations necessary for complex items. While the time required for each step may change with stimulus variations, the processes themselves remain constant. As such, learning, developmental and individual differences are all characterized by reference to these four processes.

Propositional Theories. There are other theoretical orientations which account for chronometric mental rotation data without reference to mental images. Connectionist

models of cognition have been constructed which can predict the linearly increasing response times and the trajectories that rotated objects pass through (Funt, 1983; Goebel, 1990). Folk and Luce (1987) equate this type of model to more conventional propositional theories of spatial ability.

Propositional theorists describe the representations necessary for solving mental rotation problems as neither imaginal nor verbal. Anderson (1978), Olson and Bialystok (1983) and Pylyshyn (1981) describe a general theory of cognition applicable to a wide range of spatial phenomena. Images become nothing more than epiphenomenal by-products derived from propositional representations. According to Anderson (1978), spatial propositions have three defining qualities. First, they are abstract and more basic than lingual propositions. Second, they have truth value, and as such can be evaluated in much the same way that propositions are evaluated in formal logic problems. Third, they follow a basic set of rules for their formation.

Spatial cognition is based on the structural descriptions of objects. The mental predicates used in spatial representations closely mirror verbal predicates like "top," "over," and "in front of" (Olson & Bialystok, 1983). In fact, these authors argue that linguistic spatial categories emerge out of the more fundamental structures of perception and thought. Like an imagery theory, a propositional account of representation includes the notion that perceptual and constructed representations share features which allow them to be compared easily.

Olson and Bialystok (1983) argue that the absence of namable parts (e.g., top, front) makes the formation of descriptions from one's own or an observer's perspective more difficult for problems of the Shepard and Metzler (1971) type than it is for familiar objects

like cars and bottles. They suggest that parts of abstract images are labeled with structural descriptions based on the object's initial orientation with respect to the observer. Solving mental rotation problems becomes the four-fold problem of (1) "naming" the structural parts in the target display, (2) finding them again in a comparison display, (3) noting the rotation angle between the structural components of the two displays, and then (4) evaluating the similarity of the relations between their structural parts. Comparisons continue until all of the features have been judged or until a mismatch is found.

As Anderson (1978) has argued, behavioral investigations may not be the most productive method of evaluating whether subjects' representations are imaginal or propositional. The goal of this study is not to attempt to resolve the debate concerning the form of the representations used to solve mental rotations problems, but rather, to better characterize performance and to evaluate different theoretical positions in light of the model of performance described below. Anderson (1978) argues that while behavioral data cannot distinguish between classes of theories which postulate different representational systems, they can be used to distinguish between specific theories.

Developmental Theory

Process theories. Much of the developmental literature concerned with mental rotation has been produced by information processing theorists. Typical of this approach, Kail and his colleagues (e.g., Kail, 1991; Pellegrino & Kail, 1982; Kail, Pellegrino & Carter, 1980) describe individual and developmental differences in terms of process differences. Learning and development are attributed to improved processing in any one of the four steps outlined above, such as encoding and comparison, both of which improve with age (Kail et al., 1980; Kail, 1991). Increased processing ability can result

in changes in strategies though, indicating a change in instance-based or procedural (rather than declarative or process-based) knowledge (Carter, Pazak, & Kail, 1983; Kail & Park, 1990; Lohman & Nichols, 1990).

Propositional theories. In contrast, Olson and Bialystok's (1983) account of spatial development is based on their propositional theory of spatial cognition. Development, according to this view, is the process of extracting forms from implicit perceptual structures and relating them to explicit representational structures. For example, linguistic spatial categories are thought to develop out of more basic implicit perceptual and cognitive structures. Olson and Bialystok (1983) take children's drawing as evidence for this view, by noting that at very young ages children represent a multitude of objects with a few simple spatial forms. This performance is consistent with the small number of lexical forms young children have for describing the world. In adulthood, however, just as language has become more enriched, so too has spatial representation.

According to Olson and Bialystok (1983), developmental differences in spatial ability are most strongly related to the ability to intentionally define arbitrary structural parts in non-canonical (e.g., abstract) objects, which young children have greater difficulty with than adults. This difficulty arises because there is little linguistic information available in abstract objects to help younger children. There is empirical evidence to support this position. Kail et al. (1980) found that the abstract characters from the test of Primary Mental Abilities (PMA) are more difficult to rotate than alphanumeric characters because the PMA characters are not as easily "labelable" as familiar letters or numbers. Similarly, Hochberg and Gellman (1977) and Alington, Leaf, and Monaghan (1992) found that the addition of

landmark features to objects facilitated their rotation. Additionally, Hoch and Ross (1975) found that item familiarity increased mental rotation performance.

According to this theory, practice provides familiarization with objects even without readily available linguistic information. As a consequence, the ability to uniquely identify the structural features of an "abstract" array and form useful structural descriptions of the familiarized objects increases with practice.

Olson and Bialystok (1983) used films depicting rotation and direct manipulation in an attempt to teach mental rotations. They found that approximately 12 minutes of task-related training resulted in improved post-test performance. This suggested that children lacked explicit identifiers for the structural features in the Shepard and Metzler (1971) type object. These findings imply that it is not simply practice that improves mental rotation performance, but practice that allows the formation of adequate structural descriptions of objects. These structural descriptions, which represent changes in declarative knowledge (Lohman & Nichols, 1990), are then used in identical mental rotation tasks. It is only when changes in procedural knowledge occur that transfer is likely to other rotations tasks, however, because the structural features of an object are specific to that object. It may be noted that the abstract items used in their study had features that were more easily labelable than those used by Shepard and Metzler (1971) and in the current study. Ease in labeling the features of the rotation objects was likely to have made structural descriptions more accessible to Olson and Bialystok's (1983) subjects, but less likely to facilitate general mastery of rotations.

Piaget's theory. Piaget's theory of development stands in contrast to the accounts previously presented. According to Piaget and Inhelder (1956, 1971), conceptual level

determines representations. It is only after projective and Euclidean concepts have been developed and coordinated that objects can be represented and transformed by rotation. Mental rotation requires the coordination of dimensions in terms of displacement about a set of axes, and would therefore be a later developing ability.

Just and Carpenter (1985) found empirical support for this position. Subjects who performed most accurately on a mental rotations test were able to define an abstract coordinate system independent of the rotated object. The least accurate subjects, in contrast, could only make reference to the coordinate axes defined by the object. These authors argue that the information available from the use of an independent coordinate system allows better performance. Performance on Piaget and Inhelder's (1956) water-level task is similar in that poor performers on that task were constrained by the frame of the vessel, while good performers were able to construct and use an external coordinate reference system.

Successful mental rotations signifies the ability to represent perspective changes in terms of operations that define the coordination of different viewpoints. These operations include reversibility and the idea that a projection in one dimension is accompanied by corresponding reduction in another dimension (i.e., a grouping of the relationships which comprise the three spatial dimensions). Ultimately, the correspondence of different perspectives are linked via a representational coordinate system. The metric properties (straight lines, distances, and angles) of an object must be conserved during the transformation, otherwise two different objects will be compared. In fact, identifying two mental rotation items as different might be viewed as the successful recognition that Euclidean relationships have not been conserved.

Development, according to Piaget and Inhelder (1956), results from the internalization of action and imitation through visual perception. Actions and perceptions allow rotational transformations to become operationalized and coordinated. According to Piaget, mental images contain neither the relationships between parts of an object nor an object's relationship to a coordinate system. These relationships must be formed constructively, thereby allowing the image to be used in different schemes. Those structures which entail the logical application of operations to many tasks could also be viewed as improvements in procedural knowledge.

According to Piaget and Inhelder (1956, 1971), the requisite operations for successful mental rotation should be in place when thought is concrete operational. As with other Piagetian tasks (e.g., the water-level task), many individuals do not seem capable of this level of performance even well after adolescence.

Practice and Training

The effects of practice and training have been equivocal, though few training studies have been undertaken (Eliot, 1987). Of those conducted, most have been implemented over short time spans, neglected generalizability, or have trained subjects to a criterion on a particular task. There are a few notable exceptions which provide insights into performance change.

Baenninger and Newcombe's (1989) meta-analysis revealed that relatively long-term training programs can be effective in improving spatial abilities. VanVoorhis (1941), for example, demonstrated over the course of a year that visualization training in freshmen engineers provided significant gains in mental rotation performance over a matched control group. Blade and Watson (1955) concluded that completion of first-year

engineering courses was correlated with improvement in spatial task performance. Additionally, Brinkman (1966) found that training in a geometry course that focused on visual problem solving improved spatial task performance. Olson et al. (1988) found a positive relationship between spatial performance and participation in a variety of technical courses. In addition, researchers have made connections between performance and activities that seem to promote spatial competency including the use of computers in both game playing and programming (McLurg & Chaille, 1987; Miller, Kelly, & Kelly, 1988).

Because boys and girls participate differentially in "spatial" activities, Sherman (1967) hypothesized that females would be more receptive to spatial training because they are farther from asymptotic performance. Several researchers have found evidence that training is of greater benefit to females (Connor, Serbin & Schackman, 1977; Connor, Schackman & Serbin, 1978; Lohman & Nichols, 1990). While Baenninger and Newcombe's (1989) meta-analysis revealed little support for this view, meta-analyses based on means which fail to take group membership into account might be misleading.

In general, training studies have the largest impact when they are task-specific, even though generalizability suffers (Baenninger & Newcombe, 1989). For example, mental rotation performance improves after inspection of physical models. The use of sequential diagrams and films which depict rotation also have a large impact on mental rotation improvement (Kyllonen et al., 1984; Olson and Bialystok, 1983; Willis & Shaie, 1988). Subjects' performance benefits when problem attributes such as the distortion of angles and relative cue size are made salient (Connor et al., 1978; Seddon et al., 1984), or

when simply provided with feedback after each response (Kyllonen et al., 1984; Lohman & Nichols, 1990).

There is evidence to suggest that training is not automatically beneficial.

Kyllonen et al. (1984) and Cooper and Mumaw (1985) have presented evidence that training is mediated by an aptitude by treatment interaction. Training success varies as a function of subjects' natural strategies. Visually inclined subjects responded positively to visual training, but negatively to verbal training. Likewise, subjects prone to use verbal methods improved when provided with verbal training, but their performance suffered from visual strategy training.

Practice effects have been documented even in the absence of training (Bethel-Fox & Shepard, 1988; Kaplan & Weisberg, 1987). McLurg and Chaille (1985) and Willis and Shaie (1988) found that practice proved more effective for females than males. Kail's (1986) finding that subjects deficient in practice demonstrate superior improvements in accuracy, implies that females, in general, do not get as much practice as males in spatial activities.

Bethel-Fox and Shepard (1985) found evidence consistent with the view that practice results in changes in declarative knowledge, in that transfer was limited to practiced items. Kail and Park (1990) made similar observations about the lack of transfer from a practiced spatial task to a novel one. Practice primarily results in improvements in declarative knowledge, yet extensive practice should provide subjects with more elaborate representations of rotation objects (Lohman & Nichols, 1990). When used in a variety of circumstances, processes become decontextualized.

Kosslyn's (1980) imagery theory predicts which kinds of practice are the most beneficial. Wallace and Hofelich (1992) found that practice related to improvement in mental rotation affected tasks which require the same processes (e.g., rotating and refreshing), but not those tasks which necessitate the use of other processes. This suggests that transfer is not necessarily restricted to identical problems, but to tasks with similar process requirements.

The effects of practice without training in mental rotation appear to be mediated by the presence or absence of feedback (Lohman & Nichols, 1990). Without feedback, reaction times decreased even though accuracy remained constant. Presumably, without feedback, subjects get faster at responding incorrectly. When feedback was provided, however, accuracy and latencies both improved.

A Model Based Approach

Motivation

The present work was, in part, motivated by an intuitive recognition that normal-based models of performance have obscured patterns of individual differences (e.g., Lohman & Kyllonen, 1983), including evidence of latent classes of performers (e.g., Thomas & Kail, 1991).

As evidenced in Linn and Petersen's (1985) meta-analysis, age- and sex-differences are almost universally interpreted as additive-shift models. In this context, the distribution of male scores is conceptualized as identical to the distribution of female scores, differing only by an additive constant. Thomas and Lohaus (1993) point out how inappropriate assumed shift models of performance can be by showing that the traditional view of sex and age differences as mean differences are misleading when compared to a

more appropriate model. Eliot (1987) has noted that sex-differences are overshadowed by much larger individual differences. Unfortunately, t-tests, analyses of variance, and correlations are not well suited to the study of individual differences. Furthermore, measures of central tendency (i.e., sample means) do not provide much useful information when distributions are not unimodal, as in the case of mental rotation data (Turner, 1991). As Pellegrino and Kail (1982), Brainerd (1979a) and many others have pointed out, group patterns may not reflect any individual's performance. Even though these analyses cannot be expected to provide answers to many of the questions demanded of them, researchers employ them by default; a situation described by Box (1976) as "cookbookery" (p. 797). Lohman and Kyllonen more generously describe it as "blind faith" (1983, p. 114). For these reasons, the normal-based model seems manifestly inadequate (Thomas & Lohaus, 1993; Thomas & Turner, 1990; Turner, 1991). Further evidence will be provided below to demonstrate that mental rotation performance is non-normal and that group (e.g., sex) differences cannot be characterized by an additive shift model. Despite the lack of alternative models presented in the mental rotation literature, a more plausible distributional model must be employed. A model attributing group differences to differential membership in discrete populations, each described by different levels of performance, will be presented below.

Evidence concerning the characteristics of mental rotation performance comes from two sources: explicit models and patterns of data from previous research. Thomas and his colleagues (Thomas & Lohaus, 1993; Thomas & Turner, 1991; Turner, 1991) found that a mixture of binomials distribution successfully characterized performance on different spatial tasks, so it is suggested that mental rotation performance may be similar.

Other researchers have postulated models of accuracy on mental rotation tasks which imply that the binomial distribution is relevant. For example, Carter et al. (1983) introduce the following notation. Let $P(e_i)$ represent the probability of an error on an identical item (one that requires a "same" response), $P(e_{mi})$ represent the probability of an error on a mirror image item (one that requires a "different" response). Now define the probability of an error on an identical image mental rotation item as $P(e_i) = 1 - (1 - \alpha)^n (1 - \beta)$, where α represents the probability of an error during mental rotation, n is the orientation of the stimulus item (in degrees), and β is the probability of an error in either the encoding, comparison, or response phases of the response. When the two items are mirror images of one another $P(e_{mi}) = 1 - (1 - \alpha)^n (1 - \lambda)$ where λ represents the probability of making an error during the encoding, comparison and response phases of a mirror image trial, and α and n are defined as above. If $(1 - \alpha)^n (1 - \beta) \approx (1 - \alpha)^n (1 - \lambda)$, and the effects of rotation angle are small, then one might reasonably view the probability of an error as

$$1 - [(1 - \alpha)^n (1 - \beta)] \approx 1 - [(1 - \alpha)^n (1 - \lambda)] \approx 1 - [\theta].$$

This is, in fact, the probability of an incorrect response in the binomial setting, where θ is defined as the probability of a correct response. The assumption of approximately equal difficulty for identical and mirror images trials is supported by data from Damos (1990) and Pellegrino and Kail (1982), while the assumption concerning the small effect of rotation angle on the present task is supported by data from Carter et al. (1983) and

Lohman (1986). As developed, however, this model cannot account for different error rates among different individuals.

Evidence hinting that there is more than one population whose performance is being sampled abounds, though it is rarely recognized as such. For example, Kail et al. (1979) found that sex-differences in mental rotation performance were characterized by the fact that a relatively large proportion, but not all, females had significantly longer reaction times than all males. This finding implies that there is more than one "kind" of individual within the population.

This type of evidence prompted Thomas and Kail (1991) to model latencies on the PMA test with a mixture of normals distribution, providing explicit evidence of two latent classes of performers. Because a larger number of strategies are found when more complex tests are administered, a more complex model structure might be expected to fit Shepard and Metzler (1971) type items (Lohman & Kyllonen, 1983). Distributional evidence based on a small number of subjects suggests that this is the case (Turner, 1991).

Patterns of performance that describe clusters of individuals have also been found (Cooper, 1982; Bryden, George, & Inch, 1990). While no models are explicitly provided for understanding these clusters, latency and accuracy data appear to support the notion that there is more than one "kind" or "type" of performer. For example, one cluster of individuals is characterized over multiple rotation tasks by the frequently observed pattern of correspondence between latencies and angular deviation, while another cluster is not. Mislevy, Wingersky, Irvine, and Dann (1991) provide a model of reaction time performance which views the population as a mixture of strategy groups. Frequencies of

errors also suggest that there are at least two qualitatively different subject clusters, perhaps differentiated by strategy choice. Cooper (1982) argues that these clusters represent stable individual differences.

Mumaw et al. (1984) found a similar pattern of clusters of performers with different latency and accuracy scores. While they identify four groups of performers, their figures seem to indicate nearly identical performance for the middle two groups, suggesting that a data-driven model may have only found three groups. Voyer and Bryden (1990) present almost identical data showing latencies clustered into three apparent categories.

Lohman (1988) describes the differential effects of practice with feedback for high and low spatial subjects, suggesting that there are qualitative differences between these two groups of subjects. Just and Carpenter (1976) found evidence based on eye fixations that high spatial individuals rotate objects holistically, while low spatial subjects use multiple rotations. Kyllonen, Lohman, and Woltz (1984) also identified two groups of subjects based on their strategy use. It might be that strategy differences distinguish high and low ability subjects, a hypothesis that latent class models are well suited to test.

Researchers implicitly assume that the spatial ability construct is continuous, yet subjects are often split into ability levels based on either an extreme groups design or an arbitrary partition of percentile scores. Without a formal model, ad-hoc groupings will be inconsistent across studies and strategy groups, making interpretations and generalizations difficult. "Natural" groupings (i.e., those derived from the data) provide a much better approach to the problem. This is not to imply that the choice of models is solely data-driven. The form of the model, a mixture of binomials distribution, has both

direct and indirect support from previous research. As with any model fitting procedure, a close correspondence between the data and the model is desirable. The values of the estimated parameters, however, will be based completely on sample data.

Benefits of Modeling

The advantages to a modeling perspective are in its general applicability, its precision and its specificity of predictions (Coombs, Dawes, & Tversky, 1970). By employing a "bottom-up" modeling procedure that allows the data to guide the form of the model, those variables important to describing individual differences can be precisely described and evaluated. Furthermore, the model proposed here is a strong mathematical model. Strong models have the advantage that they are easily falsifiable.

As Thomas and Lohaus (1993) point out, to specify causative variables without knowledge of the structure of the phenomenon of interest is to place the cart before the horse. To begin with a suitable model structure, and then let the model resolve variables of importance would seem to be a more productive approach.

Bejar (1990) points out that there is often conflict between psychometric and information processing approaches because each perspective emphasizes different performance aspects. It is argued here that this need not be the case. Any successful model of task performance should elucidate the issues relevant to task success from a substantive (i.e., information processing, developmental, personnel selection, etc.) perspective. Otherwise, a model is of little more than academic interest.

Description of the Current Model

The approach to mental rotation performance suggested here is conceptually neutral in the sense that it can be applied to any circumstance where its assumptions are

met. The current model is based on the theory of mixture distributions, specifically a mixture of binomials distribution. Performance on this and other spatial tasks has been successfully modeled by a mixture of binomials distribution (Thomas & Kail, 1991; Thomas & Lohaus, 1993; Thomas & Turner, 1990; Turner, 1991). In these studies, two or more "kinds" of individuals were found whose performance was being measured. Once the structure of the data is evident, the model can be used to evaluate current theories or construct new ones. Further, individuals can be identified with respect to their performance group. Subjects' performance can be followed across time or across tasks providing insights, for example, as to whether improvement is best described as either incremental and gradual (i.e., does an entire group improve slightly on average?) or abrupt and stage-like (i.e., do individuals move from one group to another?).

A Model of Task Performance - The Univariate Case

The probability model is developed as follows. Let j represent the individual trials ($j=1, \dots, n$), and let i represent the individual subjects ($i=1, \dots, m$). Define a discrete random variable U such that $U=1$ for a "success" and $U=0$ for a "failure" for a given trial or item. Let $P(U=1) = \theta$, and $P(U=0) = 1 - \theta$, $0 < \theta < 1$. Now, define $X = \sum U_i$, where each U_i has the distribution of U , so that X represents the sum of the successful trials for any individual subject. Under the assumption that θ is constant across all j trials and all j trials are independent, X is binomial in distribution, and is described by

$$f(x) = b(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \quad (1)$$

The mean of X is given by

$$\mu = E(X) = n\theta, \quad (2)$$

where the symbol E denotes an expectation. The variance of X is given by

$$\sigma^2 = V(X) = n\theta(1-\theta), \quad (3)$$

and the symbol V denotes a variance. This model represents the performance of each individual i , but it must be expanded to include all m individuals. By assuming that θ describes the probability of success for all m individuals and that their responses are independent, the simple binomial model could be used. However, as anticipated by the research cited above, this model is expected to give a poor fit to the data. All individuals do not appear to have the same value of θ .

Different clusters or groups of individuals, conventionally termed components in this setting, might be characterized as having the same probability of success. The binomial structure could be preserved within each component, but different θ -values would be used to describe each group. This situation is described as a mixed binomial because two or more populations (each with its own binomial distribution) are mixed together in the general population.

For example, suppose the proportion π_1 of the population has scores distributed binomially with success probability θ_1 , and the remaining ($\pi_2 = 1 - \pi_1$) proportion had a different success probability θ_2 . The two component mixture model is formally given by

$$f(x; \theta_1, \theta_2, \pi_1, \pi_2, n) = \pi_1 b(x; \theta_1, n) + \pi_2 b(x; \theta_2, n) \quad (4)$$

Note that this model is simply the sum of two simple binomial distributions weighted by their respective proportions. The number of individuals in the first component is $m(\pi_1)$, and similarly $m(\pi_2)$ is the number in the second component. In general, any number of components is possible. The k component model is given by

$$f(x; \theta_r, \pi_r, n) = \sum_{r=1}^k \pi_r b(x; \theta_r, n), \quad (5)$$

where $0 \leq \pi_r \leq 1$, $\sum_{r=1}^k \pi_r = 1$, $0 < \theta_r < 1$, for all r . Within this model individual

differences are described by differences in θ -values. The main goal of this analysis is to estimate the different values of θ and their corresponding proportions in the population.

By convention, the θ_r -values are ordered such that $\theta_1 < \theta_2 < \theta_3 < \dots < \theta_k$.

It will be useful to refer to the mean and variance of Equation 5. The mean of X is given by

$$\mu = \sum_{r=1}^k n \pi_r \theta_r \quad (6)$$

and the variance of X is given by

$$\sigma^2 = \sum_{r=1}^k \pi_r [n\theta_r(1-\theta_r) + (n\theta_r)^2] - \mu^2 \quad (7)$$

Parameter estimation. The binomial $b(x; n, \theta)$ has two parameters, n , the number of trials which is always known, and θ , which must be estimated. A hat (^) over a parameter will denote an estimate of it throughout. The maximum likelihood estimate of θ is given by

$$\hat{\theta} = (1/mn) \sum_{i=1}^m x_i, \quad (8)$$

and an estimate of the variance is

$$V(\hat{\theta}) = [\theta(1-\theta)]/mn \quad (9)$$

with the estimated values replacing the parameters.

With k values of θ and $k-1$ values of π in the k component mixture model (Equation 5, p. 28), there are $2k-1$ independent parameters to estimate. The equations necessary to find maximum likelihood parameter estimates are given by Everitt and Hand (1981). Because they cannot be solved in closed form, iterative solutions must be obtained by using an Expectation Maximization (EM) algorithm. Given the vector of responses and guesses of the parameter values, less than 15 iterations usually produce stable estimates. As long as $m \geq 2k - 1$, as is the case here, identifiability is not a problem (Everitt & Hand, 1981).

Tests and confidence intervals. Blischke (1964) provides methods for computing estimates of the variances and covariances that permit hypothesis tests and confidence intervals. Hypotheses regarding any pair of θ parameters can be tested, for example

$$z = (\hat{\theta}_1 - \hat{\theta}_2) / [\hat{SE}(\hat{\theta}_1)^2 + \hat{SE}(\hat{\theta}_2)^2]^{1/2} \quad (10)$$

where SE denotes a standard error, and z is approximately standard normal under the null hypothesis that $\theta_1 = \theta_2$. Comparisons for π -values proceed similarly, using the appropriate covariance terms. Using the same quantities, the point estimate of any parameter ± 2 SE constructs an approximate 95% confidence interval.

Model selection and assessment. Model fit is generally improved with the addition of components. Parsimony and utility, however, dictate simpler models. Deciding on the number of components is usually regarded as a problem of selection, because no estimation standard exists for making this determination. To maximize both fit and simplicity, two indices were employed in model selection: Pearson's χ^2 goodness-of-fit statistic, and Variance Accounted For under the model (VAF). These indices are not always in agreement, so some judgment must be used. Fortunately, these model assessment criteria provide intuitive gauges of model viability.

Chi-squared. Pearson's goodness-of-fit χ^2 statistic measures the correspondence between the observed data and expected values computed from parameter estimates. While small expected cell sizes usually prohibit the assignment of p-values to computed χ^2 's, smaller χ^2 values indicate better model fit. In the case where the assumptions of the Pearson χ^2 are met, the degrees of freedom are equal to the number of cells minus one minus the number of estimated parameters. Degrees of freedom are provided throughout for all computed χ^2 's.

The likelihood ratio χ^2 , as defined by

$$L^2 = 2 \sum \hat{F} [\ln(\hat{F}/F)], \quad (11)$$

where \hat{F} refers to observed cell frequencies, and F refers to the cell frequencies expected under the model, and the cells are the outcomes of the summed variable X as defined above (Hagenaars, 1990). When the expected and observed frequencies are in close correspondence, L^2 is small relative to its degrees of freedom (which are the same as the Pearson chi-squared statistic). The likelihood ratio chi-square will be computed for two reasons. Because the Pearson χ^2 values cannot always be referenced to tabled values the likelihood ratio chi-squared statistic (L^2) will be provided for comparison. When both statistics are approximately equal, they provide stronger evidence for the acceptance or rejection of a model.

More importantly, the likelihood-ratio chi-squared statistic can be partitioned once the number of components (or latent-classes) has been determined (Hagenaars, 1990; McCutcheon, 1987). When restrictions are placed on a model's parameter values (e.g., $\theta_1 = .50$), degrees of freedom are recaptured and the chi-squared values increase.

The increase in L^2 for a "restricted model" is acceptable if it is small relative to the corresponding increase in degrees of freedom (i.e., if the additional variation is less than would be expected by chance) (McCutcheon, 1987). In this fashion, parameter restrictions can be evaluated.

Variance Accounted For (VAF). The familiar sample variance, s^2 , estimates the population variance. By replacing model parameters (π_r, θ_r) with their estimates in Equation 7 (p. 28), an estimate of the model variance σ^2 , can be computed. The ratio $\hat{\sigma}^2/s^2$ provides an estimate of the amount of variance accounted for by the model. Better fitting models will account for a higher proportion of variance. Occasionally, VAF is greater than 1 as a result of sampling error.

Complex model estimates occasionally suggest simpler models. For example, four component $\hat{\theta}$ -values might appear as $\hat{\theta}_1=.52, \hat{\theta}_2=.52, \hat{\theta}_3=.71, \hat{\theta}_4=.98$. These models are summarily rejected in favor of simpler models, regardless of fit statistics.

Classification. Once performance has been modeled by a mixture of binomials distribution, it is natural to consider partitioning subjects into the model components to which they "belong". In other words, for individuals with score $X=x$, which component is most likely to have produced that score? By computing posterior probabilities based on a maximum likelihood estimate of the parameters θ_r, π_r , individuals can be assigned to the components from which their scores have the highest probability of having been drawn. The probability that score $X=x$ "came from" the r th component is given by

$$P(r|x) = \frac{\pi_r b(x; \theta_r, m)}{\sum_{o=1}^k \pi_o b(x; \theta_o, m)} \quad (12)$$

where $o = 1, 2, \dots, r, \dots, k$. Parameter values are, in practice, replaced by their estimates.

Component assignments on two or more tasks provide bivariate data useful in assessing between task correlations.

Model assumptions. The assumptions of constant θ and independent trials which underlie the proposed model have been usefully employed in similar settings (Thomas & Lohaus, 1993; Thomas & Turner, 1991), and would seem to be appropriate here as well. At issue is not whether the assumptions have been violated, certainly they have. At issue is whether they have been violated to the extent that they render the model useless. It is argued here that they have not, and evidence to this effect will be presented below.

Constant θ . The idea that each subject has an equal probability of success on each trial has some intuitive appeal. An individual with a given ability level might be expected to perform similarly on items of equal difficulty. Egan (1979) argues that items of like complexity should be subject to the same mental processes, while items of differing complexity are likely to be subject to different processes. The items of the mental rotation task are quite similar to one another because only a small number of similar objects and their mirror images are used. This implies that an equal probability assumption is not unreasonable. However, there is some evidence that accuracy is lower for larger rotation angles (Berg, Hertzog, & Hunt, 1982; Just & Carpenter, 1985; Robertson & Palmer, 1983), and that foils are more difficult than identical items (Pellegrino & Kail, 1982). As noted above, however, these differences may be small enough that they do not affect the model in any important way.

Ideally the equal probability assumption would be tested using multiple identical trials for each subject. Unfortunately, replicates (i.e., multiple responses to the same item within the same test administration) were not obtained. However, a rough estimate of the degree to which this assumption is violated can be obtained by calculating the proportion of correct responses for each item within each component (Thomas & Lohaus, 1993). If the proportion of correct response estimates (\hat{p} 's) are constant within sampling error, confidence can be placed in the model's descriptive power.

Independence. For the binomial or mixed binomial model to hold, each subject's trials are assumed independent of his or her other trials. Unfortunately, this issue is not easily decided. Consider the subjects selected by Shepard and Metzler (1971), whose accuracy performance was nearly perfect. In some sense their accuracy on early trials predicts their performance on later trials, suggesting a lack of independence. However, this dependence is a functional dependence, not a stochastic one. It is the latter that is of concern here. Stochastic independence demands that the outcome of a trial is neither influenced by previous trials nor influences subsequent ones. Again reflect on Shepard and Metzler's (1971) subjects. It seems unlikely that responses somehow affected each other in the absence of feedback, because high ability subjects would have no reason to refer to previous problems. Lohman and Kyllonen (1983) cite research to support this contention. Thomas and Lohaus (1993, p.145) provide the following analogy. Imagine a bag of marbles each marked with a 1. Choose a marble at random and define a success as selecting a marble with a 1 on it. Repeat this process. The probability of drawing a marble marked with a 1 is 1, but there is certainly no trial to trial dependence: the outcome of trial n in no way affects the outcome of trial $n+1$. The distribution of

accuracy scores demonstrated by Shepard and Metzler's (1971) subjects could be viewed similarly.

It is difficult to decide the issue of functional versus stochastic dependence reflectively. Under an assumption of stochastic independence, performance on individual trials should be uncorrelated with one another. In other words, within sampling error each trial Y_j should be uncorrelated with each other trial Y_l , $l \neq j$, implying the correlation between l and j , $\rho_{lj} = 0$, within each component. For example, for the lowest performing group, $\rho_{Y1, Y2} = \rho_{Y1, Y3} = \dots = \rho_{Yn-1, Yn} = 0$. Under the null hypothesis $\rho = 0$, $r \pm 2 SE_r$, where $SE_r = [1/(n-1)]^{1/2}$ should contain 95% of the probability mass of $\rho = 0$ (Huber, 1977). The proportion of observed correlations for all of the trials outside this confidence interval can provide information about the validity of the independence assumption, and thus the amount of confidence to place in the model. But again, at issue is whether the general conclusions concerning the structure of the data are in jeopardy because the model's assumptions have been violated. Simulations by Thomas and Lohaus (1993) provide reason to believe that minor violations have little consequence.

The model presented here attempts to provide insight into the nature of mental rotation performance. It is argued that while the assumptions necessary for a binomial mixture are undoubtedly not met, they are perhaps not so flawed that the basic conceptualization of performance groups should be rejected.

Data analysis strategy. The general strategy is quite simple. Male and female subjects from both Penn State and Cooper-Union responded to 24 Vandenberg and Kuse (1978) type mental rotation items (Old) and 24 more complex items (New). Half of the items required a "same" response and half a "different" response. Individual trials for each subject's "same" and "different" responses to the 48 mental rotations items were scored as either successes or failures (See Chapter II below for a more complete description of the dataset). The number of correct items for each subject were summed for each of two mental rotation tasks (Old and New), yielding two summary scores between 0 and n , where n is the total number of within subject trials for each task. Estimates for the model were generated by an EM algorithm enabling models of increasing complexity up to four components to be fitted for males and females separately on each task. Three measures of model fit were used to determine the most appropriate model. Estimates from the best fitting models were used to evaluate the hypotheses described below.

A Model of Performance Change - The Bivariate Case

The data pairs (x_i, y_i) , $i=1, \dots, m$, are of focus. Here, X and Y denote the two tasks of interest, either task 1 and task 2 at a fixed time or either task at time 1 and time 2. The joint distribution of X and Y , $f(x, y)$ is naturally conceived as bivariate mixed binomial given that the marginal distributions $f(x)$ and $f(y)$ are mixed binomial. In general, the number of components in the bivariate model is the product of the number of components in the marginals. Assuming that the number of components, k , is the same for both tasks (or across time), the bivariate model is given by

$$f(x, y) = \sum_{a,b=1}^k \tau_{ab} \pi_{xa} b_{xa} b_{yb} \quad (13)$$

where $a=1,\dots,k$ and $b=1,\dots,k$; $b_{xa} \equiv b(x; \theta_{xa}, n_x)$ and $b_{yb} \equiv b(y; \theta_{yb}, n_y)$ are familiar binomial distributions for each task. The π_{xa} 's represent the univariate mixing proportions on task X, and the τ_{ab} 's represent transition parameters or state change probabilities across tasks (or time). These are conditional probabilities. That is, τ_{ab} is the probability of "moving" to component b of the Y variable having been in component a of the X variable. Note that $\sum_{b=1}^k \tau_{ab} = 1$ for all a. For example, τ_{12} might denote the probability of to the high group at time 2 given having been in the low group at time 1. More useful in this case than transitions from a component on X to a component on Y are the joint proportion estimates $\hat{\pi}_{ab}$ which represent the proportion of subjects simultaneously in component a on X and b on Y. For example, $\hat{\pi}_{11}$ might represent the proportion of subjects in the low performance group at times 1 and 2. As with the univariate model, $\sum_{a,b=1}^k \pi_{ab} = 1$. The full development of the models for two and three component marginal distributions is presented in Appendix B.

Parameter estimation. Maximum likelihood estimated probabilities of success, $\hat{\theta}_{rs}$, and mixing proportions, $\hat{\pi}_{rs}$, for the bivariate distributions are chosen from the best marginal models. State change parameter estimates are solved using iterative techniques. Maximum likelihood estimates can then be used to test specific hypotheses.

Figure 1 graphically depicts transition parameter values under a no-learning hypothesis, testable under the model. In this simple example, τ_{11} , τ_{22} , and τ_{33} , are the proportions of subjects who, given having been in components 1, 2, or 3 remain there on the second assessment. Note each value is 1. Other, more interesting hypotheses can also be evaluated, for example it is possible to test whether performance decreases over time.

Model assessment. Model complexity is limited by the marginal models (i.e., two component marginal models imply four joint components). However, simpler models are preferable. The Pearson χ^2 goodness-of-fit statistic was used as before to distinguish between bivariate models. In general, lower values are better. Considerations of parsimony dictate that reasonably fitting simpler models should be accepted in favor of more complex models. Even though small cell sizes make it hazardous to provide p-values, computed χ^2 's that are small in relation to their degrees of freedom are preferred.

Intertask correlations. In the mixture setting, bivariate correlations have a very different interpretation than the usual Pearson r. This model assumes local independence, a feature of latent class models in general which specifies that by conditioning on any marginal component (or latent-class) of one variable the scores of variables (X and Y) are independent. For example, from Equation 13 (p.36) suppose that $k=2$, resulting in 4 bivariate components. Of these 4, focus on the component defined by $a = 1$ and $b = 1$. The observations on X and Y from within this bivariate component (and within the remaining 3 components) are independent and thus uncorrelated. In effect, the overall observed correlation between X and Y is the result of component membership correlation rather than an association between the two task variables within a component of X and a

component of Y. This example does not imply that the two variables X and Y are not related in the mixture setting. To the contrary, it suggests that the relationship is caused by an unobservable latent variable: that is the "class-like" structure of X and Y. Both of which are the product of the same latent variable responsible for the functional dependence across trials.

Summary of Model Approach

The approach outlined above attempts to find optimal univariate and bivariate mixtures of binomials models to describe individual differences in performance on two mental rotation tasks collected over two time periods. The procedure separates individuals into performance groups in an effort to measure performance change. The model parameter estimates can then be used to evaluate specific hypotheses concerning performance and performance change predicted by various theories in a way that traditional, normal-based statistics cannot.

Connections to Other Models

This type of model can be viewed as a form of cluster analysis (McLachlan & Basford, 1988) or from an item response theory framework (Lord, 1965). Its relation to item response theory will be developed in greater detail in the following chapters. This model also shares many similarities with more general latent class analyses, where each mixture component is a latent or unobservable class of performers. The kind most familiar to the social sciences are those developed in Lazarsfeld and Henry (1968) and McCutcheon (1987).

Thomas and Kail (1991) posited a conceptually similar mixture model to explain mental rotation latencies. They regarded reaction times within a mixture of normals

distribution framework, and provided an X-linked genetic explanation to account for differences between performance groups. Their model, like this one, recognizes that some subjects of both sexes perform at very high levels, while others perform quite poorly. In past modeling studies of spatial ability, sex-differences seemed to be accounted for largely in the relative proportions of each sex within those performance groups; the same is expected here.

The bivariate binomial mixtures employed here are also similar in spirit to Markov chains (e.g., Brainerd, 1982). Performance in a Markov model is viewed in terms of performance states similar to the performance components in the mixture model. Both parameterize transitions between those performance categories. However, there are fundamental distinctions between the two as well. State changes in Markov models are based on subjects' rule-sampling. It is difficult to see what rules could be successfully applied to the mental rotation task, especially when no feedback is provided during testing.

Statement of Purpose

Purpose

A model is developed to describe accuracy on mental rotations tasks, necessitating a fundamental change in the way performance and individual differences are traditionally viewed. Having characterized performance from a modeling perspective, long-standing substantive questions regarding individual differences in performance and performance change over time can be addressed. This study builds on the work of Thomas and Turner (1990), Turner (1991) and Thomas and Lohaus (1993). It will attempt to model individual differences in mental rotation performance, but it will

do so on two tasks of differing complexity to assess the effects of item difficulty on performance. It will also attempt to describe spatial task performance change over time, providing more insight as to the nature of change than was possible using cross-sectional data (Thomas & Lohaus, 1993; Thomas & Turner, 1991). While the model proposed here will not be applied to developmental data per se, because performance change over time is modeled, one natural extension of such a model would be to evaluate developmental phenomena such as the development of spatial skills. Past research (Turner, 1991) suggests that a mixed binomial model will successfully describe mental rotation performance. Because there is some indication that a three component binomial mixture model will provide the best fit to the data, a three component bivariate binomial mixture model was developed (See Appendix B).

Research Questions

At the most basic level, this research is guided by whether the proposed model can adequately describe performance and offer insights into the nature of individual differences. If so, issues related to theoretical predictions can be addressed, including whether or not the novel CAD curriculum provides a training advantage over traditional graphics course-work. The following five inter-related questions guide the research:

Performance Under the Model.

- 1. How does performance change over time (within each treatment group) and how can that change be modeled?**
- 2. How does stimulus complexity affect performance (with respect to component structure, change over time, sex, and treatment group)?**

Strategy Use.

3. Can strategies be inferred from the proposed model structure?

4. How is strategy use related to item complexity?

5. Are there changes in strategy use over time?

Design Overview

To answer these questions, first-year engineering students were trained in object rotations over the course of a 12 week semester using either a traditional or novel curriculum. This novel treatment is intended to take advantage of the long-term, self-guided practice of rotations with feedback on a CAD-based computer program (See Table 1 and Chapter II). This intervention should be general enough that it will not force non-visual problem solvers to change their strategies, but allow them to refine their current ones. As a result, it should promote more automatic and efficient processing in relation to natural strategies for all subjects (Kyllonen et al., 1984). Assessment consisted of an adapted version of the Vandenberg and Kuse (1978) mental rotation task administered at the beginning and end of the academic semester. The current mental rotation task used 24 items taken from the Vandenberg and Kuse (1978) task and 24 items made of more complex shapes (See Appendix A).

Hypotheses

1. How does performance change over time (within each treatment group) and how can that change be modeled?

Because the tasks used during training (engineering-related design objects) and testing (Shepard & Metzler, 1971 abstract shapes) vary to such a large degree, those

theories positing changes in procedural knowledge should predict performance change, while those espousing changes in declarative knowledge should not.

Olson and Bialystok's model predicts that practice of the sort used here, which does not focus on the same objects for training and test, should prove ineffective. There is not enough experience with the objects used at test to become familiar enough with their structural features. As a result, the structural features should be unavailable, at least for initially poorly performing subjects, when attempting transformations like rotation.

Piagetian theory, however, might predict otherwise. If development proceeds based on the internalization of actions which result in the formation of operations, then the same activities which produce change in childhood might be expected to be effective even in college students if the length of the intervention is sufficient in the same way that hypothetico-deductive reasoning might be taught at the college level. The opportunity to manipulate object transformations and form internalized imitations based on perceived images from the computer screen should allow rotation to become internalized. The most important aspects of this training is that it is subject controlled, thus allowing for exploration, manipulation (i.e., action) and imitation, and that it is long-term. In sum, support is expected for Piaget's theory but not for Olson and Bialystok's theory.

Performance change over time will not be identical for both the treatment groups, however. Because Piaget's theory predicts that the type of interaction offered by the novel curriculum will allow for the formation of operations, it is hypothesized that the subjects who received the novel intervention will have state change proportions that reflect greater improvement than their traditional curriculum counterparts. Because Piaget's theory predicts that the type of interaction offered by the novel curriculum will allow for the formation of

operations, it is hypothesized that the subjects who received the novel intervention will have state change proportions that reflect greater improvement than their traditional curriculum counterparts. This type of improvement is characterized in Figure 2, which shows consistent probabilities of success (θ 's) and increasing proportions for the "high" level group (π_2).

2. How does stimulus complexity affect performance (with respect to component structure and change over time)?

Previous research has demonstrated that stimulus complexity adversely affects accuracy (e.g., Yuille & Steiger, 1982). Presumably, the increased processing demands associated with complex problems cause subjects to resort to less efficient strategies, which in turn lower accuracy. If subjects' strategies are well captured by θ -values in the mixture setting, then their strategy shifts will also be well described by the model. For any fixed time, subjects whose performance decreases on the more complex task should shift from a higher component to a lower one. In other words, the same components should be observed over both tasks, but the proportion of subjects in the highest performance groups on the standard task will be lower on the more complex task. This hypothesized relationship will be observed in the state change probabilities. Those probabilities will be highest for joint components which reflect performance declines. This relationship is expected to hold because strategy use is affected by complexity.

Alternatively, as Kail and Pellegrino (1982) point out, a decrease in accuracy might simply be the result of a greater number of processing operations when stimuli become more complex. If this is true, a very different set of results would be expected. Since each process would have a small likelihood of error, the compounding of the processes would imply that complex problems will be associated with lower probabilities

of correct response in general. From an analysis of variance perspective, this might be expected. In general, the average score is lower for more complex items pooled across subjects. In this case, subjects are likely to remain in their relative groups, but the probabilities of success for all groups are likely to be lower for more complex items. This, however, is not anticipated. Instead, it is hypothesized that the pooling of subjects is providing an inaccurate picture of the effects of stimulus complexity, and that a more abrupt state change (i.e. a change from one component to another across task) is more likely.

Strategy Use.

3. Can strategies be modeled?

One of the most significant questions concerns the discovery of the different strategies used to solve mental rotation problems. There is some research to indicate that there are multiple strategies used in solving this type of problem (Just & Carpenter, 1985; Sedon et al., 1984; Mislevy et al., 1991). If there are multiple strategy groups, can the quantitative differences in model parameter estimates be linked to qualitative differences in processing?

The data used in the present study are not amenable to direct evaluation of the strategies subjects use. However, some implications of strategy use might make themselves apparent. Recall that the model posits that within each component, each item has the same probability of success associated with it. Should a mixture of binomials model describe performance, then it is difficult to see how any single component (defined by a single probability of success value) could be the result of an amalgamation of strategy groups (i.e., that two different strategies would have the same probability of

success over all items). With sufficient power, however, strategies with similar probability of success should be identifiable. A similar model has already been shown capable of clarifying strategy differences by viewing them in terms of component groups (Thomas & Lohaus, 1993).

For instance, Turner (1991) found a significant proportion of subjects whose probability of success hovered around .50. On a binary choice task like the one used here, this performance is at chance levels. It is argued that this was exactly the strategy employed by these subjects, although other types of response bias should be excluded before this interpretation is accepted. This example seems the most intuitively obvious instantiation of how the current model framework might be used to understand subjects' strategies, but other hypotheses concerning subjects' strategy use can also be investigated.

If the probabilities of success (θ -values) remain constant over the two difficulty levels of the task (e.g., differences between the standard items and the more complex items are characterized by π -value differences), then evidence that each component represents a strategy group will have been found. Each strategy is assumed to have a constant probability of success associated with it, regardless of item complexity. This is similar to the way rule use in other tasks is conceptualized in that item complexity will only affect subjects' strategy choice. While the probability of success values (θ -values) are expected to remain constant over different versions of the test, the mixing proportions (π -values) should change, indicating that subjects' strategies are dependent on item complexity. This assumption is doubtlessly incorrect, but hopefully not importantly so.

Lohman and Kyllonen (1983) provide evidence in support of the hypothesized effect of item difficulty on strategy use.

Moreover, differences in general aptitudes in visual and verbal modes of processing (Kyllonen, Lohman, & Snow, 1981, cited in Lohman and Kyllonen, 1983) have been implicated in strategy differences. These authors suggest that visualizers are more likely to use rotation strategies to solve mental rotation problems, while verbalizers are more likely to use piecemeal rotations or less efficient verbal codes. This finding implies that component groups should be distinguishable based on verbal and mathematical aptitude and achievement test scores like the SAT. SAT, cumulative GPA's and college placement scores were available for Penn State subjects. It is hypothesized that the highest performing component groups will have the highest math-related scores, while one or both of the lower performance groups will have superior verbal-related scores. Finding differences between component groups on these achievement scores also adds some validity to the components, indicating that individuals within the same component groups are, in fact, similar to one another and different from individuals in other component groups.

4. How is strategy use related to complexity?

As Linn and Petersen (1985) point out, some of the differences in accuracy scores over different versions of mental rotation tasks (e.g., PMA space versus Shepard & Metzler objects) are thought to be caused by the failure of less able subjects to implement the same strategies on difficult items that they apply to simple items. Lohman and Kyllonen (1983) provide additional support for this hypothesis, noting that strategy changes are related to item characteristics. It has been suggested that stimulus

complexity can reduce imaging efficiency and exceed short-term memory capacities in less proficient subjects. For example, Carpenter and Just (1978; Just & Carpenter, 1985) outlined an account of mental rotation that describes different mental rotation strategies. One of these strategies is the part-by-part rotation of figures. They argue that more complex figures will require longer "rotation" times because several rotations are required. Similar findings were reported by Bethel-Fox and Shepard (1988). This idea also receives some support from Kail et al. (1979) who found that a significant subsample of females had much longer rotation times than males. Tapley and Bryden (1977) suggest that when subjects are forced (due to the task demands) to use piece-by-piece strategies, individuals may have difficulty keeping the parts of an object in proper relationship to one another. Considering the relationship between rotation time and accuracy, it follows that more complex stimuli will be judged less accurately. Linn and Petersen (1985) note that tasks requiring more analytic strategies, such as a part-by-part rotation, show minimal sex-differences. Additionally, in tasks which use relatively simple objects, no sex-differences in strategy use are detected (Kail et al., 1984). If females are more likely than males to adopt a part-by-part strategy as item complexity increases, as this evidence suggests, then one would hypothesize that sex-differences in accuracy will increase with corresponding increases in task complexity. This hypothesis will be investigated with reference to the univariate analyses. It is hypothesized that the performance of both sexes will change similarly within component groups. In other words, the univariate component structure will look similar at both time points for both men and women.

This type of hypothesis is easily tested under the current model. Specifically, an increase in sex-differences due to stimulus complexity would be observed by comparing male and female π -values for the standard and complex items. A greater relative decrease in the π -value for the highest performing group and a corresponding increase in the π -value(s) for the lower performing group(s) of women across the two tasks would provide evidence for this effect. The relevant transition parameters should also be revealing.

The current hypothesis accounts for the apparently sex-related complexity effect in terms of a function of general performance level. Because a greater proportion of males are in the highest performing levels this effect appears to be a between-sex difference when mean-based analyses are used. If stimulus complexity affects poor performers more than good performers regardless of sex, then a different (but still testable under the current model) pattern in the parameter estimates will be observed. One potential instantiation is consistent θ -values over standard and complex items, but lower proportions of subjects in the highest performance group, regardless of sex.

In summary, subjects' θ -values will decrease (from a higher component to a lower one) as task difficulty increases, and increase (from one component to a higher one) with practice. If probability of success values (θ -values) remain constant across task and proportion estimates (π -values) change over task, one explanation for this result would be that subjects who perform at high levels on an easier task are unable (due to processing demands) to implement the same strategies on more difficult items. Presumably, those subjects performing at uniformly high levels on both versions of the

task are able to use the same strategy regardless of what item is presented. In this study, each strategy group is expected to be defined by a constant θ -value.

5. Are there changes in strategy use over time?

Because performance change on other spatial tasks was found to be the result of changes in component membership rather than incremental changes in probabilities of success (θ -values), the same is predicted here. If strategy use defines component performance, then changes in strategy use will also be seen as a result of changes in component membership. It is difficult to conceive of strategy changes that would be manifested in small shifts in θ -values. If the CAD-based intervention allows for improvements in procedural knowledge that affect strategy use, performance will improve abruptly on both versions of the mental rotation task, as manifested by changes in π -values.

If the current intervention is too visually laden for subjects in the lower performance group(s) (who may be using verbal rather than visual strategies), then one would expect to see good performers get better and poor performers get worse (i.e., more extreme θ -values over time). Model parameters consistent with this hypothesis are provided in Figure 3. The groups of performers defined by strategy use at pretest and identified by the mixture model should maintain their members, while each group is affected (either positively or negatively) by the treatment.

Finally, this study should also provide suggestions for directions in future research which address important questions concerning the relationship between spatial abilities, strategy use, and sex-differences in spatial competence.

CHAPTER II

Method

To answer the research questions posed in Chapter I, an existing dataset was analyzed. Data collection procedures from a curriculum evaluation in The Pennsylvania State University's College of Engineering are summarized below. Note that while the focus of the curriculum evaluation was on pedagogic technique, this study proposes to evaluate a model for performance and performance change and relate model-based findings to psychological theory. As such, the intervention will only be briefly summarized.

Participants

First-year engineering students ($n=556$), enrolled in design courses at The Pennsylvania State University and The Cooper-Union College, were provided with either a traditional or novel engineering graphics curriculum. Table 1 summarizes the composition of the participants. The sample was comprised of 409 men and 147 women, of whom 393 were attending the Pennsylvania State University, and 163 were attending the Cooper-Union College. Because the data were gathered from naturally occurring course sections, no effort was made to control for the sex, age, or institutional affiliation of the students. In addition, not all subjects were available for testing at both time points in these naturally occurring classroom groups. As such, subjects available only at pretest were included in all univariate analyses, as were subjects available only at posttest. Those subjects who provided data at both pretest and posttest (represented by the

overlapping portions of the ellipses in Table 1) provided suitable data for the bivariate analyses.

Because treatment group and university affiliation were confounded, basic descriptions of the two populations from which the subjects were sampled are provided. Average admission SAT scores and middle 50% range SAT scores are provided for both Penn State University and Cooper-Union College first-year student populations in Table 2, demonstrating the potentially higher ability level of the Cooper-Union sample subjects. One concern is that performance differences at post-test are confounded because the two treatment groups were not identical at pretest. It is argued that because the Cooper-Union students were, in general, academically superior to their Penn State counterparts, performance differences would be expected to favor the Cooper-Union subjects in the event that the curriculum intervention was unsuccessful (e.g., a greater proportion of Cooper-Union subjects would be expected to be in the component associated with high-level performance at both time 1 and time 2 if there is no advantage to the new curriculum). However, if the novel curriculum proves more effective (in the sense that a significant proportion of subjects change from components associated with lower performance to ones associated with higher performance), it is argued that such a change was in spite of the advantage that the Cooper-Union subjects appeared to have as the result of the college's potentially higher selectivity.

Tasks

To evaluate the effectiveness of the curriculum change, a modified version of the Vandenberg and Kuse (1978) mental rotation task was used. The adapted version contained 12 items: six taken from the Vandenberg and Kuse (1978) version of the

Shepard and Metzler (1974) mental rotation task and six made from more complex three-dimensional structures (See Appendix A). Each item contained four comparisons to a target figure. Figures appeared within a 2.5 cm (1 in.) circle, reproduced so that there were five items on each line, with the target item always presented first. Subjects were asked to judge whether the comparison items were the same as or different from the target items. Comparisons were presented four items to a page on three pages.

One significant change in the Vandenberg and Kuse (1978) mental rotations test format was made. Each item required the comparison of four figures to a target figure. Of every four comparisons, two and only two figures were identical to the target item (albeit rotated). As a consequence, solution strategies not related to spatial ability are possible. Once a subject has determined which two items of the set of four were the same as the target item, the remaining two need not be rotated because they can be logically eliminated as potential matches to the target figure. In effect, once any three items of the four have been solved, the fourth is logically, as well as spatially determined.

The mental rotation tasks used here were adapted so that half of the items overall were the same as and half different from the target stimulus, but each block of four contained a random number of "same" items. As a result, each block contained from 0 to 4 items which required a "same" response, and there were approximately an equal number of each of the five possible block types. Consequently, subjects were forced to solve each of the four items individually.

The Vandenberg and Kuse (1978) paper and pencil task is typically administered with a strict time limit in an attempt to induce rotational strategy use. Overall scores are then computed from the number of items answered correctly of those attempted, minus a

penalty factor for guessing. This scoring scheme was not adopted in the present study for three reasons. First, because subjects do not attempt the same number of items, differences between protocols in the number of items attempted make interpretations difficult. Also, because the model proposed above requires the same number of items for each subject, this scoring system is a priori inappropriate. Second, researchers have found that some of the male superiority in accuracy is caused by a greater number of errors of omission by females (Goldstein, Haldane, & Mitchell, 1990; Stumpf, 1993). It can also be argued that a time limit changes the focus from accuracy to speeded accuracy, which is a different skill entirely (Lohman, 1979, cited in Eliot, 1987; McGee, 1979). Both problems were reduced by allowing subjects enough time to complete all items. Finally, the model presented above is a stochastic one, and as such, can account for guessing responses, obviating the need for a correction factor. Items were scored for the number of correct "same" and "different" judgments from 0 to 24 for each set of items, standard and complex.

In addition to mental rotation performance, 10 measures of achievement were also collected from Penn State subjects. SAT Math and Verbal scores, three college placement tests in mathematics, one in chemistry, and one in English, and high school and college cumulative GPA's were gathered because performance on these measures has been related to spatial ability.

Procedure

Intervention

The curriculum change employing the novel CAD program attempted to incorporate the findings of previous training studies into a comprehensive procedure.

Usually, a relatively crude wire-frame CAD program or manual drawing instruction is used in teaching the design courses. Improvement in spatial visualization skill was attempted by capitalizing on the ability of a CAD-based solid-modeling computer program (Silver Screen) to provide constant feedback while demonstrating three-dimensional rotations in motion on a two-dimensional screen. In addition to rotations, the software has the ability to show sections of solid objects on screen and provide multiple views of the objects simultaneously. Moreover, the experimental opportunity was provided for a full semester, allowing for long-term practice. Each section of the Engineering Graphics 50 course spent approximately 22% of class time (approximately 8.25 hours) interacting with the novel software.

Design

A quasi-experimental design was used to evaluate the usefulness of the solid modeling technique as a training method. Students from Cooper-Union College were given either standard wire-frame CAD software or manual drawing instruction, providing a contrast group (traditional course), while Penn State students were given the solid-modeling-based instruction (solid-modeling). The courses at the two universities were similar in almost every other regard. A pre-test was administered to all students within the first two weeks of the semester, prior to any in-depth study of engineering design. The pre-test consisted of the 12-item mental rotation task described above. The post-test, a second administration of the mental rotation task, was conducted within the last two weeks of the semester after the graphics curriculum had been presented. Subjects' performance was tracked over time using Student I.D. numbers when given. Unfortunately, a large proportion of subjects chose not to provide this information. As

such, many of the subjects were tested on both occasions, yet their performance could not be evaluated over time. No other demographic data were collected.

Assessment

Data were collected over the course of three semesters from subjects in classroom groups of 25 to 35 students for both assessments. Subjects completed the tasks without a time limit. This was done for three reasons: (1) to ensure that nearly all subjects would complete the task (2) so that accuracy judgments would not be confounded by response time differences, and (3) because spatial visualization problems that naturally occur in the engineering field seldom require timed rotation, any attempt to improve or measure performance should be as similar as possible to the skills used in engineering.

Response Measures: Reaction Time Versus Accuracy

Both reaction time and accuracy scores are commonly used response measures for assessing mental rotation performance (Linn & Petersen, 1985). There is evidence to suggest that reaction time judgments and accuracy judgments measure different aspects of mental rotation skill. It is argued below that accuracy measurements are superior in the present context.

It is possible that a speed-accuracy trade-off is responsible for some of the sex-differences typically found when latencies are measured. However, Cooper (1982), Lohman (1986) and Tapley and Bryden (1977) have all provided evidence that this is unlikely because speed and accuracy are correlated. Evidence suggesting that both accuracy and speed of rotation are dependent on representation quality supports the notion that both measures characterize ability (Kail, Stevenson & Black, 1984; Lohman, 1986).

It has also been shown that while speed and accuracy of rotation are correlated, latencies will not measure analog rotation processes in subjects who use non-rotational strategies. (Tapley & Bryden, 1977). Researchers using reaction times to measure performance on mental rotation tasks often infer that larger angular deviations require more time to rotate. Because reaction time for subjects using non-rotational strategies cannot be viewed in the same way, reaction time may not be as good a measure of ability as accuracy.

In studies where speed and accuracy were not related, accuracy was found to be a superior measure of ability. Egan (1978, cited in Cooper and Mumaw, 1985; 1979) detected independent factors for accuracy and latency scores, and found that accuracy scores were more predictive of success in aviation than reaction times were. Lohman (1988) has found that individual differences in the speed with which problems are solved does not predict accuracy on complex problems. Merriman, Keating, and List (1985) provide psychometric support for preferring accuracy judgments in measures of spatial ability as well. In sum, performance models based on response times from a subject pool restricted to those who perform almost without errors (e.g., Shepard and Metzler, 1971) can not provide a generalizable account of performance (Lohman & Kyllonen, 1983); accuracy must be assessed.

A final consideration is the practical advantage to using accuracy scores. Reaction time studies can only use data from "same" comparisons. Because there is no rotation that can bring different items into congruence, response latencies for these items are difficult to interpret in terms of mental rotation. In effect, half of the data are unavailable to analysis.

Some of the objections leveled at latency measures could be directed at studies using accuracy judgments, especially because accuracy does not imply that any mental rotation has taken place. The focus of this study, however, is not whether an analog imaging process occurs as much as that an important spatial ability is being measured. Lohman (1979, cited in Cooper & Mumaw, 1985; 1986) points out that superior accuracy in rotation of complex objects is, in part, what it means to have superior spatial ability.

CHAPTER III

Data Description

The data for this study were gathered from two groups, each provided with a different curriculum in their first year engineering courses. Participants from the Pennsylvania State University had training which involved special emphasis on rotation skills, while participants from The Cooper-Union College received no such emphasis. Subjects from both groups responded to 24 Shepard and Metzler (1971) mental rotation objects and 24 more complex objects of the same type, denoted Old and New items respectively (see Appendix A). Correct responses to the old and new items involved either a "same" or "different" response to a target. To measure the effect of the curriculum difference, pre- and post-test data were gathered. In addition, the sex of the subject was recorded.

These five variables (curriculum - Penn State/Cooper-Union, Sex - Male/Female, Item type - Old/New, Item status - Same/Different, and Time - 1/2), each with two levels, provide 32 basic-level groups (item-sets) used in the modeling procedure below. Curriculum (University) and sex are between subjects variables, while Item type, Item status, and Time are within subjects variables. Each of these samples is referred to using the short-hand notation outlined in Table 3. Within each of the 32 basic-level groups, the number of correct responses was summed for each individual as described above (p. 26; i.e., an observation on X), to provide a score indicating the total number of correct

responses for that individual on each subset of items (e.g., old-same items at time 1, old-different items at time 1, etc.).

To provide an overall sense of the data, descriptive statistics are provided below for each of the 32 samples. That is, for each sex and curriculum (university) group Table 4 indicates the mean number of items correct, sample size, variance, and range of responses for each item type, item status, and test time. The possible range of responses (i.e., the number of items attempted) is not equal for all of the 32 basic-level groups, and as such the means and ranges presented in Table 4 are directly comparable only in the cases where the number of items are equal. Recall that while half of the items were old and half new, and half were the same and half different, item type and item status were not evenly counterbalanced. Of the old items, 15 were different from the target items, and nine were the same. Conversely, of the new items, 15 were the same as the target item, and nine were different. The overall model interpretation, however, is not affected by the number of trials per task.

More revealing than the descriptive statistics, frequency histograms of the number of items correct are provided for four of the 32 basic-level groups in Figures 4-7. These particular figures were chosen only because they are representative of the 32 basic-level histograms in all important respects. In addition to the observed frequencies provided in the figures, a normal curve is shown. Estimates for the normal were obtained by substituting sample means and variances for their corresponding population values. Best fitting two and three component binomial mixture estimates are also provided for comparison. The mixture estimates will be considered in more detail in Chapter IV.

Most conventional analyses rely heavily on two assumptions: normality and the idea that group differences are manifested as additive differences. These issues are considered in turn.

Figures 4-7 suggest that these data were not sampled from a unimodal population like the normal. In all cases, the mixture models seem to preserve the features of the data more adequately. Chi-squared goodness-of-fit test statistics are almost always lower for the two component binomial mixture model, indicating a better fit relative to the normal. Even in cases where normality cannot be rejected on the basis of chi-squared values alone, the histograms demonstrate the clear superiority of the mixed binomial model. Goodness-of-fit statistics are provided for each of the three models presented in Figures 4-7 and the other 28 basic-level performance groups in Appendix C.

One might, at this point, argue that t-tests are robust to departures from normality, and that an argument against conventional analyses on those grounds is unwarranted. In many cases this is true, however, the problems with conventional analyses are more serious than the failure of the data to have come from normal distributions. In the case of sex-differences, a t-test procedure assumes that male and female distributions are identically normally distributed, except that one of the distributions is shifted to one side. Figure 8 graphically illustrates the t-test model usually applied to the sex-differences found in spatial abilities. Clearly, if the female performance distribution in Figure 8 is shifted to the right, it is identical to the male performance distribution, and consequently an additive shift model is appropriate. It was argued earlier that this type of model is inappropriate for the data under investigation because shift-models do not adequately describe differences between performance groups for the data presented here. The data

presented in Figure 9 show Penn State Females' performance on New items that required a "different" response at time 1 and time 2, with the time 2 data scaled to account for the difference in sample sizes. The mean number of items correct increased over time from 6.1 to 7.0. A shift-model, such as a t-test, indicates that, within sampling error, the observed frequency distribution at time 2 is identical to the frequency distribution at time 1 when it has been shifted to the right by .9 items correct. It is difficult to see, however, what constant could be added to time 1 performance in Figure 6 that would make the two frequency distributions "line-up." In fact, as Thomas and Lohaus (1993) have pointed out, any time observations are bounded on some interval, a shift model is likely to fail, simply because scores cannot be shifted past the boundaries of the maximum and minimum number correct. In other words, t-tests and models like it, which account for mean differences in terms of distributional shifts require unbounded population distributions. Consequently, they fail when the measurement system prohibits this. In addition, the very use of a mean in describing a distribution implies that a measure of central tendency captures some important characteristic of the data. In the case of the data presented in Figure 9, this spirit seems to have been violated. The analysis of these data from a binomial mixture model perspective is considered next.

CHAPTER IV

Univariate Binomial Mixture Analyses

The overall strategy is to fit the data to binomial mixtures with one, two, three, and four components. Model estimates generated by the best-fitting, simplest models are compared across different samples and used to isolate variables related to individual differences in task performance. Finally, the assumptions underlying binomial mixture models are evaluated.

Basic-level Analyses

Results of the one, two, three, and four component mixed binomial models fitted to each Curriculum x Sex x Item type x Item status x Time grouping are presented in Appendix D. The best-fitting, simplest and therefore preferred model estimates and fit indices for each group are summarized in Tables 5 through 8. Parameter estimates and their standard errors were obtained as described above. Sorting subjects according to each of these five independent variables represents the most basic-level of partitioning available. The resulting sets of items are referred to as basic-level item-sets, so for example, the $n = 93$ Penn State Females' performance on Old items requiring a "Same" response at time 1 (PFOS1(93)) represents a basic-level group. Summary-level analyses, which are composed by collapsing across independent variables, will be considered separately.

A few features of Tables 5 to 8 are notable. First, the minimum VAF was 85.3%, while the average was 92.4%, indicating a high correspondence between the model and

the data. In comparison to measures of explained variance usually seen in regression or analysis of variance settings, a model explaining over 92% of the variance is substantial. Equally remarkable is the stability of the model structure across the different groups: in the majority of cases, two component models provide the best fit. For three of the basic-level groups a one-component model provides satisfactory fit, however these represent the groups with the smallest sample sizes (CFOS1(28), CFNS1(28), and CFOS2(27)) and therefore the lowest power to detect more than one group. Without exception, the four component models were rejected. Although there are a few cases where a three component model does appear to agree best with the data, under closer inspection, the lowest performing component of these three component models seems to pick up only a very few subjects, usually 1 or 2 out of a sample of between 75 and 258 individuals. Recall that the π -estimates reflect the proportion of individuals within a component. When a three component model best fits the data, typically the π value is quite low, as can be seen in Table 5. In fact, for most estimates, the confidence interval of $\hat{\pi}_1$ includes 0. For example, the three component model for Penn State Females' performance on Old-different items at Time 1 has $\hat{\pi}_1 = 0.011$. The interval of $\hat{\pi}_1 \pm 2 \text{SE}(\hat{\pi}_1)$ includes 0. When the outliers that these parameter estimates measure are removed and the models re-fit, the two component estimates invariably fit as well as the actual three component models, and the estimates are practically identical to those of the original two component model. This can also be seen in the fact that the VAF index is fairly high for the two component model, and does not increase substantially with the addition of a third component.

Rarely does VAF substantially increase with the addition of a third component, even though one or both of the χ^2 test statistics is often significantly reduced. Overall, two components seem to model the data well. The large two component χ^2 values are partly indicative of just how much power these tests have here, especially when sample sizes are large. When the three component models fit best according to a χ^2 criterion, often the values are below their expected values, suggesting that the third component is simply picking up a few subjects whose scores are responsible for the high χ^2 values in the 2 component model, and that the two-component model fits almost all of the data quite well. It is often the case that just a few cells of the two component model are responsible for high χ^2 values, while the model captures the general spirit of the data. For example, consider Penn State Females' performance on Old-Different items at Time 1 (PFOD1(93)). A three component model seems to provide a better fit according to VAF and both χ^2 statistics than the two component model. Table 8 shows the individual cell χ^2 contributions for the two component model, indicating that 1 subject is responsible for nearly all of the computed χ^2 value. Eliminating that one subject's score results in a χ^2 value that fails to reject the model at the $\alpha = 0.05$ level. As also shown in Table 9, each of the χ^2 's are unrestricted. Without pooling adjacent cells some of the expected values fall below 5, also potentially increasing the observed χ^2 values. This information supports the high correspondence between the models and the data demonstrated in Figures 4-7. For these reasons, two component models were chosen for all of the basic-level item-sets.

Figures 10 and 11 graphically show the probability of success (θ) estimates and proportion (π) estimates, respectively for each of the 32 basic-level two component models.

In Figure 10, the front row of bars indicates probability of success estimates (θ) for the lower performing group, while the back row of bars shows probability of success estimates (θ) for the higher performing group. While this figure shows that the success estimates fluctuate across each of the 32 item-sets, it does not appear that there are any systematic fluctuations in the success parameters for either the high- or low-level estimates. The fluctuations are more pronounced for the lower performance groups' estimates, as might be expected given that $\hat{\theta}$'s closest to 0.50 have the largest standard errors. Figure 11 presents the proportion estimates for each of the 32 basic-level item-sets. In contrast, these estimates do seem to change more systematically across item-set. For example, within each of the Sex x University samples, the lower performing group has the largest proportion of individuals for new-same items at either time 1 or time 2, and for two groups the two largest "low" proportions are for new-same items regardless of time. Similarly, old-same items appear to have smaller proportions of low-level performers at both time 1 and time 2 for 3 of the four Sex x University samples, while old-same items have smaller proportions of low-level performers at time 2 for the fourth group. In general, it also appears that there were more high-level performers at time 2 than time 1 across all of the task conditions.

Probability of Success Estimates, $\hat{\theta}$. Of preliminary interest is whether the probability of success parameters are significantly different from one another for different

item-sets given that the number of components is the same. It was hypothesized earlier that the same component groups would be found across old and new as well as same and different test item-sets and for both sexes and curriculum (university) populations across time. For example, this hypothesis indicates that all of the θ_1 parameters are the same within sampling error. This is, however, an idealized hypothesis. The probability of success parameters are almost certainly different, but perhaps at least similar to one another.

In order to assess the consistency of the probabilities of success, approximate z -tests were conducted within component groups for each Sex x University sample. Of interest is consistency of the item-type x item status x time $\hat{\theta}$'s (e.g., Old Same items at Time 1, Old Same items at Time 2, etc.) within the 4 independent Sex x University groups (e.g., Penn State Females, Penn State Males, etc.). It would be natural to expect that probabilities of success would be more similar to each other within each group than they would be across groups. Recall that approximate z -tests can be constructed to compare parameters given the parameter estimates and their standard errors (see Equation 10, p.29, and Table 5). With 8 groups there are $\binom{8}{2} = 28$ pairs of estimates within each component. Under the null hypothesis that all of the parameters are equal, 5% of the 28 (or 1.4) z -scores would be expected to be greater than 2 by chance. The best-fitting two component model estimates for each Sex x University were compared against each other within each Sex x University sample (See Appendix E). Of the 224 z -tests conducted for these 4 groups, 101 (or roughly 45%) were significant at the $\alpha = 0.05$ level, indicating that perhaps the probability of success estimates were not uniform within sampling error. It

will be argued later, however, that this does not necessarily indicate the groups are markedly or importantly different from one another. The lack of similarity between probabilities of success as measured by the z-tests appears to be due, in part, to the tests' relative power. From Appendix E it can be seen that the number of significant z-scores is highly related to each Sex x University group's sample size in that Penn State Male's have the largest sample size and the highest number of significant differences, while the Cooper-Union Female's have the least number of significant z-scores and the lowest sample size.

It might also be reasonably expected that the parameter estimates would be similar within time and item type for all of the Sex x University samples. For example, the Old-same items at time 1 should be similar for Penn State Females, Penn State Males, Cooper-Union Females, and Cooper-Union Males (See Appendix F). Consistent with the within Sex x University sample z-tests, 42 of the 96 \underline{z} -tests were significantly different, indicating that they were not alike. Again, given the power involved it is remarkable how similar they are.

To give an overall sense of the probability of success estimates, \underline{z} -tests conducted on the $\binom{32}{2} = 496$ possible lower performance group pairs and the 496 higher performance group pairs were significantly different at a much higher rate than the 5% expected due to chance. For the lower performing group, 232 (or 47%) of the z-scores were significant, while for the higher performing group, 221 (or 45%) were significant. It is clear that all θ -estimates are not the same, as had been hypothesized. It does not appear, however, that there are many systematic differences with respect the independent

variables as seen in Figure 10, and this is the most important point. In general, it appears that the lower component θ -values fluctuate more than expected, but not systematically.

To illustrate this point, imagine that the same two component population model strictly holds for each of the 32 basic-level groups, and that the two types of performers are described by the probability of success parameters θ_1 and θ_2 . Under the assumption that the $\hat{\theta}_r$ ($r = 1, 2$) are approximately normal in distribution and each of the 32 basic-level groups have common θ_r , then the estimates of θ_r should all fall within ± 2 standard errors of θ_r . A histogram of the $\hat{\theta}$'s should therefore show two bell-shaped "humps", with each "hump" centered on θ_r . With sample sizes varying from 27 to 258 and the number of trials either 9 or 15, the computation of the standard error is problematic. One rough method of assessing this is to use the average $\hat{\theta}_r$ to provide estimates of the population values for θ_r , and the average of the estimated standard errors for $\hat{\theta}_r$ to estimate the population standard error. Figure 12 provides a histogram of the 32 $\hat{\theta}_1$ and $\hat{\theta}_2$ values and identifies the contribution of each basic-level group. The curves shown in Figure 12 denote the expected distributions of the $\hat{\theta}$'s under the assumption that the $\hat{\theta}$'s are approximately normally distributed. While the $\hat{\theta}$'s are more spread out than would be expected, they do fall into two clusters, indicating that each of the $\hat{\theta}_r$ do seem to represent similar if not common θ_r values.

Figure 10 suggests that while the $\hat{\theta}_r$ vary more than anticipated, they do not do so in any systematic fashion. Figure 12 allows this hypothesis to be investigated further. If there is any systematic variation in the $\hat{\theta}_r$ caused by the different levels of the five independent variables, then it should be the case that the probability of success estimates associated with those variables will fall above or below the mean $\hat{\theta}_r$ value more than would be expected by chance. For example, if male θ_r values are higher than female θ_r values, then a binomial test should show that the male $\hat{\theta}_r$'s fall above the mean $\hat{\theta}_r$ value significantly more than 50% of the time. Likewise, female $\hat{\theta}_r$ values would be expected to fall below the mean significantly more than 50% of the time. Binomial tests were conducted for each of the five independent variables: Curriculum (University), Sex, Item-type, Item-status, and Time. In all cases, $p > .05$ indicating that there were no systematic influences acting on the $\hat{\theta}$'s.

This evidence suggests it reasonable to view performance from a common model structure, but that estimates fluctuate. As a further test of whether a common probability of success vector (i.e., $\theta_1 = 0.6283$ and $\theta_2 = 0.9336$) underlies performance across all 32 basic-level item-sets, the two component models were refit with a common, fixed probability of success vector (i.e., θ_1 and θ_2 were both fixed) while the proportion parameters (π_1 and π_2) were both free to fluctuate. Because it was hypothesized that performance differences across item-type and time would be seen as differences in

proportion values, these parameters were not fixed across models. The common θ_r estimates were obtained from the overall two component model solution. Results indicate that, while the fit as measured by χ^2 , L^2 , or VAF was reduced in many cases, the models did fit reasonably well overall. In order to assess the acceptability of the model restrictions, the likelihood ratio chi-square and its degrees of freedom were partitioned as described in Chapter I. When the two θ_r estimates are fixed, two degrees of freedom are gained in the restricted model. If the increase in L^2 due to the restrictions of the θ_r is small with respect to the increase in degrees of freedom, then the model restrictions are acceptable. Appendix F provides the parameter estimates and fit statistics for both the restricted and unrestricted two component models for all 32 basic-level groups. For each of the basic-level groups, the partitioned L^2 and degrees of freedom are provided, with asterisks indicating non-significant increases and acceptably well fitting restricted models. Except for the Penn State Males, 20 of the 24 restricted models provide adequate fit, suggesting that the restricted models capture the important features of the data and that a common probability of success vector describes performance. However, only two of the eight Penn State Male basic-level groups showed a comparable fit. It must be remembered, however, that the Penn State Male samples are the largest and the resulting increased power to reject any restricted model is substantial.

The overall fit of the restricted model lends further support to the notion that a common model structure describes performance. *As such, the restricted two component*

model solutions will be used to describe the basic-level item-sets for the remainder of the analyses.

Understanding the probabilities of success, however, only considers half of the model parameters. As hypothesized earlier, individual differences will be most readily apparent in the proportion parameters (i.e., the π 's) which describe "how many" individuals are in each performance group.

Proportion Estimates, $\hat{\pi}$. It was hypothesized earlier that, consistent with previous research findings, observed mean differences would be the result of differences in the proportion parameters. As noted above, new same items appear to be especially difficult across the four Sex x University samples. Tests of significant differences are reported in Appendix G. Of the 59 significantly different independent z's, over half (35) are due to differences between the new-same item estimates and other estimates. On the other hand, only slightly more than one-quarter of the differences (17) are due to the new-different items, suggesting that they are not much more difficult than the old-different items with 14 significant differences. These tests provide further evidence that the proportion estimates vary more consistently across the different samples than the probability of success estimates, and that old items were easier than new items while performance at time 2 exceeded performance at time 1.

Sex, time, item-type, and item status differences all appear to be the result of proportion differences. For example, both males and females show two groups of performers. The high-level males and females perform at similarly high levels, while the low-level males and females perform at similarly low levels. Sex differences are the result of the fact that there are relatively more males who belong to the better performing

component and relatively more females whose scores were drawn from the lower performance group. These between group proportion differences are displayed in Figures 13-16 which show restricted model π_2 -estimates and standard-error bars for the four Curriculum by Sex groups over time on Old-different, New-different, Old-same, and New-same items respectively.

Figure 13 shows the proportion of subjects in the "high" ability component on Old-different item over time for each of the four Curriculum by Sex samples. Because the low-ability component's proportion estimates ($\hat{\pi}_1$) are simply $1 - \hat{\pi}_2$, only the $\hat{\pi}_2$'s are shown. The two male samples' performance is indistinguishable. Both improve over time as seen in the increased proportion of subjects in the "high" ability component by time 2. In contrast, there are fewer Penn State females in the "high" ability component at both time 1 and time 2, even though the rate of improvement is consistent with male improvement. In other words, approximately 10% of the Male and Penn State Female samples shifted from the "low" ability to the "high" ability group on Old-different items over time. The Cooper-Union females, on the other hand, performed at the same level as their male counterparts at time 1, yet failed to show any improvement in performance on the Old-different Items over time.

Figure 14, which shows the proportion of subjects in the "high" ability component on New-different items at Times 1 and 2, is somewhat similar. Unlike the Cooper-Union females, a significantly higher proportion of Penn State females are found in the "high" level group at time 2 than time 1 ($z = 3.1, p < .05$), even though performance levels for Penn State and Cooper-Union females are non-significantly different at both time points. Similarly, the Penn State males showed a significant level of improvement not seen in the

Cooper-Union males ($z = 2.3$, $p < .05$). Additionally, both female groups have a lower proportion of members in the "high" ability component than the two male groups at both times.

Consistent across both types of Different items, females have a significantly lower proportion of members in the "high" ability component than males at both Times 1 and 2, while there are no differences between Penn State and Cooper-Union students. Further, the Cooper-Union females show the least amount of improvement, while the Penn State subjects, both male and female, improved on both types of Different items.

As shown in Figure 15, none of the four Curriculum by Sex groups showed improvement on the Old-same items over time. However, a higher proportion of males than females were found in the higher performing group at both times (at Time 1 ($z = 2.5$, $p < .05$) and Time 2 ($z = 2.5$, $p < .05$).

While a higher proportion of males belong to the "high" ability component at both Times 1 and 2 (as shown in Figure 16), only Penn State males and females showed a significant improvement over time ($z = 3.8$, 3.3 , p 's $< .05$, respectively). The effects of Curriculum and Sex on performance over time are considered further in the bivariate setting.

If samples are pooled and analyzed in terms of Curriculum and Sex differences something akin to "main-effects" can be investigated. Two methods of combining subjects' responses are used here. The term "pooling" is used here to indicate the concatenation of scores. For example, if the first four Penn State females scored 9, 6, 3, and 2 on the 9 Old-same items at Time 1, and 14, 12, 3, and 4 on the 15 Old-different items at Time 1, eight scores would be analyzed as though they represented independent

observations, resulting in scores 9, 6, 3, 2, 14, 12, 3, and 4. Scores can also be summed within subjects, increasing the number of trials. In this example the four Penn State females would have scores on Old items 23, 18, 6, and 6. Unless otherwise noted, combinations of basic-level item-sets are pooled using the first method. The summary-level items represent the only exception. Because items are not generally separated according to whether they require "same" or "different" responses, the summary-level item-sets represent summed scores over the item-status variable.

The estimates for the mixing proportions are not the same across Figures 13-16. It appears that some items are more difficult than others based upon the proportion of subjects who belong to the "high" level group. To test this hypothesis, Figure 17 presents $\hat{\pi}$'s for all four of the restricted two component Item-type by Item-status sets (e.g, Old-same items at Time 1, Old-different items at Time 1, etc.). As shown in Figure 17, the proportion of subjects in the "high" component drops at both Time 1 and Time 2 across these problem sets, indicating that Old-same items are the easiest, followed by Old-different, New-different, and New-same items. Z-tests indicate that $\pi_2^{OS(454)} >$

$$\pi_2^{OD(454)} > \pi_2^{ND(454)} > \pi_2^{NS(454)}$$

for both Time 1 and Time 2 samples (with one exception: New-same problems at Time 2 were non-significantly more difficult than New-different problems at Time 2). Over both sample times each item set is significantly more difficult than the one preceding indicating the order of problem difficulty.

As with the same items, a significantly lower proportion of females belong to the higher performing group for both Different items at both Times 1 and 2, while there is no effect for Curriculum. Overall, it appears as though the Penn State subjects showed more

improvement than the Cooper-Union subjects, and that females improved more than males. To test this hypothesis, performance across the combined set of all items (Old-same, Old-different, New-same, and New-different) were modeled.

Tests of other hypotheses can be designed as well. For example, the differential effects of stimulus complexity on males and females can be evaluated. This hypotheses can be evaluated for each of the Item-types (Old and New) by looking at differences in proportion estimates. Figure 18 shows restricted model "high" component proportion estimates for all females on Old and New items at Times 1 and 2. The difference between the estimates at Time 1 is significantly greater than the difference at Time 2 ($z = 2.64$, $p < .05$), even though a significantly larger proportion of individuals belongs to the higher performing component for Old items than New items at both Times. The picture is similar for males. As shown in Figure 19, the difference between the proportion of subjects in the "high" component on Old and New items at Time1 is significantly larger than the difference at Time 2 ($z = 2.72$, $p < .05$). This suggests that for both males and females, the effects of stimulus complexity are reduced with experience. Further, performance increased more for the more difficult items. And while the Cooper-Union students did not receive the same curriculum the Penn State students did, both groups were involved in Engineering.

As a direct test of the effects of Curriculum, Figure 20 shows the proportion estimates of "high" level performers over all items for Penn State and Cooper-Union subjects at Times 1 and 2. Although probably partly due to the power derived from larger number of Penn State subjects, the figure shows a significant increase in performance for Penn State subjects over time ($z = 5.44$, $p < .05$) and a non-significant increase in

performance for the Cooper-Union subjects over time ($z = 1.7$, $p = \text{n.s.}$). A significantly greater proportion of Penn State subjects than Cooper-Union subjects shifted from the "low" level component to the "high" level component over time.

Sex differences over time are presented in Figure 21 which shows the proportion estimates of males and females over time. It was hypothesized that female improvement over time would be greater than male improvement for reasons cited by Sherman (1967), contrary to the findings of other researchers (e.g., Baenninger & Newcombe, 1989). This was not, however, borne out by the data. Both sexes improved similarly significantly over time, and a greater proportion of males belonged to the "high" level component at both times. Figures 20 and 21 are, in effect, Curriculum x Time and Sex x Time interaction terms, respectively. Figure 22 shows what might be viewed as the three way interaction between Curriculum, Sex, and Time in the following manner. The "high" ability component proportion estimates for each of the four Curriculum by Sex groups were examined at Times 1 and 2 for all items. The difference between the proportion estimates at times 1 and 2 were then computed. These differences were then graphed in Figure 22. So, for example, the point indicating 0.082 in Figure 22 represents the proportion of Penn State Males in the "high" level component at Time 2 ($\hat{\pi}_2 = 0.742$) minus the proportion of Penn State Males in the "high" level component at Time 1 ($\hat{\pi}_1 = 0.660$). In essence this figure represents the relative increase in the proportion of subjects in the "high" level component over time. While Figure 21 shows that there is no overall improvement difference between males and females, Figure 22 demonstrates that the Penn State Females show the largest increase, followed by Penn State Males. Cooper-Union subjects, performed at a uniformly high level over time, showing less improvement.

The results presented in Figures 20 and 21 are interactions subsumed by the interaction in Figure 22. In sum, the females who received the solid-modeling curriculum showed the greatest increase in membership in the "high" level component, followed by the males who received such training. Both males and females without the solid-modeling curriculum showed no such increases in performance.

General Model Findings

In general, it appears that probability of success estimates vary in a less systematic fashion than the proportion estimates, suggesting that their differences could be more the result of sampling error than any real population differences. It was hypothesized that a mixture model would show mental rotation performance change to be the result of changes in component membership rather than changes in probabilities of success, and most between group differences appear to be the result of consistent proportion differences. In other words, there does indeed seem to be two groups or types of performers: those who perform slightly, but significantly above chance ($\hat{\theta}_1 = 0.6283$; where chance is $\theta_1 = 1/2$), and those who perform at near ceiling levels ($\hat{\theta}_2 = 0.9336$).

Summary-level Analyses

While the models listed above separate items based on whether they were the same as or different from the target item, Vandenberg and Kuse (1987) type mental rotation tasks items are rarely separated into same and different problems for analysis. For this reason, the tasks were modeled without regard to same-different item status as well so that model results could be compared to previous research. In addition, combinations of variables were modeled as well. All task combinations were modeled

with 1, 2, 3, and 4 component binomial mixtures. Fit indices, parameter estimates and their standard errors are presented in Appendix H, with the two component models outlined for clarity. Often it was difficult to decide on whether two or three component models should be used to fit the data because fit indices were mixed. Three component models were accepted, however. While the justification for doing so will become more apparent in Chapter V, consider the following rationale. The basic-level group PMOS1(258) has nine items, while the basic-level group PMOD1(258) has fifteen items. Many subjects perform at the lower-level on both Old-same and Old-different items. As such, their probabilities of success on the nine Old-same items is 0.6283. Similarly their probability of success is 0.6283 on the 15 Old-different items. Their overall probability of success on the 24 Old items remains 0.6283. Those subjects who perform at the higher level on both Old-same and Old-different items is consistently 0.9336 on all items. Those subjects who perform at the "high" level on the "same" items and at the "low" level on the "different" items have an overall probability of success on the set of all Old items equal to the weighted average of the two high- and low-level probability of success values. In effect, an intermediate probability of success between 0.743 and 0.819 (the weighted averages of the two restricted model $\hat{\theta}_r$'s). The result is a three component model with the same "low" and "high" probabilities of success as the two component model and the addition of a third, "intermediate" component which captures inconsistent performers. In addition to the unrestricted models, Appendix I shows the three component model solutions with the "low" and "high" values of θ set equal to the two component restricted θ 's, and partitioned likelihood chi-square statistics to assess fit.

These restricted three component models can be used to test hypotheses about the effects of each of the variables on performance. For example, to test the hypothesis that new items are more difficult than old items for each component group, the dataset is partitioned by item-type only. In this fashion, the high and low performance groups are compared across item-type (i.e., $\hat{\pi}_1^{(O)}$ is compared to $\hat{\pi}_1^{(N)}$, $\hat{\pi}_2^{(O)}$ is compared to $\hat{\pi}_2^{(N)}$, $\hat{\pi}_3^{(O)}$ is compared to $\hat{\pi}_3^{(N)}$, where the letter in parenthesis identifies the group - in this case Old and New items). Figures 23-27 portray three component π -estimates for each of the five variables of interest, item-type, curriculum, sex, item-status, and time respectively. For each of the five variables, observations were pooled over the remaining four variables, and modeled by two, three, and four component mixtures. *In all cases, restricted three component models were accepted as the best-fitting, simplest models.*

Figure 23 shows a significantly smaller proportion of subjects in the highest performance group on New items as compared to Old items ($z = 13.7$, $p < .05$). And while the intermediate group contains more individuals on the New items ($z = 4.6$, $p < .05$), the lowest performing group is significantly larger for the New items than Old ($z = 10.7$, $p < .05$). Differences in task complexity are accounted for by large differences in component membership. Figure 24 compares parameter estimates for Penn State and Cooper-Union subjects, indicating little overall difference between the two curriculum groups (z 's < 2.0 , p 's = n.s.). While this suggests that there is little overall difference between the two engineering curricula, these estimates are pooled across Time. As seen in Figures 20 and 22, Penn State subjects showed greater performance gains over Time

than the Cooper-Union subjects. Figure 25 shows the overall effect of sex. In this figure, the proportion estimates show the relatively higher proportion of males the highest performance group ($z = 10.6, p < .05$) and a higher proportion of females in the lowest performance group ($z = 8.6, p < .05$). Overall, sex differences, like differences in task complexity, appear to be well captured by differences in group membership.

It was hypothesized that items that required a "same" response would not be different than those requiring a "different" response, even though there is some evidence in the literature to suggest that each item required different solution times. As seen in Figure 26, when items are pooled across Curriculum, Sex, Item-type, and Time, there are no significant proportion differences between Same and Different items (z 's < 2.0 , p 's = n.s.). Note, however, that while there are no differences between Same and Different items there is a significant interaction between Item-status and Item-type. Figure 28 shows the "high" component proportion estimates (π_2) from Figure 17 arranged to show how Item-status and Item-type interact. By comparing the difference between Same and Different "high" component proportion estimates ($\hat{\pi}_2$'s) for Old and New items

$(\pi_2^{(OS)} - \pi_2^{(OD)} = \pi_2^{(NS)} - \pi_2^{(ND)})$ the interaction can be tested. "Same" items more than "different" items become more difficult as complexity increases (at Time 1, $z = 6.4, p < .05$; at Time 2, $z = 3.4, p < .05$). In other words, complexity affects "same" and "different" items differentially.

The effects of time are shown in Figure 27. In effect, the number of subjects performing at high levels increases over time ($z = 6.1, p < .05$), and a reduction is seen in the proportion of subjects in the "intermediate" group ($z = 5.7, p < .05$). When

performance differences exist, they appear to be well described by changes in component group membership. In effect, the number of subjects performing in the "high" level component increases over time. In sum, old items are easier than the newer more complex items, males perform at higher levels than females, and performance increases over time.

A more complete analysis of the relationships between variables will be considered in the bivariate context in the next chapter. For example, questions concerning whether old-same items are more difficult than old-different or new-same items are most easily answered from a bivariate perspective. In this fashion, subjects responses over item type, item status, and time can be assessed in comparison with each other and the two between subjects' variables.

Partitioning Based on Component Membership

Evidence presented to this point indicates that a two-component mixed binomial distribution describes mental rotation performance when Same and Different items are examined separately. One of the advantages of the current modeling perspective is that it allows for the partitioning of subjects based on the posterior probability (defined in Equation 12, p. 32) that a score came from one of the two components. Subjects can be jointly classified as belonging to either the "high" or "low" level groups on both Old and New items or for Time 1 and Time 2. This bivariate classification will be considered further in the next chapter. By classifying subjects based on the component from which their score most likely came, it becomes possible to test a range of hypotheses concerning individual differences in performance by analyzing what makes subjects in one component different from subjects in another component beyond the number of items correct. If the components are psychologically valid, then subjects classified as high-level

performers might be different than those classified as low-level performers on other tasks that relate to spatial ability. It is also possible to evaluate the performance of each kind of performer with respect to the effects of rotation angle on performance. While it was hypothesized earlier that each item's rotation angle would not affect performance, it may be the case that high and low-level performers respond differentially to different angles of rotation. One might envision that "high" level performers' accuracy is not dependent on the angle of rotation because their ability is sufficiently developed so that all rotations use the same set of well-mastered procedures. Individuals in the lower group, on the other hand, may find some angles of rotation more difficult than others because their procedures are more object dependent. Issues concerning factors that might covary with component membership are considered next.

Covariant Analysis

Achievement Tests. It was hypothesized that when different subjects groups show similar model structures, there are underlying similarities between the individuals in those groups. High school and college cumulative GPA's, SAT scores, and college placement scores (one for English, one for Chemistry, and four for Mathematics) were obtained for Penn State subjects. It should be the case that if similar parameters define performance for both males and females on a certain item-set (e.g. Old items at Time 1), then performance should be similar on these achievement measures as well. In other words, if the best performing groups for males and females on the Old items at Time 1 have similar θ -values, then there should be no difference between their SAT math scores. Conversely, groups differentiated by their θ -parameters should have different outcomes on these measures as well. To answer this question, t -tests were performed for each of the above

measures, with either sex or component group membership independent variables. Table 10 provides sample sizes, means, standard errors, t -statistics, and p -values for males and females without regard to performance group. Interestingly, there was only one sex difference for all of the comparisons. Females had a higher high-school GPA than males. It might have been expected that when conditioned on sex, there would be differences on many of the 10 variables because there is a greater relative proportion of males in the higher performing group than females. It must also be remembered that the samples here are not representative of the general population. Engineering students, both male and female, are selected for admission based, in part, on these criteria. As a result, sex differences are less likely to be found here than they might be with a different sample.

The t -test, however, is not an omnibus test of distributional differences, but rather a test of mean differences in the normal setting. A potentially stronger claim about the differences between males and females may be possible. For each t -test conducted, a Kolmogorov-Smirnov two-sample test was conducted to identify distributional differences on the achievement tasks (Roscoe, 1975). Under the null hypothesis that there are no differences in the distributions associated with the achievement scores, the statistic

$$K_D \leq 1.36 [(n_1 + n_2) / n_1 n_2]^{1/2}, \quad (14)$$

at the $\alpha = 0.05$ level, where n_1 and n_2 represent cell frequencies (Roscoe, 1975). K_D is simply the maximum relative frequency difference for all levels of the observed distributions. In the analyses which follow, only the presence or absence of significant differences are reported for the Kolmogorov-Smirnov tests. While not as powerful as a t -test at detecting mean differences when normality holds, this test is designed to measure any distributional differences. In no case, however did the test detect a significant

difference between males and females at the $\alpha = 0.05$ level, suggesting that there were no distributional differences between males and females on any of the 10 achievement tasks.

Sex-differences on each of these variables can also be evaluated within performance group. Tables 11-14 provide sample sizes, means, standard errors, t -statistics, and p -values for males and females within each performance group on a representative sub-sample of basic-level item-sets. Performance groups were defined by posterior probabilities such that each subject's score was placed in the component from which it was most likely drawn. Each of the four tables defines "high" and "low" performance based on a subset of items. Because there were many models that could have been used to assign individuals to either the "low" or "high" performance groups, four of the two component models were used so that the relationship between external variables and the conceptualization of mental rotation performance forwarded here was not dependent on any one model. Thus, "high" and "low" performers are compared without regard to sex. Results are similar across all four partitions, indicating that there is probably some validity to the model conceptualization proposed here and that subjects are reasonably partitioned. The remaining possible partitions were similar in all respects to those presented here, so these effects are not unique to any particular partition. Overall, there were almost no sex differences within component. The only variable which showed a consistent difference was High-school GPA, consistent with the overall sex difference on this variable. It is perhaps not surprising that the choice of models is not especially crucial because the same data are being partitioned only slightly differently by the posterior probabilities of the different models. Additionally, Kolmogorov-Smirnov two-sample tests were conducted, however, only 1 of the 80 comparisons reached significance.

As shown in Tables 11-14, sample sizes for the achievement test comparisons ranged 3 to 49 and averaged approximately 25. Some of the smaller samples doubtlessly show no differences as a result of insufficient power rather than a lack of actual differences. In general, however, it appears that within component, there are few sex differences on achievement tasks.

To provide another perspective, the three component model partitions were also examined. Two are presented in Tables 15 and 16 showing performance based on Old items at Time 1 and New items at Time 2, but all four are similar. Consistent with the two component models, none of the t-tests and only 1 of the 60 Kolmogorov-Smirnov two-sample tests showed a sex difference within component.

However, when the components are compared *without* regard to sex, a very different pattern emerges. Tables 17-20 provide sample sizes, means, standard errors, *t*-statistics, and *p*-values for four of the eight possible partitions defined by posterior probabilities from the two component models, without regard to sex. Chemistry, English, and GPA seem to be generally unrelated to component membership, but SAT and Math achievement tests are strongly related to group membership. In two of the four cases, the "high" performers outscored their "low" performing counterparts as measured by *t*-tests. Further, the Kolmogorov-Smirnov two-sample test indicates distributional differences on many of these variables. While none of the variables differed, according to this test, across all four partitions, the math achievement tests seem to be best at differentiating between the two performance groups. The fact that not all of the partitions show between component differences will be considered further in the next chapter. In sum, while there seems to be no difference between either "high" performing males and females or "low"

performing males and females, the "high" and "low" performance groups show strong differences, at least with regards to math and science achievement test scores.

Rotation Angle. It was assumed when the binomial mixture model was first outlined that rotation angle would have little measurable effect on accuracy scores, however this is an empirically testable question. If the amount of rotation required to align "same" items to their target for each item has no effect on accuracy, then the overall proportion of correct responses for an item should not vary as a function of rotation angle. Each of the "same" items were rotated on all three (e.g. X, Y, and Z) axes. For each object, the X, Y, and Z axes were defined by the orthogonal faces of the cubes that make up the objects. The decision as to which axis to assign the label X, Y, and Z was made on the target drawing by the graphics program on which the objects were created. In general, the target objects' Y axis corresponded to the vertical axes, while the X axis was primarily a horizontal axis and the Z axis an axis in depth, as shown in Figure 29. As a result, the Z-axis primarily describes rotations in the picture plane, while the X- and Y-axes describe vertical and horizontal rotations in depth, respectively (at least until after the first rotation when the axes are shifted). In measuring the amount of rotation required to align "same" objects, each of the three axes were fixed with respect to the orthogonal faces of the objects, and rotated as necessary to achieve congruence with the target. Unfortunately, when rotating an object on three axes, the rotations are non-commutative. That is, a 45° rotation along the X-axis followed by a 45° rotation along the Z-axis is not the same as a 45° rotation along the Z-axis followed by a 45° rotation along the X-axis, as demonstrated by a randomly oriented object used in this study (Figure 30). This leads to a difficulty in measuring the angle of rotation for each item with respect the three axes because there are

many possible rotations within each frame of axes, and many frames of axes (e.g. axes that remain stationary while the object is rotated vs. axes that remain fixed with respect to the faces of the object during rotation). This has presumably been the impetus for most researchers to vary the rotation items along only one axis at a time (e.g., Shepard & Metzler, 1971). In real-life situations, objects may require rotation about more than one axis for proper orientation. The simple act of fitting a key into a lock, for example, requires that the key be rotated on at least two of its three orthogonal axes. For each of the 24 "same" items, one of the many possible rotations was arbitrarily chosen and then measured about the three axes. In each case, rotations were made in either a positive (clockwise) or negative (counter-clockwise) direction, and never exceeded 180° (a 190° rotation on an axis is equivalent to a -170° rotation on the same axis, and the -170° rotation was used under the assumption that subjects would choose to rotate an object -170° rather than 190°). For analysis, the absolute angle of rotation was used, assuming that it is equally difficult (or easy) to rotate an object 170° as it is to rotate it -170° and that labeling one rotation as negative and the other as positive is somewhat arbitrary. Using this framework, accuracy can be assessed relative to the absolute angle of deviation for each of the three axes separately.

In addition, the sum of the angular deviations about the three axes of rotation should also be considered, especially given the non-commutative nature of the rotations. By adding the three absolute angles of rotation, one measure of the total amount of rotation necessary to align the objects to the targets can be assessed. In some sense this captures part of the spirit of rotation, in that objects which require many degrees of rotation will do so regardless of the order of axes used in rotation, while those that only

deviate slightly will require minimal rotation regardless of the axes, even though there is a less than perfect correspondence between the sum of absolute rotation angles under different rotation orders. As seen in Panels B and C of Figures 30, both objects are rotated a sum of 90° , and both are approximately equally different from the initial orientation shown in Panel A of Figure 30.

As Carpenter and Just (1985) note, in addition to separate rotations about a set of orthogonal axes, objects rotated in three space also have a unique, object-defined axis that defines the minimum amount of rotation necessary to align the objects. Carpenter and Just (1985) found that this rotation axis is preferred by high ability subjects. For each of the 24 "same" objects, the angle of rotation about this unique axis was measured according to Funt (1983). In addition to accuracy assessments relative to the three orthogonal axes and the sum of the deviations about these axes, accuracy was assessed relative to the unique axis as well. It might be expected, for example, that the accuracy rates of "high" performing subjects as classified by posterior probabilities, would have no relationship to rotation angle except perhaps the unique axis angle. "Low" performing subjects, might be expected to have some dependence on rotation angles about the individual axes or the summed angles of rotation if they use a rotation strategy to solve the mental rotation items. If, on the other hand, the "low" performing subjects have scores completely unrelated to rotation angle while the "high" performing subjects have scores related to rotation angle, evidence which suggests that the "low" performers are using a non-rotational strategy will have been found. To test these hypotheses, posterior probabilities from the two component model for "same" vs. "different" response items was used to partition subjects into the "high" and "low" performance groups. Once subjects

were classified according to performance group, the proportion of subjects responding correctly to each item was plotted against the summed rotation angle, the rotation angle of each of the three orthogonal axes separately, and the object-defined minimum axis (Figures 31-35). Each of the plotted samples was fit by linear, quadratic, cubic, and quartic least-squared regression lines. Regression equations were chosen based on the lowest possible number of terms which explained a significant amount of variance (i.e., when a linear term fit, the quadratic, cubic, and quartic terms were rejected). Often, however, none of the regression equations was significant. Regression lines which predict a significant amount of variance have their r^2 's outlined by ellipses. In addition to the least squares regression line, 95% confidence bands are also displayed for each regression line (Snedecor & Cochran, 1980).

Obviously, the least squares regression lines for the "high" performers will be above the lines for the "low" performers because the groups have been partitioned according to performance levels. It remains to be determined, however, whether the slopes of those lines will be different and whether the form of the lines (e.g. linear, quadratic, etc.) will be different. With the exception of the summed rotation angle (Figure 31) where the proportion of correct responses decreased slightly with summed rotation angle for "high" level performers, the proportion of correct responses is unrelated to rotation angle for both "high" and "low" level performers.

Earlier model results indicated that within-sex differences are much more dramatic than between-sex differences (i.e., there is little difference between males and females within component). To examine this result more fully, regression lines and confidence bands were generated for males and females separately within each of the two

components. Figures 36-40 show the proportion of correct responses plotted against summed rotation angle, the object-defined rotation angle, and the three orthogonal rotation angles for male and female "high" and "low" performers. As seen in these figures, males and females have non-significantly different regression lines for all of the rotation angle measures when compared within component. The only significant regression line belongs to females in the "high" level component when proportion of correct responses is compared to summed rotation angle, and it is still non-significantly different from the male regression model.

Summary. Results from the "high" and "low" performers show that, in general, "low" subjects' accuracy is much more variable than had been anticipated; the data are certainly not homoscedastic. Overall, the analysis of rotation angle with respect to accuracy indicates that rotation angle has little effect on performance, but a linear dependence comes closest to describing the relationship between accuracy and rotation angle, at least for the "high" level performers. There was no relationship, however, between accuracy and the minimum angle of rotation about a unique axis as had been anticipated.

Assumptions of the Binomial Mixture

The binomial model is based on two assumptions: that the probability of success for each trial is the same and that each trial is stochastically independent from all other trials. The plausibility of these assumptions can be gauged by measuring the proportion of subjects responding correctly on each item and the inter-item correlations, both within component. If both assumptions hold, then the proportion of correct responses will be

constant across all items and the inter-item correlations will be zero within sampling error within each component. The constant θ assumption will be tested first, followed by a test of trial to trial independence.

Constant θ . Results to this point suggest that there are two different types of performers on the mental rotation task when Same and Different items are considered separately, but that the probability of success estimates for the unrestricted two component models vary across Curriculum, Sex, and Time to a greater than anticipated extent. One possible reason for this phenomena is a failure of the constant θ assumption. In other words, some trials may be more easily solved than others.

As with the analysis of rotation angle, subjects were partitioned based on posterior probabilities into high- and low-level performers for this analysis. In addition, items were segregated based on item-type (Old, New), item-status (Same, Different) and Time. For each of these samples, the proportion of subjects who responded correctly was measured for each trial. Within each component the expected proportion correct is \hat{p} , the binomial parameter, where $[\hat{p}(1 - \hat{p})/n]^{1/2}$ estimates the standard error of \hat{p} . For each component, 95% of the observations should fall within ± 2 standard errors of \hat{p} , and the average \hat{p} will be used to find an estimate of the standard error. Figures 41-45 provide histograms of the item success rates for a two component and a three component example. These figures were chosen because they show the mean proportion of subjects responding correctly in each component for the cases that most egregiously violate the constant probability assumption. Figures 42, 44 and 45 which represent the performance of subjects from the intermediate and highest level groups are not at all inconsistent with

spirit of the constant probability assumption in that most of the proportions are clustered about a single mean value. In contrast, Figures 41 and 43 show proportions that are bimodal or at least not clustered together. In fact, very few of the proportions fall within a 95% CI of the mean of the proportions. Interestingly, the only variable that seems to have any association with these clusters is summed rotation angle. By partitioning proportions from Figure 41 at the mean, the easier items (\bar{X}_e) have a marginally higher summed rotation angle than the harder items (\bar{X}_h) ($\bar{X}_e = 168.5$ vs $\bar{X}_h = 141.9^\circ$, $t = 1.85$, $p < .11$), suggesting that the poorest performers' accuracy was affected by rotation angle for some, but not all items.

The majority of the data, however, are more similar to Figures 42, 44, and 45 than Figures 41 and 43, indicating that the constant probability of success assumption is likely to be valid enough to use a mixture of binomials model. In many cases 95% of the data were contained within ± 2 standard errors though less often for the "low" performers, as can be seen in the figures. In general, there appears to be much more variability in accuracy than expected. But again, at issue is not whether the assumptions of the model have been violated, but whether they have been violated to the extent that the model is no longer useful. Based on the overall model findings above it would appear that it has not. The "low" component seems to describe performers with an overall low accuracy rate. The "high" component, on the other hand, seems to reflect the fact that some individuals perform at extraordinarily high levels regardless of the complexity of the rotation object.

Independence Across Trials. Beyond the intuition that it is difficult to see how getting an item right or wrong (especially without feedback) could influence how a subject responds to the next item, assessing independence is a difficult issue. One might expect

that some subjects use a strategy which involves solving the easiest item (in a row of items, see Appendix A), then evaluating the remaining items against both the target and the first-solved item. In this way, solutions to the remaining three items are dependent on whether the first item was solved correctly, as the third item is dependent on whether the first two were solved correctly, and so forth. This plausibility makes a check of the independence assumption more important. However, there is evidence to indicate that moderate departures from independence are not especially damaging to the model conclusions (Thomas & Lohaus, 1993). In order to test this assumption, each item, scored as either correct (1) or incorrect (0) was correlated with all other items within Item type, Item status, Time, and component for the two component basic-level models and the three component summary-level models. In other words, success or failure on the first rotation item was correlated with success or failure on the second, third, etc. items. Frequency histograms of these correlation coefficients should be centered about 0 if the independence assumption holds. In addition, Huber's (1977) distribution free result provides a means of estimating a standard error of r . As with the constant θ assumption, we would expect 95% of the correlations to be found within ± 2 standard errors.

Figures 46-50 provide examples of correlation coefficient histograms for the two components on Old-Different items at time 2 and the three components for Old items at time 2, which most strongly appear to violate the independence assumption. As can be seen in the figures, the highest and intermediate components conform to an independence expectation and approximately 95% of the correlations are within two standard errors of 0 (Figures 47, 49 and 50). The two low components, on the other hand, contain between 65 and 71% of the correlations in these examples, suggesting that only the "low"

performance group's inter-item correlations present a departure from a mean correlation of 0 (Figures 46 and 48). As hypothesized, it appears that the independence assumption has been violated, but only to a moderate degree and to the greatest extent for the "low" performers. Certainly the lack of independence presented here is within a tolerable range. As a result, it does not appear as though the model should be rejected based on violations of either the constant probability or independence assumptions, because while neither assumption strictly holds, neither appears to call into question the structure of the data to any serious degree.

CHAPTER V

Bivariate Binomial Mixture Analyses

Results to this point suggest that at least the major features of the data are captured by a mixture of binomials distribution. In addition to performance on the individual tasks, subjects' joint performance across task and over time are of interest. Using the joint model of performance outlined above (see Equation 13, p. 36) the effects of curriculum differences and how item difficulty affects performance can be assessed. This chapter examines the relationship between variables (e.g., Curriculum and Sex) across the different levels of the within-subject variables (Item type, Item status, and Time). Usually within subject variables are considered from a correlational perspective, where data are assumed to be roughly bivariate normal. In this case, the joint distributions are considered from a mixture perspective. Joint frequency histograms show bivariate performance for three representative variable pairs in the top panels of Figures 51-53: One which varies over Item-status, one which varies over Item-type, and one which varies over Time. The top panel in each figure shows the observed joint frequencies, while the bottom panel shows the expected frequencies under the model. In both panels, shaded regions indicate the joint component groups, and marginal histograms have also been included. Two aspects of these figures are noteworthy. First, there is a strong correspondence between observed and expected frequencies under the joint model. The marginal and joint frequencies both appear to correspond well with expected values, and each of the four observed joint probability masses seem to agree with expected values in shape and overall appearance. Altogether, the figures indicate

that the model successfully captures the spirit of the data. Second, the data are at variance with the kind of structure usually assumed in a correlational setting. The observed data are clearly inconsistent with a bivariate normal perspective which would predict symmetric marginal distributions and a single "mound" of data somewhere near the center of the figure's floor.

Before the bivariate results are presented, an unusual feature of the data requires consideration. As noted in Table 1, the sample sizes at Time 1 and Time 2 are much larger than the number of subjects who participated at both times. This was due to subjects' reluctance to use their Social Security Numbers as identifiers. For example, of the 28 Cooper-Union females who participated at Time 1, all but 1 participated at Time 2. However, only 12 provided a subject identifier which allowed their performance to be tracked over time. As such, the univariate analyses which evaluate performance at both time points are based on somewhat different samples than the bivariate analyses. As such, the results often show minor discrepancies. The univariate results have much greater power than the bivariate results, so it is difficult to evaluate which sample provides the "best" view of performance change. Doubtlessly results that are consistent in both the univariate and bivariate settings are the most reliable.

A first look at how performance changes over time, by item-type, and by item-status can be observed by using the categorizations provided by the univariate posterior probabilities. Figures 54-59 provide an graphic view of performance change, while the bivariate model framework is implemented below. Figures 54-59 show the transition of subjects from one basic-level item-set to another for 12 of the item-set pairs based on the posterior probabilities from Chapter IV. While there are 28 possible pairings of the 8 basic-level groups, only those that were psychologically interesting (i.e., those constant

across two of the three variables which describe each item-set) are presented. For example, Old-same items at Time 1 are compared to Old-different items at Time 1 because they differ only on the item-status dimension, but not with New-different items at time 2 because this kind of comparison seems less interesting.

In each of the figures, sample sizes for each component are provided in boxes which are proportionate to sample size. Arrows show changes from one item-set to the other. For example, the left panel of Figure 54 shows that 51 subjects were classified as most likely belonging to the "low" component on Old-same items at Time 1. Of those 51, 38 most probably remained in the "low" component for Old-different items at Time 1, while 13 most probably came from the "high" component on Old-different items at Time 1. The figures which show item-sets within time provide one measure of the relative difficulty of each item-set. Consistent with the univariate results, both panels in Figure 54 show that Old-different items are more difficult than Old-same items because a relatively small number of subjects classified as "Low" performers on the first item-set are classified as "High" performers on the second item-set at both Time 1 and Time 2. Similarly, a relatively high proportion of subjects change from the "High" level on Same items to the "Low" level on different items at both times. In contrast, Figure 57 shows that Old-different items were not easier than New-different items because there is much more consistency from set to set than with Old-same and Old-different items.

The figures which show item-sets across time provide a measure of the amount of learning or performance increase over time. For example, the panels on Figure 56 show that many of the subjects classified as "Low" at Time 1 changed component by Time 2. This increase in performance was offset somewhat by a decrease in performance by some

of the individuals in the "High" group. The corresponding figure for New items are shown in Figure 59. In this case, performance appears to be much more stable.

Parameter Estimation and Interpretation

The probability of success and proportion parameters for the joint mixture model are the same as those for the univariate models, and do not require further estimation. Though they were taken from the complete sample, they should provide the best estimates. The joint model has additional parameters not found in the univariate models which reveal important information: π_{ab} and τ_{ab} (where a , ranging from 1 to k , refers to a component on X , and b , ranging from 1 to k , refers to a component on Y). The joint proportion π_{ab} describes the proportion of subjects who belong to component a of variable X and component b of variable Y , while the transition parameter τ_{ab} describes the proportion of subjects from component b of variable Y who were in component a on variable X . The transition parameters represent shifts in subjects' performance from a given level on the first item set to a level on the second item set.

Basic-Level Analyses

For all of the panels in Tables 21 - 25, the X variable is on the vertical axis and the Y variable is on the horizontal. For example, from Panel A of Table 21, Old-same items at Time 1 is the X variable and Old-same items at Time 2 is the Y variable. The choice of which variable is X and which is Y is in many ways arbitrary, except when comparisons of the same item-set are being made over time. In this case, the transition from Time 1 performance to Time 2 performance is a natural one. Note that the joint proportion estimates ($\hat{\pi}_{ab}$) sum to 1, while the transition estimates ($\hat{\tau}_{ab}$) sum to 1

across each row (variable X).

For Tables 21 - 25, 2 x 2 contingency tables are provided for the 12 basic-level item pairs (labeled panels A-L). The panels were constructed based on the posterior probabilities of component membership from the univariate models. The maximum likelihood estimates for the joint π -weights are presented at the top row of each cell with their standard errors in parentheses immediately below them. In addition, the observed proportion of subjects in each of the joint components (denoted \hat{p}_{ab}) using the univariate posterior probabilities are presented to the right of the $\hat{\pi}_{ab}$'s. For example, Panel A of Table 21 indicates $\hat{\pi}_{11} = 0.059$, $\hat{\pi}_{12} = 0.058$, $\hat{\pi}_{21} = 0.119$, and $\hat{\pi}_{22} = 0.763$. According to the joint model, 5.9% of the subjects performed at the "low" level on Old-same item at Time 1 and Time 2, while 76.3% of the subjects performed at the "high" level on these items at both times. These four estimates sum to 1 within rounding error. According to the posterior probabilities, the observed proportion of subjects who performed at the "high" level on these items at both Time 1 and Time 2 was $\hat{p}_{22} = 0.775$; very close to the parameter estimate of $\hat{\pi}_{22} = 0.763$ ($z = 0.4$, $p = \text{n.s.}$) suggesting that the model is in close correspondence with the data.

The estimates of the transition weights are in the third line of each cell of Tables 21 - 25 with their standard error presented in parentheses immediately below. In addition, the observed proportion of subjects from component a of variable X who "move" to component b of variable Y, denoted \hat{t}_{ab} , are presented to the right of the $\hat{\tau}_{ab}$. For example, from Panel A of Table 21, of those subjects from the

"low" component on Old-same items at Time 1, $\hat{\tau}_{11} = 0.506$ and $\hat{\tau}_{12} = 0.494$.

This indicates that of those subjects who perform poorly on the Old-same items at Time 1 (the "low" component) approximately half remain in the "low" component at Time 2 while approximately half change to the "high" component on Old-same items by Time 2.

Again there is close correspondence between the observed proportion based on the univariate posterior probabilities and the parameter estimate. For example,

$\hat{\tau}_{12} = 0.494$ is non-significantly different from the observed proportion $\hat{t}_{12} = 0.700$

($z = 1.6$, $p = \text{n.s.}$). In most cases there is a close correspondence between

the observed and estimated transition values. The bottom row of each cell contains the frequency of subjects in each joint component as computed by posterior probabilities.

However, when the observed and estimated quantities presented in each panel correspond less than perfectly, the model values are more reliable estimates of the population because they avoid the errors made from classifying subjects using posterior probabilities. Correlation and fit statistics to be evaluated below are also presented for each panel.

Many of the 2 x 2 contingency tables in Tables 21-25 are of three basic types described in Table 26. Each of the three panels in Table 25 represent possibilities in joint performance, where the "X's" are used to denote a relatively large number of subjects and the "O's" a relatively small number of subjects. When the two variables represent item-sets consistent over Time, Type A of Table 25 suggests that Variables X (left side) and Y (top) are of approximately equal difficulty, while Type B suggests that Variable X is easier than Variable Y because there are a greater number of subjects classified as "high-

level" performers on Variable X and "low-level" performers on Variable Y than the reverse. Likewise, Type C suggests that Variable Y is easier than Variable X.

Response Bias. Panels B, D, J, and K of Table 21 all measure performance on Same vs. Different items, within Time and Item type. Consistent with the univariate results, Panels B and D of Table 21 show that Same items are more easily solved than Different items for the original Shepard and Metzler (1971) items, but panels J and K show the opposite; that Different items are easier than Same items for the newer, more complex items. One interpretation of this finding is that Old items evoke a "Same" response bias, while New items evoke a "Different" response bias. Panels B and D show a "same" bias on the Old items as seen in the fact that the π_{12} are all non-significantly different from 0 (all z 's < 2.0 , for instance in Table 21 Panel B, $\hat{\pi}_{12} = 0.005$, $SE(\hat{\pi}_{12}) = .011$, $z = 0.5$), while the π_{21} are significantly different than 0 (all z 's > 2.0 , for instance again in Panel B, $\hat{\pi}_{21} = 0.177$, $SE(\hat{\pi}_{21}) = .011$, $z = 16.4$).

This pattern of significance indicates that while there are a significant number of individuals who respond "Same" to Old-same and Old-different items, there are none who respond "different" to both Old-same and Old-different items. This response pattern indicates a bias to respond "same" on the Old items by the individuals who perform at seemingly high levels on the Old-same items and low levels on the Old-different items. Panels J and K show a "different" bias on the New items in that the reverse pattern of significance is observed (that the π_{21} are all non-significantly different from 0, while

the π_{12} are significantly different than 0). Again, this indicates that while there are individuals with a "different" bias on New items, there are none with a "same" bias.

Panels C and G of Table 21, which show joint performance on Old and New items would also be classified as Type C, corroborating the differential response bias interpretation. The observed component memberships in these panels are consistent with subjects demonstrating a "Same" bias on Old items and a "Different" bias on New items in that subjects who respond "same" on Old items also respond "different" on New items. For the Old-same items this is a correct response, while on New-same items it is incorrect. As a result these subjects perform at the "high" level on Old-same items and at the "low" level on New-same items. The fact that panel I of Table 21, which shows joint component membership on New-same items over time, is also of Type B, while panel L of Table 21, which shows joint component membership on New-different items over time, is closer to Type A, suggests that there is reduction in that bias over time as might be expected with learning.

It was argued earlier that males' and females' performance is indistinguishable within component group. If this hypothesis is true, then it should also be the case that males and females within these groups are indistinguishable in terms of their biases. Tables 22 and 23 provide the twelve 2 x 2 tables identified as Panels A through L with model estimates for males and females, respectively. Males and females show the same response biases in Panels B, D, J, and K of Tables 22 and 23. For example, Panel B of Tables 22 and 23 both indicate that π_{12} is non-significantly different from 0 ($z_M = 0.1$, $z_F = 0.2$, p 's = n.s.), while π_{21} is significantly greater than 0 for both males and

females ($z_M = 11.0$, $z_F = 7.4$, $p's < .05$). Panels D, J, and K are identical in terms of this bias. In fact, the proportion estimates for the zero-valued cells are non-significantly different for males and females in all cases (all $z's < 2.0$, for example $\hat{\pi}_{12}$'s from Panel B in Tables 22 and 23 are 0.001 and 0.025 respectively; $z = -0.8$, $p = n.s.$).

Improvement Over Time. Panels A, F, I, and L in Table 21 provide information about the rate of improvement. In this case, the τ_{ab} 's are of interest. Recall that the $\hat{\tau}_{ab}$ estimate the proportion of the population in component a on variable X who shift to component b on variable Y. For example, from Panel A of Table 21, of those individuals who performed in the "low" component on Old-same items at Time 1, $\hat{\tau}_{11} = 50.6\%$ remained in the "low" component on Old-same items at Time 2. Additionally, $\hat{\tau}_{12} = 49.4\%$ changed to the "high" component at Time 2. By the same token, only $\hat{\tau}_{21} = 13.7\%$ of subjects started in the "high" component at Time 1 and shifted to the "low" component by Time 2. This panel shows that a significant proportion of subjects improved (testing the null hypothesis that $\tau_{12} = 0$, $z = 3.9$, $p < .05$). Panel F of Table 21 shows performance on Old-different items, also indicating that subjects improved over time ($z = 85.5$, $p < .05$). However, on Panels I and L, which shows improvement on New-same and New-different items, none of the transition parameter estimates which indicate improvement are significantly different than 0. In other words, subjects improved on Old items, but not New items.

As with the measure of response bias, male and female performance changes over time can also be compared. Panels A, F, I, and L of Tables 22 and 23 indicate that male and female performance was similar to the overall pattern. For Panels A and F, which show performance change for Old-same and Old-different items respectively, the τ_{12} 's are significantly different from 0 for both males and females, while for Panels I and L, the τ_{12} 's are not significantly different than 0 for either sex. These results indicate that both males and females who began in the "low" component improved on the Old items, but not the New items.

Males and females did not, however, improve identically. As suggested by the univariate results in Chapter IV, more females changed from the "low" component to the "high" component. By comparing π_{12}^{Male} with π_{12}^{Female} in panels A, E, I, and J (those panels which show improvement over time), only Old-different items (panel F) shows a significant difference (favoring females, $z = 2.7$, $p < .05$), indicating that the proportion of subjects who begin in the "low" component and finish in the "high" component is higher for females than males on the Old-different items, but no others. Interestingly, by comparing τ_{12}^{Male} and $\tau_{12}^{\text{Female}}$ in panels A, E, I, and J, there are no significant differences except on the Old-different items (Panel F) which favors males ($z = 3.4$, $p < .05$). It appears that while a greater *number* of females show improvement, a greater *proportion* of males show improvement. Because there are more females than males in the "low" component initially, their smaller proportion still proves to be a larger total number. In order to examine this hypothesis further, the $\hat{\pi}_{12}$'s were summed

for both males and females to test whether there was any overall difference between the proportion of males and females who move from the "low" component to the "high" component across the four item-sets. The total proportion of females who improved was 0.69, while the corresponding proportion for males was 0.51 ($z = 2.0$, $p < .05$). A similar gross measure of the rate of improvement can be estimated by examining the sum of the relevant transition estimates ($\hat{\tau}_{12}$) for males and females. Results indicated that the average of the transition parameters in the cells which show improvement (the upper-right cell in Panels A, E, I, and J) was higher, but not significantly so, for males than females ($z = 1.2$, $p = \text{n.s.}$). While males and females improve at the same rate, more females change from the "low" to the "high" component over time because a greater proportion of females are performing at the "low" level initially.

Item Complexity. The remaining panels, C, E, G, and H, feature bivariate performance on Old and New items. These panels indicate that the New items were more difficult than the Old items in that the joint proportions (π_{ab}) show a greater proportion of the population perform at the "high" level on the Old items and at the "low" level on the New items than at the "low" level on the Old items and the "high" level on the New items ($\pi_{21} > \pi_{12}$ for all four panels, z 's > 2.0 , $p < .05$). Again, males and females perform nearly identically in this regard. Panels C, E, G, and H of Tables 22 and 23 show that for both males and females, $\pi_{21} > \pi_{12}$, for all but males in Panel G. In that case, both π_{21} and π_{12} are non-significantly different from 0.

Curriculum. The effects of curriculum can also be evaluated in the bivariate setting. Tables 24 and 25 provide bivariate model estimates for Penn State and Cooper-

Union subjects, respectively. Panels A, F, I, and L, which show improvement over time on the four Item-type by Item-status problem sets are the most relevant in assessing the effects of curriculum. Comparisons of the π_{12} , which indicate the proportion of subjects who demonstrate improvement, show that a greater proportion of Penn State subjects improve on Old-different and New-same items (Panels F and I; z 's > 2.0 , p 's $< .05$), but not on either Old-same or New-different items. The general improvement gains attributed to Penn State subjects in the univariate results appears to be the result of improvements on these two types of items.

Summary-level Analyses

It was suggested in the previous chapter that the relationship between the two component basic-level models and the three component summary-level models could be understood in terms of how items were summed within subject. That is, subjects who perform at the "low" level on both, say Old-same item at Time 1 and Old-different items at Time 1, would perform at a lower level on Old items at Time 1. Subjects, however, who performed at the "high" level on one type and the "low" level on the other would perform at some "intermediate" level with a probability of success equal to the average of the "high" and "low" probabilities of success. Uniformly "high" performers would perform at the "high" level on the combined set of items.

These three kinds of performers would be seen in Panels B, D, J, and K of Tables 21-25. The proportion of jointly "low" performing subjects are represented by

$\hat{\pi}_{11}$, "high" performers by the $\hat{\pi}_{22}$, and the "intermediate" performers by the $\hat{\pi}_{12}$

and $\hat{\pi}_{21}$. As noted earlier though, these four panels fit either pattern B or C of Table 26.

In fact, any panel which fits patterns B or C (of Table 26) implies that there are, in effect, three joint components. For example, panels B and D of Table 21 (which are of Type C) suggest that there are three "kinds" of subjects: those who do poorly on both Same and Different items, those who do well on the Same items but do poorly on the Different items, and those who do well on both items. There are very few subjects who are successful on the Different items and unsuccessful on the Same items. This suggests that when Old items are examined independently of Same/Different Status, a three component model should best describe the data. This was the primary reason behind the use of three component models to describe the summary-level items in the previous chapter.

To test this hypothesis, membership in the four bivariate groups of Panel B (for Old-same and Old-different items at Time 1) were identified and compared with component membership for the three univariate summary-level components for Old items at Time 1. The top panel of Table 27 provides frequencies for the cross-classification of component membership. For example, Panel B of Table 21 shows that there were 38 subjects jointly classified as "low" performers on both Old-same and Old-different items at Time 1. Of those 38, Table 27 shows that 30 were at the "low" on Old items at Time 1, 8 were identified as "intermediate" level performers, and 0 were at the "high" level. The encircled frequencies represent the modal frequency for each column and indicate correspondence between the two model formulations. A chi-square test of independence (between the two classification schemes) rejects a lack of association between the two models. Table 27 clearly demonstrates how subjects in the bivariate two component model are partitioned in the univariate three component model.

New-same and New-different items (panels J and K of Table 21) are similar in structure, but are more like Table 26's Type B. The bottom panel of Table 27 compares

joint two component membership on Same and Different New items at Time 1 with Univariate three component performance for both New items at Time 1. As seen in the table, there is a very high correspondence between component membership groups.

Of course, part of this association is the result of the measurement scheme. It would not be possible, for example, for an individual who performed in the "low" group on both Old-same and Old-different items to be in the "high" category on Old items. However, the fact that a three component univariate solution provides a good fit and corresponds well to the bivariate model of performance lends additional credence to the notion that there are three different components when same and different items are collapsed. There was no a priori reason to assume that the joint components would have this structure, and it is certainly not specified by the model in any way as seen in the fact that not all of the 2 x 2 tables can be classified as Types B or C. While not provided, the correspondence is equally compelling at Time 2.

Three component bivariate models are presented for the summary-level Old and New item sets in Table 28. The univariate estimates were taken from the restricted three component univariate models using all subjects. Parameter estimates, standard errors, and cell frequencies follow the same format as the 2 x 2 tables. Panels C and D of Table 28 show performance over time for Old and New items respectively. Consistent with the finding presented in Chapter IV, subjects did improve significantly over time. Summing

$\hat{\pi}_{12}$, $\hat{\pi}_{13}$, and $\hat{\pi}_{23}$ to estimate the total proportion of subjects who

improved over time (i.e., those who moved from a lower component to a higher one), the proportion of subjects who improved over time is significantly greater than 0 for both

panels (for example from Panel B, $\hat{\pi}_{12} + \hat{\pi}_{13} + \hat{\pi}_{23} = 0.182$, $z = 4.5$, $p < .05$).

While the proportion showing improvement for the New items is greater than the proportion showing improvement over the Old items, it is non-significantly so. The univariate results, in contrast, demonstrated that subjects improved to a greater extent on the New items. This result is most likely due to both the differences between the univariate and bivariate samples and the fact that the $\hat{\pi}_{12}$ values estimate gross increases in performance, while the $\hat{\pi}_2^{(\text{Time } 2)} - \hat{\pi}_2^{(\text{Time } 1)}$ difference in Chapter IV estimates net performance increases.

Panels A and B reflect performance over Item type for Times 1 and 2 respectively. These panels measure the relative difficulty of Old and New items. The proportion estimates from the cells which show improvement over time in Panels C and D ($\hat{\pi}_{12}$, $\hat{\pi}_{13}$, and $\hat{\pi}_{23}$) indicate that the New items are more difficult than the Old items because there are no subjects who move from a lower component group on Old items to a higher one on New items. Table 29 shows cells of Table 28 with τ_{ab} values significantly different from 0 (although Panel D - cell 3,1 - is non-significantly different from 0). Table 29 defines the joint proportions of subjects on Old and New items (at Time X - either Time 1 or Time 2) who fall into these cells. For example, those subjects performing at cell 2,1 are at the "intermediate" level on Old items and the "low" level on New items. While there are two groups of subjects who perform at the "intermediate" level on the Old items, this cell is predominantly made up of subjects who perform at the "high" level on Old-same items and at the "low" level on the Old different items (see Panel B of Table 21, OS1 "high", OD1 "low" outnumber OS1 "low", OD1 "high" by a 10 to 1 margin). These subjects perform at one of three levels on the New items as

described by the transition parameter τ_{2b} , where b indicates which of the three New item components subjects are drawn from. The only τ_{2b} significantly greater than 0 is τ_{21} , indicating that the majority of subjects who perform at the "intermediate" level on Old items perform poorly on both New-same and New-different items. The most common subject performances for the remaining four cells in Table 29 are identified similarly. In effect, the specific pattern of five nonzero τ_{ab} 's on the four Item-type by Item-status item sets (e.g. Old-same, Old-different, New-same, New-different) is apparent. This pattern is consistent with a hierarchy of item difficulty, where New-Same items are the most difficult, followed by New-different items, Old-different items, and Old-same items which appear to be the easiest (see Figure 17). Read from top to bottom and left to right, each of the five groups adds one more item set at the "high" level. If a subject is likely to belong to the "high" category on any item set, it is the Old-same items. If a subject belongs to the "high" component for only two item sets, it is the Old-same and Old-different items. Given the univariate results, which also show this pattern, this appears to be a robust finding. However, because each cell of the 3 x 3 bivariate tables is made up of four subgroups, the relationships between variables are not as clear as it is when examining the bivariate basic-level item-sets. As such, no other analyses will be conducted on the 3 x 3 tables.

Correlation

Correlations within the present model have a very different interpretation than the conventional bivariate normal correlation. In the current model setting, the interpretation of correlations between two sets of items is the result of component membership, not the conventional linear structure seen in the bivariate normal case. For each of the modeled

item-sets (i.e., panels) in Tables 21-25 and 28 three correlation values are given. The first correlation value, identified as the raw score correlation, is the familiar Pearson product moment correlation between the scores on variables X and Y. Panel A of Table 21 shows the Pearson correlation between Old-same items at Time 1 and Old-same items at Time 2 as $r = 0.205$ ($p < .001$) indicating that the scores are correlated. The second correlation presented, identified as the component score correlation, is r_{ϕ} a measure of the correlation between the dichotomous component scores for variables X and Y. Subject's scores were classified as having been drawn from the "low" component (component score = 1) or from the "high" component (component score = 2) and the component scores were correlated across the two variables X and Y. In Panel A of Table 21, the correlation between the component scores is $r = 0.148$ ($p < .014$) indicating that the component scores are significantly correlated across time for Old-same items. It was argued earlier that component membership drives the correlation between the two variables X and Y, not the correlation between the raw scores per se. The component score correlation estimates the effect of the components on the overall correlation.

Appendix B outlines the model's expected correlation as a function of the proportion (π) and probability of success (θ) parameters. When estimates replace parameters, $\hat{\rho}$, measures the association between variables X and Y as predicted by the model. For example, based on the univariate model estimates for Old-same items at Times 1 and 2, and the estimated transition parameters, the model predicts the correlation between X and Y to be $\hat{\rho} = .102$ ($p = \text{n.s.}$). Estimates for all three correlations can be obtained and compared to one another. If the joint model provides a sufficient description of the data, all three will be similar, especially the predicted model and the

observed raw score correlation. This is because the model correlation and the component score correlation both, in some sense, measure the association of the two variables X and Y as a function of the component groups. If the raw score correlation is significantly larger than the model correlation, then the component groups cannot account for the observed relationship between X and Y . If, on the other hand, the model correlation successfully predicts the observed raw score correlation, then the model gains added credibility. Each of the three correlation coefficients in Tables 21-25 and 28 have superscripts A, B, or C to indicate significant differences. Correlation coefficients with the same letter are non-significantly different from one another within each panel.

For the majority of panels in Tables 21-25 and 28, the model correlation is in very close agreement with the raw-score correlation, indicating that the raw score correlation is due to the proportion of subjects within each latent class (and that the model is sufficient to make predictions about the data). The component score correlation, which takes into account only the correlation of the latent class levels also closely approximates the observed raw-score correlation for nearly all of the contingency tables. Almost without exception, the three measures of association are in close correspondence, even though significance tests do detect a greater number of differences than would be expected by chance.

The joint model used here also predicts that within each of the joint components the correlation between the two item-sets should be 0 as a consequence of local independence; a feature common to latent class models. Local independence specifies that *within* each joint component, all variance is random error, thus the correlation between the two joint variables is zero. This sometimes counter-intuitive notion can be explained by the following example. Consider voting behavior for U.S. Senate and

Representatives. Were a random sample of the voting population taken, the correlation between these two variables would probably be very high. However, if a latent variable like political affiliation were found, one might expect that within each level of the latent variable (liberal and conservative), the correlation would disappear. That is, when only liberals are examined, there would likely be no correlation between Senate and Representative votes. Most liberal individuals would presumably vote democratic in both races, and idiosyncratically for either mixed or solely republican candidates. In effect, the correlation is caused by the levels of the latent variable. The same should be true with regards to mental rotation performance. Within each component group, the correlation between items should be zero, while the correlation of the groups (component score) should estimate the raw-score correlation between the variable pairs.

The within component correlation between scores on task A and B can be evaluated to test this. For example, from Panel A of Table 21, 32 subjects were jointly classified as performing at the "high" level on both Old-same items at Time 1 and the "low" level on the Old-Same items at Time 2. The number of items correct on the Old-same items at Time 1 can be correlated with the number of items correct on the Old-same items at Time 2 for these 32 subjects (in this case, the within component correlation of performance on Old-same items at Time 1 and Old-same items at Time 2 is $r = 0.201$, $p = \text{n.s.}$) and each of the other remaining 47 joint cells in Panels A-L of Table 21. These correlations are typically non-significantly different from 0, as hypothesized. Figure 60 provides a funnel graph of the within cell correlation coefficients. The graph gets its name from the funnel-shaped ± 2 standard error region. The standard error lines are curved because as sample size increases, smaller and smaller r values become significantly different than 0, using Huber's (1977) distribution-free result. The interior

area of the funnel graph depicts correlations non-significantly different than 0, while points falling outside of the funnel are significantly different than 0. As seen in the figure, there are a few correlations which do fall in the region of significance (outside the funnel). However, the small number of violators indicates that within cells, most of the variance is simply random error and that the model is on the right track.

Model Fit

The top panels of Figures 51-53 show observed joint frequencies, while the bottom panels show frequencies expected under the bivariate model. The close apparent correspondence between the two suggests that the bivariate mixture model provides an excellent description of the data. Observed and expected values appear to correspond very closely. The usual method of assessing model fit in this type of joint setting involves some kind of χ^2 goodness-of-fit procedure. The two χ^2 statistics used in the previous chapter become problematic in the current setting, however, because a large number of observed and expected frequencies less than 1 in the $m_x \times m_y$ celled joint frequency histograms (see Figures 51-53). Commonly, adjacent levels of the variables are pooled together so that frequencies are sufficiently large to permit the Pearson χ^2 goodness-of-fit test. While these fit statistics are not presented, the majority reject the bivariate mixed binomial model. However, there is some recognition that what appears to be a close relationship between a model and data can be rejected by χ^2 fit statistics, at least in tests of independence (Diaconis & Efron, 1985). Diaconis and Efron (1985) argue that with sufficient power, virtually all models will be rejected using standard hypothesis testing techniques. The rejection of a model based on a χ^2 test statistic does

little to assist in finding a correct alternative, and often the alternatives are less well supported by the data. In other words, while a specific model may be rejected, it can still be both the most appropriate and the most useful model for the data. For this reason, and because so many of the $m_x \times m_y$ cells have values less than 1, an alternative method of computing χ^2 was used. Thomas (1977) suggested that when both the observed and expected frequencies are less than 1 for a given cell, that cell's contribution to the χ^2 test statistic should be negligible because of the agreement between observed and expected values. Following this approach, the χ^2 test statistics employed here were modified such that if both the expected and observed frequencies for a cell were less than 1, then the cell χ^2 was set to 0. The degrees of freedom remained unchanged.

Adjusted Pearson goodness-of-fit and likelihood ratio chi-square statistics are presented with their associated degrees of freedom for each of the joint models in Tables 21-25 and 28 below the panels containing the bivariate parameter estimates which generated the fit statistics. The adjustment for small cell values precludes referencing computed χ^2 values to tabled critical values, but from an intuitive perspective, the overall fit of the models appears satisfactory. In many cases, for example Panel F of Table 21, the test statistic is approximately equal to the degrees of freedom, indicating a close correspondence between the model and the data. In many cases, however, the joint binomial mixture models are rejected at the $\alpha = 0.05$ level. However, given the relative power of these tests and the apparently high correspondence between the observed and expected frequencies in Figures 51-53 the data fit remarkably well. Certainly the fit is better than many alternatives, such as the commonly used bivariate normal model.

Summary

The joint binomial mixture model provides interesting insights into the nature of mental rotation performance. The 2 x 2 tables which detail performance on the basic-level item-sets suggest that when Same and Different items are combined, there are three kinds of performers, lending support to the evidence for three components found in the univariate analyses. These joint tables also serve to illustrate an as yet unexplained response bias that differs for Old and New items. Additionally, the joint model supports the pattern of sex differences over time seen in the previous chapter. The joint model also successfully predicts the observed correlations between item sets, yet explains them in a very unconventional way. These findings are interpreted in detail in the following chapter.

CHAPTER VI

Discussion

The model presented above takes the two component joint mixture model posited by Thomas and Lohaus (1993), expands it to fit three components jointly, and applies it to longitudinal data. In addition, the model is applied to two types of items (Old and New), and by examining response type (same or different), illustrates the importance of gender and curriculum on mental rotation performance changes over time. The major findings of the current study suggest that a two component mixture of binomials model of performance best fits the item-sets which differentiate between same and different items. The model finds fairly consistent but not identical probability of success estimates for each component or latent class across curriculum (Penn State/Cooper-Union), Sex (Male/Female), Item-type (Old/New), Item-status (Same/Different), and Time (1/2). When items are collapsed over the Same/Different variable, a three component model best captures performance. While the mixture often provides a less than ideal fit, it is a substantial improvement over the normality often assumed for these kinds of data. Evidence supporting a latent class view of mental rotation performance is consistent with a growing body of literature suggesting that there is something fundamentally unique about spatial cognition in this regard (Thomas & Lohaus, 1993).

One might, however, argue that the model had not fulfilled its promise, that the fit was somewhat less than stellar, and that perhaps, since it had not been uniformly accepted, it should be entirely suspect. Like any model of performance it has its

shortcomings. Furthermore, the large sample sizes presented here give tests so much power that any model will be rejected. In many ways the model used here is exploratory. Still, it is a better way to view performance on mental rotation tasks than the alternative normal distribution models. Clearly, while the mixed binomial does not fit perfectly, to argue that its fit is imperfect is to simultaneously argue for the rejection of the normal model because while the mixed binomial model is not perfect it certainly appears much closer to reality than any normal model. While conclusions based on the mixed binomial may be less than certain, conclusions based on the normal model are even more so. The mixed binomial model is more believable than the normal based models because it seems to capture the spirit of the data: for any given set of items there are two "types" or "kinds" of performance. One of the most useful features of the model is that it can help guide thinking about performance-related issues.

The use of the word "types" or "kinds" has been used to refer to individuals, but it is probably best applied to *performance*, rather than *performers*. In the present context the latent class terms do not imply that each person could have a stamp placed on the forehead describing them as "high" or "low" ability subjects because individuals often perform at a "high" level on "same" items and a "low" level of "different" items or vice-versa depending on the complexity of the items. The terms might be more accurately characterized as "states" or "performance levels" that are likely related to strategies or problem solving methods (perhaps more nebulous and less conscious than a strategy). The components of the mixture model describe two "performance levels" when the item status (i.e., "same" vs. "different") is taken into account, and that proportion differences are the most important (i.e., the state in which a subject performs for a given type of

mental rotation problem). Because this model better describes performance than those traditionally employed, it should lead to better evaluation of psychological theories of mental rotation performance.

In general, between group differences are the result of differences in the proportions of subjects belonging to each latent class (π 's), not differences in probabilities of success (θ 's) which are generally consistent across item-sets. For example, sex differences were characterized by a larger percentage of males than females in the highest performance group and a lesser number of males than females in the lowest performance group (see Tables 5-8 and Figure 25). Penn State and Cooper-Union subjects perform at generally equivalent levels, although Penn State subjects did appear to improve more over time than Cooper-Union subjects, at least on some items (see Figures 20 and 22). Subjects were more accurate on the Old items taken from the Vandenberg and Kuse (1978) mental rotation task than they were on the newer, more complex items, as expected. This difference was manifested in the large proportion differences in "high" component membership between Old and New items (see Figures 17-19). Finally, there was a uniform improvement in performance over time for Penn State subjects, especially the females (see Figure 22).

When compared on different achievement measures, subjects classified as "high" performers were similar to one another regardless of sex, as were those classified as "low" performers (see Tables 11-16). Likewise there were no within component individual differences with regard to accuracy over different rotation angles (see Figures 36-40).

The modeling framework for understanding mental rotation performance presented here provides a means for integrating many previously unrelated or seemingly contradictory findings in the literature. Lohman and Kyllonen (1983) argue that no good models of individual differences in spatial abilities exist. This model attempts to fill that gap by providing interpretations for varied empirical findings. The discussion focuses primarily on the hypotheses outlined in Chapter I, beginning with the issue of performance change over time. The issue of stimulus complexity is considered next, followed by strategy-related issues.

Performance Change over Time

The model presented here could have found performance change to be the result of (1) consistent probabilities of success (θ) paired with changing proportion (π) values, (2) changing probabilities and constant proportions, or (3) jointly changing probabilities of success and proportions. Each implies different processes governing change. The first alternative appears to be the most likely. The probability of success estimates are largely consistent over time, but fluctuated across item-sets. Although not as consistent as had been assumed, the θ -estimates do not appear to vary in any systematic fashion. The variability of the $\hat{\theta}$'s was undoubtedly due in part to a failure of the model assumptions concerning the consistency of trial to trial success rates. However, the probability of success estimates are still fairly similar within a component group, while the proportion estimates appear to change in ways consistent with hypotheses. The bivariate binomial mixture model of performance presented here describes performance change in terms of transition parameters which describe shifts in component membership. Individuals within each of the two or three latent classes at Time 1 are associated with a

probability of either changing performance groups or remaining in the same performance group at Time 2. Were performance changes the result of small changes in probabilities of success, the joint model would not have fit as well as it did.

The fact that performance improved over time supports the idea that training was effective. From a psychological perspective, the question of which curricula evokes the most improvement is moot. It would be expected, based on numerous studies which find a link between study in engineering and mental rotation performance, that both would improve (e.g., Brinkman, 1966). Of most interest is how and why subjects improve, and whether that improvement is constant across item complexity (item type) and item status (same/different status). The issues of how and why are addressed in this section.

The fact that the training phase did not center on the kinds of items used at test or retest fails to support Olson and Bialystok's (1983) description of spatial cognition. Olson and Bialystok (1983) argue that improved spatial performance requires that the parts of objects be labelable, which is not likely in this case here. Because training did not focus on the same objects used during testing, it seems unlikely that the test items became substantially more familiar to subjects, though it is possible that subjects' ability to provide labels for abstract, unfamiliar objects improved with experience. On the other hand, Piaget's notions of more general operations, which require the conservation of length and angle and an internalization of an object's permissible transformations (i.e., invariance of angle and length) are better suited to explain the current data. The abrupt nature of performance change is more consistent with Piaget's theory.

The data do not provide unconditional support for Piaget's theory, however. As with the water-level task, many adult subjects appear to lack the operations required to

solve problems of imagery yet these should be evident by the end of concrete operations. Piaget was less specific about mental rotation issues than issues of horizontality and verticality, but the two are certainly related. One interpretation for less than optimal performance on this task is that subjects lack operations concerning invariant properties and reversible displacements of the rotation items.

In contrast to Piaget's conception of performance change, Lohman (1990) argues for a gradual accretion model of performance on mental rotation tasks. Model results here are at variance with this hypothesis. One reason for this apparent contradiction is that Lohman (1990) presents averaged data which do show gradual improvements over time. As Brainerd (1993) argues, average data can be very misleading because abrupt performance shifts by individuals will often appear to be gradual transitions when data are grouped.

The evidence which suggests that subjects identified as "high" and "low" level performers at time 1 differ on achievement tests would lead to the expectation of an Aptitude by Treatment Interaction of the kind specified by Kyllonen et al. (1984) and Cooper and Mumaw (1985). Such an effect, however, was not observed. For this to have occurred, the better subjects would have shown an improvement in their probability of success estimates, while the poorer subjects would not. According to the results of the unrestricted models, the higher performing subjects had the most consistent probabilities of success, while the poorer performing subjects had the most variable (and not always higher) probabilities of success over time. The fact that improvements in performance were seen as shifts in component membership makes the idea of a mental rotation aptitude by treatment interaction very unlikely.

Stimulus Complexity

Consistent with previous research which shows that item complexity adversely affects accuracy, the newer, more complex items showed lower average accuracy rates than the old items. However, this effect was not uniform across all component groups. The joint frequency tables (Tables 21-25) are useful in understanding the relationship between complexity and accuracy. For example, Panels C, E, G, and H of Table 21 show joint performance on Old (simpler) and New (more complex) items. Panels C and E, which show performance on Old- and New-same items indicate that New-same items were more difficult to solve. Many subjects with near ceiling accuracy on Old-same items were performing at the "low" level on New items. Panels G and H, however, show more consistent performance on Old-different and New-different items. The joint model suggests that subjects who succeed on Old-different items are also successful on New-different items.

The data reveal an interaction between item complexity and item-status such that new-same items are the most difficult (see Figure 28 and Panels B, D, J, and K of Table 21). Hochberg and Gelman's (1977) finding that landmark features improve accuracy provides one possible reason for this result. It would appear that New-different items would have more distinctive features than New-same items. Subjects may use a strategy whereby they respond "different" to an item if visible landmarks are absent because the target and comparison items are more likely to be different. This strategy would cause them to erroneously identify some "same" items as different when items are more complex because of limitations in working memory. Because the Old items lack the

same number of potentially distinctive landmark features, Hochberg and Gelman's (1977) findings might also explain why Old-different items are not easier than Old-same items.

Complexity effects are not universally found (e.g. Cooper, 1982; Yuille & Steiger, 1982). Results here suggest that this may in part be due to sampling differences. The highest level performers were unaffected by complexity (as were uniformly low performers), yet intermediate level performers were strongly affected by complexity. This result is due to response bias changes over item-sets of varying complexity, and will be discussed in more detail below.

Strategy Use

Lohman and Kyllonen (1983) argue that a model of mental rotation performance must take into account the fact that subjects may use different strategies for different tasks. The current model suggests just how this might be characterized. While the particular strategies subjects use are not directly penetrable, different strategies do imply different rates of success across the different item-sets, and different strategy groups imply a latent class structure. While the existence of latent classes does not provide a logical basis for assuming different strategy groups, there are enough effects consistent with strategy use to suggest a correspondence between the two. Further, the results discussed below provide reason to believe that each of the latent classes of performers are using different strategies. It would be very unlikely (although admittedly possible) that different strategies would lead to similar probabilities of success. In addition, the latent class structure of mental rotation performance found here can rule out certain strategies.

For example, it was originally hypothesized that the lowest component of a three component model would be associated with a guessing strategy and have a probability of

success of about .50. This appears not to have been the case. In addition to the fact that the low performers' probabilities of success are significantly larger than .50 for nearly every item-set, histograms of within component item accuracy (Figures 41 and 43 - the worst fitting examples) show that response accuracy is not focused at .50. Subjects appear to be responding below chance on some items at about the same level as intermediate-level performers on others. This finding is consistent with evidence provided by Just and Carpenter (1985) who found four discrete strategies. The lowest performing subjects in their study rotated objects about three axes sequentially. This strategy is consistent with the current data which suggest that the summed rotation angle over all three axes distinguishes the two clusters of Figure 43, with more difficult items (left-most cluster) averaging larger trajectories. This finding is also consistent with Lohman and Kyllonen (1983) in that the lowest performing subjects may shift strategies within item-set. In other words, the fact that the θ_1 -estimates were so much more variable than the θ_2 -estimates could be the result of subjects changing strategies on, for example, Old-same items. The fact that the items of each set were mixed when presented further supports a strategy shifting hypothesis.

It also appears that something related to strategy use affects performance when item-sets are collapsed across the Same/Different variable. Subjects respond as if they have a "same" response bias on the Old items and a "different" response bias on the New items. This bias is only apparent for one of the three latent class groups (or one of the four in the two component bivariate setting). There is one set of subjects who perform at near ceiling rates on both Old and New items and another group that performs at very

low levels regardless of the item-type. The intermediate-level performance group does well on the one type, but not the other.

This response bias may be caused by a feature-matching strategy. The pattern of response bias seen in Panels B, D, J, and K of Tables 21-25 is also consistent with a feature matching or piecemeal rotation strategy. Subjects using either strategy might be more likely to identify more complex (New) items as different based on visible features, even when the items are the same, by confusing the features of the target and stimulus items. Carpenter and Just (1978) have shown using eye fixation data that as mental rotation items become more complex, some subjects fixate on non-corresponding target and stimulus segments. As a result, as segments become confused, and subjects are more likely to respond "different" to same items. This was exactly the pattern found here. With increased item complexity, one sub-set of subjects was more likely to respond "different" to same items. When the items were less complex, however, subjects were more likely to provide incorrect "same" responses. Carpenter and Just (1978) also argue that incorrect "same" responses are the result of a failure to identify segments which distinguish targets and stimuli. For example, there are relatively fewer distinguishing features on the original (Old) items, so subjects looking for landmark features and not finding them may assume that the items require a "same" response. Because the less complex Old items have fewer features to compare and the features are more similar to one another than they are for the more complex New items, subjects' "same" bias on the Old items is consistent with their interpretation. In essence, the data presented here argues that subjects showing response biases are rotating stimuli in a piecemeal fashion. In contrast, those who performed at the "high" level on both items (for all four panels),

show no response bias and are likely to have used holistic rotation strategies. In addition, subjects of both sexes appear to have shown these patterns.

The findings presented here argue that subjects generally use the same strategy across trials, but that these strategies are not equally effective for all items. For example, feature-matching strategies should be more successful on Old-same items than on New-same items because there are fewer features to confuse on Old items. The larger number of features on the New items may make it difficult for subjects using either feature matching or piecemeal rotation strategies to keep the features straight. Subjects' confusion of features on the New items is one likely cause of a "different" bias on these items that is absent on the "same" items.

Strategy Use and Rotation Angle

All conclusions about angle of rotation must be made with caution since there was no systematic manipulation of rotation angle and each object was rotated on more than one axis. As a result, analyses of the three orthogonal axes of rotation are always confounded by rotations along the remaining axes. Furthermore, because there are no unique solutions for sequential rotations about three axes, subjects may have used trajectories other than the ones used in this analysis. As a consequence, the summed rotation angle (of the three orthogonal axes) can be different depending on the order of rotation axes used to solve for the angular deviations. With different orders, rotation angle about each axis is changed, as is the sum of the three angles. The object-defined axis is unique and can be used with more confidence, but caution must still be used because there is little evidence to suggest which subjects, if any, may have used this axis.

While many authors (e.g., Shepard & Metzler, 1971) have demonstrated that reaction time is affected by rotation angle in good performers, it does not appear that accuracy is affected to any great extent. Accuracy for the highest level performers in the current study were not obviously affected by rotation angle, which suggests that they were able to holistically rotate the items. Bethel-Fox and Shepard (1988) argue that complexity affects mental rotation in proportion to how well the rotation image is integrated into a unified whole. Additionally, these authors argue that high spatial subjects provide evidence for more effective encoding of spatial information than low spatial subjects. The fact that highest component group's accuracy was unrelated to rotation about the object defined axis, even when they use this axis, can be accounted for this way (see Figure 32). Conversely, subjects in the lowest component, who are the least likely to use an object-defined axis, would also not be expected to have accuracy correlate with this measure of angular deviation. The fact subjects from both components showed little relationship between accuracy and rotation angle is probably related to the unusual measures of rotation angle in this study.

It is easy to imagine that "different" items would be easier to distinguish from one another (as they are for the intermediate/poor subjects who may use a feature matching strategy), but not if a rotation strategy is used. For the highest level performers it also seems to be the case that same and different items are equally difficult, which also suggests a rotation strategy. Because the features are different for "different" items a feature-matching strategy implies different performance levels for "same" and "different" items. It appears that "high" component subjects use a consistent algorithm (e.g. Kail et al., 1984) for all problems since subjects do not know until the end of the algorithm

whether items will turn out to be the same as or different from the target (in the comparison phase). Carpenter and Just (1979, cited in Kyllonen et al., 1984) used eye-fixation data to differentiate subjects who use sequential rotation (piecemeal) strategies, and found that they have higher error rates than subjects who use holistic rotation strategies. Furthermore, Tapley and Bryden (1977) found that accuracy decreased with degree of stimulus rotation. Bethel-Fox and Shepard (1988) argue that longer rotation times result in a degradation of the to-be-rotated mental representation. Consequently, image degradation hurts those who rotate figures on a part-by-part basis more than those who do so holistically. High ability subjects' performance is not affected by increases in the angle of rotation about an object-defined axis partly because rotation strategies require less time to implement, and are thus less sensitive to the image degradation. Intermediate subjects, who performed well on some item but not others, were unaffected by rotation angle, suggesting that they used a feature-by-feature comparison strategy. In addition, it is more difficult to keep track of the individual pieces when image degrades, so part-by-part comparisons are even more difficult. Cooper and Mumaw (1985) argue that low ability subjects lack the ability to maintain mental images long enough to successfully transform them over multiple partial rotations.

Alternatively, these findings could be taken as evidence in favor of Lohman's (1986) research on speed-accuracy trade-off curves. It is conceivable that the higher level performers are at the top of the speed accuracy curve, in that the time provided them was more than sufficient to complete all items without sacrificing accuracy. While the task was not timed, it is possible that the less able subjects imposed their own deadline, hurried their performance, and became less accurate as a result. This becomes especially

possible given that the data were collected in a group setting. While it was one of the specific aims of this study to collect mental rotation data that were not influenced by solution time, the likelihood that some students saw their peers finishing and felt compelled to compete is a realistic concern. In effect, this may have become an untimed task for the better performing subjects and a timed one for the less able subjects. While their accuracy was not dependent on rotation angle, the subjects from the lowest component may have been at one of the lower points on the speed-accuracy curve.

Strategy and Complexity

Many authors have theorized that as items become more complex, the number of observed strategies should increase (e.g. Kyllonen et al., 1984). No support for this hypothesis was found. The fact that a two component model fits both Shepard and Metzler (1971) and more complex items equally well when same and different items were examined separately and a three component model fits both well when same and different items were examined together is compelling evidence against the notion that item complexity breeds an increased number of strategies for subjects. Complexity did appear to impact subjects' strategies, however.

Just and Carpenter (1985) found that high and low spatial ability subjects used different strategies for solving mental rotation items. While high spatial subjects rotated holistic mental images, one strategy common to those subjects with higher error rates was a piecemeal rotation strategy. Because subjects are thought to use non-holistic strategies as a result of inferior quality mental images, complexity should differentially increase error rates for this group. In fact, complexity effects should only be seen when either piecemeal or propositional strategies are used, but not rotation strategies (Bethel-Fox &

Shepard, 1988; Folk & Luce, 1987). Pellegrino and Kail (1982) argue that the likelihood of correct responses is directly related to the number and type of processing operations a task requires. For a holistic mental rotation strategy the operations are constant from trial to trial, as is the resulting probability of success (and therefore component group). For subjects who use piecemeal rotation strategies, the number of required rotations increases with item complexity. An increase in the number of operations increases the likelihood of an incorrect response because there are more opportunities for error, and a single process error will cause an incorrect response. Additionally, Cooper and Mumaw (1985) argue that subjects with lower quality mental images show a greater number of erroneous "same" responses. As a result, performance drops (from $\hat{\theta}_2 = 0.9336$ to $\hat{\theta}_1 = 0.6283$) as complexity increases for those who use piecemeal rotation strategies.

These hypotheses are consistent with the data found here. As expected by theories which posit analog mental rotation strategies (e.g., Shepard & Cooper, 1982), one sample of performers remain unaffected by complexity as would be expected if a holistic rotation strategy were being used. If a non-rotational feature matching strategy were used by the intermediate level performers (i.e., subjects whose joint categorization indicated "high" levels of performance on one task and "low" levels on another) their strategy might be expected to have differential impact on Old and New items which differ in complexity. Evidence from the current study is consistent with this hypothesis in that the proportion estimates for complex items are lower for New items than they are for Old items. In fact, this hypothesis implies that a proportion of individuals performing at "high" levels on the Old items will fall into the low level group when they attempt the

New items, as was found. Subjects identified as intermediate-level performers manifest complexity effects with a differential response bias for Old and New items. Panels B and J of Table 21 show that the intermediate-level subjects respond as if they have a "same" response bias on the Old items, and a "different" response bias on the New items. The use of both part-by-part and holistic rotation strategies is consistent with the current results.

Old items have fewer features to match and often the features are very similar even when the items are different. As a consequence, subjects using a feature matching strategy might be expected to respond "same" regardless of whether the items were the same as or different from the target. A feature matching strategy, therefore, would likely generate more "same" responses than a rotation strategy for Old items. In other words, a feature matching strategy would predict, that Old-same items were easier than Old-different items. Subjects using this strategy, however, would be more successful on items different enough to have visibly or remarkably different features. There would be, however, relatively fewer of these items on the original Shepard and Metzler style items than in the set of newer, more complex items. As a result, these subjects would also be expected to perform better on Old-different items than on either New-different or New-same items because there are fewer features to confuse on Old-different than on New-different items. Furthermore, subjects using a feature matching strategy would also be likely to show better performance on New-different items than New-same items because the New items have a large number of features which become confused. This is because both New-same and New-different items are likely to generate "different" responses when the same feature on an item and its target are confused. For the New-different items this is a correct response, but for the New-same items it is incorrect. Further, New-

different items were more easily solved than New-same items because complexity makes features more readily identifiable than in comparison to Old-different versus Old-same items which have fewer features to distinguish them.

It is also possible, that subjects change strategies depending on whether they are attempting Old or New items. Kyllonen et al. (1984) point to a range of results which suggest that subjects often change strategies across tasks. Lohman and Kyllonen (1983) provide evidence which indicates that some individuals use rotational strategies on less complex items and analytic strategies when item complexity increases. The pattern of results in Panels C and E of Table 21 indicate that the differential response bias found in this study could be explained if the intermediate-level subjects were changing from an effective rotation strategy for the Old items to a less effective part-by-part strategy for the New items as the result of increased complexity. Unfortunately, there is no evidence available here to distinguish these possibilities.

Strategy Change Over Time

Panels B, D, J, and L of Table 21 indicate that the same/different response biases were slightly reduced over time. Over time, subjects for the intermediate groups seem to migrate to the highest-level component. While there is no direct evidence that subjects can respond to strategy change over time, the reduction in the numbers of subjects who have a differential response bias suggests that strategies do change over time. The model also suggests that these changes are abrupt in the sense that they are the result of component membership changes, which are more likely the result of strategy shifts than incremental refinements in existing strategies.

Linn and Petersen (1985) point out that females are more likely than males to use part-by-part rotation strategies and that sex differences are smallest on tasks which require their use. If females are more likely to adopt a piecemeal rotation strategy, then sex differences should be largest on the most complex items. The current data show this to be the case at Time 1: The largest sex differences in performance were seen on the New items (see Figures 14 and 16). By Time 2, however these differences were significantly reduced. The fact that latent class membership changes were responsible for this reduction implies that change over time was the result of strategy shifts on the part of lower level performers. Although it is also possible that subjects simply became more efficient in their implementation of part-by-part rotations as suggested by Lohman and Nichols (1990) this explanation seems less consistent with the data.

Sex Differences

As has been found in other spatial tasks (e.g. Thomas & Lohaus, 1993) there appear to be multiple discrete groups defined by performance level. Sex differences reside not in the performance levels that define the mixing distributions, but in the differences in relative proportions of males and females within those distributions. Although the terms strategy and algorithm are not necessarily interchangeable, Kail's thesis that there are no sex differences in the algorithms used to solve mental rotation problems must be modified to admit that there are no sex differences in the algorithms used to solve mental rotation problems within a given latent class. Consistent with this view, males and females were not different on achievement tests (see Table 10), but high and low performance groups were (see Tables 11-17). However, this may partly be the result of the unique sample here. Engineering students of both sexes are academically

superior, especially in math and sciences, to the general student body. A broader sample of the population would probably have shown greater sex and component differences.

Further, the notion that the locus of sex differences is in reaction time is pervasive (e.g. Lohman, 1986). The current dataset suggests that this conclusion is unfounded.

Many of the studies which suggest this use much simpler two-dimensional rotation stimuli. The much higher accuracy levels reported by Kail et al. (1979) are probably due to the fact that the tasks administered here are much more difficult. Tapley and Bryden (1977), for example, note both accuracy and reaction times show significant sex differences, with males "on average" more accurate than females. Were data from the current study analyzed using traditional analysis of variance techniques, the same conclusions would have been drawn. However, the mixture model clearly shows that sex differences in accuracy result from component membership differences. The much more interesting question revolves around component group differences. In addition to sex differences in accuracy, Tapley and Bryden (1977) found a greater number of male "visualizers" and female "verbalizers," but no significant differences in terms of accuracy between strategy groups. One explanation for this was the relatively low proportion of women performing at the "high" level in their sample. This conceptualization agrees with Paivio's (1971) finding that imagery skill predicts spatial performance for males, but not females.

Kail et al. (1979) suggest that a significant subset of females employ a piecemeal rotation strategy because their rotation times are slower by a whole number factor (as opposed to some fractional number) than males' times. Current results suggest both males and females (but a higher proportion of females) use this strategy. It should be the

case, for example, that if subjects are partitioned based on their accuracy scores, the within sex reaction time differences should be greater than between sex differences within latent class. This hypothesis seems especially plausible given that Thomas and Kail (1991) found support for similar latent classes of mental rotation reaction times in both men and women.

The fact that no within component sex differences were found on the achievement test scores (Tables 11-16), accuracy rates across increasing rotation angle (Figures 36-40), or in the response patterns in the bivariate data pairs (Tables 22 and 23) supports the position that the most important differences are within rather than between sex. Liben (Sholl & Liben, 1995; Vasta & Liben, 1996) has argued, however, that at least on the Piagetian Water-level task, subtle differences may exist between males and females of the same latent class. This contention is likely true for mental rotation tasks as well. The fact that females' data were better modeled by the restricted two component model is evidence in favor of the notion that males and females may differ slightly even though their performance is quite similar. As noted earlier, however, the point is not that males and females are identical within latent class, but that their performance is almost indistinguishable, suggesting that their methods of solving mental rotation problems is also similar. As such, one would expect similar strategies and processing algorithms to be used. The current evidence supports such a contention. Again, while similar probabilities of success (θ 's) for males and females does not entail similar problem solving strategies, it is consistent with this hypothesis. Furthermore, it is very difficult to imagine that two radically different strategies would provide such similar probabilities of success.

While there was no overall difference in the rate of improvement for males and females (Figure 21), Penn State females' superior improvement (Figure 22) suggests a Sex by Treatment interaction similar to the one hypothesized but not supported by Baenninger and Newcombe (1989). Figure 61 provides a copy of their model of two potential Sex by Treatment interaction outcomes in the upper panel, and an alternative view in the lower panel which supports their hypothesis and the current data. While Baenninger and Newcombe (1989) view increases in spatial ability as ending abruptly after a certain level of experience, perhaps the logarithmic curve in the bottom panel might present a more realistic view of learning. The data presented in Chapter IV show that while both males and females improve, females who received the intervention improved to a greater extent. This condition is not possible in Baenninger and Newcombe's (1989) model, but it is perfectly consistent with the logarithmic curve in the bottom panel of Figure 61. In addition, a logarithmic curve like the one shown in Figure 61 can also account for why some studies support Sherman's (1967) hypothesis while others do not. For instance, the lower panel in Figure 61 implies that sex differences in improvement are related to initial performance differences. The greater the initial performance superiority for a group, the greater the subsequent relative improvement by the lower performing group. The Sex by Treatment interaction envisioned by Sherman (1967) would only be seen on tasks where females' initial performance is significantly lower than males'.

Because New items are more difficult (for males and females) performance begins at a lower point on the ability curve. A logarithmic curve predicts the greater rate (but not absolute level) of improvement for these more difficult items as compared to the

easier items. Figures 18 and 19 are consistent with this conception. For both males and females, the difference between Old and New items decreases with experience.

Figure 61 implies that "low" performers will show greater improvement than "high" performers. This is usually taken to mean that women will show greater improvement than men, because, on average, men outperform women on spatial tasks. The current model framework provides an explicit way to define "high" and "low" performance, regardless of sex. Studies of gender difference which use mean spatial test scores for males and females draw samples from the "high" and "low" components without knowing the composition of "low" and "high" performers directly. As a result, the current study is more apt to find a meaningful relationship between initial performance and improvement. The parameters of the mixture model can be directly related to the figure. If the amount of "spatial ability" for each sex is defined in terms of the proportion of individuals within the "high" component and improvement by the increase in that proportion, then females should improve more than males (at least for the Penn State subjects who received the curriculum intervention). Figure 62 shows the proportion estimates for the "high" performing males and females at Time 1 and Time 2 ($\hat{\pi}_2$). While only four data points does not provide extremely convincing evidence of a logarithmic relationship between ability and experience, it is certainly suggestive and requires further investigation.

Directions for Future Research

The major drawback of the current study was its inability to directly evaluate the strategies subjects use. Only two kinds of items were used in the current study, and no attempt was made to control for the complexity of each object. A broader and more

systematic manipulation of complexity, as measured by either the number of blocks or the number of arms on each item would facilitate a deeper understanding of the role of item complexity. The approach used by Siegler (1981) might hold special promise for untangling the different strategies subjects use, but designing objects in this paradigm is difficult. For example, the use of "different" objects which are not always mirror images (e.g., Yuille & Steiger, 1982) should be easier for subjects classified as intermediate-level performers if they do, in fact, use a feature-matching strategy. Subjects who rotate objects holistically may find both foils equally difficult from an accuracy perspective.

Additionally, only accuracy scores were measured here. If the hypothesis that "high-level" performers use holistic rotation strategies and "intermediate-level" performers use piecemeal strategies, then the two groups' reaction times should show a convergent latent class structure.

The fact that "same" and "different" items have different proportions but the same probabilities of success suggests that investigation from a signal detection framework might prove fruitful. The relationship between accuracy and latent classes might be captured by a latent class signal-detection model of performance which captures performance groups based on different distributions of d' . It may be the case that subjects' strategies involve different criteria setting for "same" and "different" items, or it may be that there are mixture distributions of d' that account for subjects' performance. It is important to rule out the possibility of shifting response criteria for Old and New items. For example, the differential response bias found for Old and New items could simply be due to a general criterion shift over different levels of item complexity. Folk and Luce (1987) have argued that a salary minus penalty for incorrect responses remuneration scale

is effective at increasing motivation. This approach may also be a useful method of manipulating subjects' response criteria from a signal detection perspective.

Another method of understanding strategy use is to identify the cognitive correlates and component processes of performance which distinguish latent classes. In other words, investigative efforts should first focus on the nature of within sex differences. The gender composition of samples should not be of major importance according to the findings of the present research (although, as Sholl & Liben, 1995 argue, both males and females should be sampled). What makes high-level performers different from low-level performers will lead naturally to explanations of sex differences. It appears to be strategy differences that result from differences in experience.

It was suggested here that strategy differences define each latent class and that visualization skill mediates subjects' strategy selection. Further examination of the visualization differences between latent classes could help decide this issue. Shepard and Cooper (1982) state that the quality of imagery is a necessary (but not sufficient) requisite for accurate performance on mental rotation tasks. This implies a non-equivalent conditional relationship between visualization to mental rotation performance. Now that high- and low-level performance can be categorized in a meaningful way (as opposed to more arbitrary methods like median splits), these questions become more tractable.

Additionally, one of the most intriguing areas of exploration relates to the effect of rotation angle on accuracy judgments. In the majority of mental rotation investigations, objects are rotated along only one of the three orthogonal axes (e.g., Shepard & Metzler, 1971). A systematic evaluation of how subjects respond to objects

rotated about multiple axes and whether the psychological and physical minimum angles of rotation correspond would provide a much more ecologically valid understanding of how individuals mentally manipulate objects. Furthermore, investigations of this type can also shed more light on whether rotation angle and accuracy are related within a latent class setting. In addition, one measure of rotation complexity might be defined in terms of the number of axes required for rotation. Some subjects may naturally rotate objects about each axis sequentially, while others may not. Results from other studies suggest that subjects with the highest level of spatial ability use an object defined minimum angle of rotation, while less able subjects rotate objects sequentially about each axis. A study involving accuracy and reaction times could test the hypothesis that "high" component subjects' reaction times linearly increase with object defined axes, while "intermediate" component subjects' reaction times increase linearly with summed rotation angle. It would not be difficult to ensure that the object-defined rotation angle and summed rotation angle were uncorrelated. This could provide further information about the nature of spatial representations and spatial cognition. It might also be the case that "high" and "low" performers change component membership (while θ 's remain constant) across X, Y, and Z rotation axes when examined singly, but this is an empirically testable hypothesis.

Finally, in order for issues of development to be more carefully addressed, a larger age range and time period should be studied. While learning was modeled successfully, it remains to be seen whether or not developmental change is similar. The model provided here is capable of answering such questions.

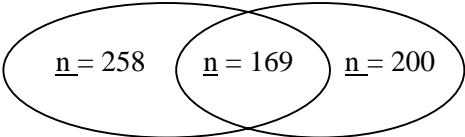
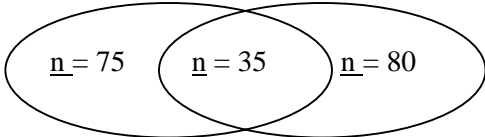
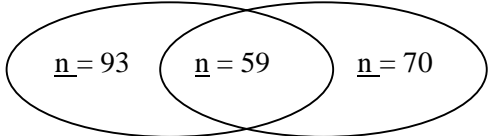
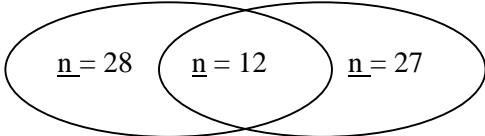
Summary

This framework for evaluating performance provides insights inaccessible to conventional modes of analyses. For example, the fact that within latent class there are no (or few) sex differences suggests that sex per se is not responsible for performance differences on mental rotation items. By looking at the proportion of items solved correctly without regard to component membership status, these findings would be washed out. The current framework allows mental rotation accuracies to be used to consolidate many seemingly inconsistent empirical findings regarding strategy use, sex differences, the effects of item complexity and rotation angle on performance.

CHAPTER VII

Tables and Figures

Table 1. Sample Sizes by Sex and Treatment Condition.

Sex	Treatment Condition	
	Solid Modeling	Traditional Course
	Time 1 → Time 2	Time 1 → Time 2
Male		
Female		

Note: Total sample size, $N = 556$. Sample sizes from Time 1 and Time 2 represent univariate analyses' sample sizes. Numbers inside both ellipses represent sample sizes of subjects who participated at both Time 1 and Time 2 and provided bivariate data.

Table 2. SAT Comparison Between Penn State and Cooper-Union First Year Students

College	Mean SAT*		Middle 50% **	
	Math	Verbal	Math	Verbal
Penn State	581	502	530-651	450-650
Cooper-Union	720	570	690-760	540-660

* American Council on Education (1992)

** Straughn & Lovejoy-Straughn (1995)

University (Curriculum)					
↓	Sex	Item Type		Item Status	Time
↓	↓	↓	↓	↓	↓
P	M	O	S	1	(258) = Penn State Males, Old Same Items at Time 1, n=258
	M	O		1	(333) = Males (Penn State & Cooper-Union), Old Items (Same & Different) at Time 1, n=333
C	F	N		2	(27) =Cooper-Union Females, New Items (Same & Different), at Time 2, n=27

Note: Letters denote variables which contributed to the samples. When letters are omitted, scores were summed across that variable. For example, within the sample MO1, University and Item Status are absent indicating that old-same items at time 1 and old-different items at time 1 were added for each subject to yield a sum of correct responses on all old items, increasing the total number of items. Similarly, MO1 denotes the fact that Penn State and Cooper-Union male samples were combined, increasing sample size.

Table 4. Descriptive Statistics

University	Sex	Item type	Item Status	Time	N	Mean	Variance	Range
Penn State	Female	Old	Same	1	93	7.80	1.75	3-9
			Different		93	11.42	8.64	0-15
		New	Same		93	10.12	6.54	4-15
			Different		93	6.10	3.63	1-9
		Old	Same	2	71	7.61	2.59	3-9
			Different		70	12.01	9.90	3-15
		New	Same		70	11.23	7.86	4-15
			Different		70	6.97	2.87	2-9
	Male	Old	Same	1	258	8.26	1.18	4-9
			Different		258	12.83	6.42	0-15
		New	Same		258	11.69	6.75	5-15
			Different		258	7.09	3.44	0-9
		Old	Same	2	200	8.00	2.59	3-9
			Different		200	12.97	7.79	4-15
		New	Same		200	12.25	7.31	4-15
			Different		200	7.52	2.20	3-9
Cooper-Union	Female	Old	Same	1	28	7.64	1.57	4-9
			Different		28	12.25	7.01	4-15
		New	Same		28	10.39	4.54	7-15
			Different		28	6.29	4.21	2-9
		Old	Same	2	27	7.74	1.43	5-9
			Different		27	12.48	6.64	6-15
		New	Same		27	10.52	7.72	6-15
			Different		27	6.26	4.82	2-9
	Male	Old	Same	1	75	8.04	1.82	3-9
			Different		75	12.84	5.81	3-15
		New	Same		75	11.92	7.99	2-15
			Different		75	7.43	3.19	2-9
		Old	Same	2	80	8.23	1.19	4-9
			Different		80	13.49	3.60	5-15
		New	Same		80	12.16	7.45	3-15
			Different		80	7.56	2.45	3-9

Table 5. Penn State Female Model Estimates and Fit Indices

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PFOS1	Penn State	Female	1	Old	Same	2	0.726 (0.027)	0.933 (0.011)		0.323 (0.048)	0.677 (0.048)		0.981	6.5	7.3	6
PFOD1					Different	3	0.000 (0.000)	0.606 (0.020)	0.895 (0.011)	0.011 (0.011)	0.428 (0.051)	0.561 (0.051)	0.962	8.0	8.4	10
						2	0.576 (0.021)	0.887 (0.011)		0.404 (0.051)	0.596 (0.051)		0.885	1.0E+04	25.9	12
PFNS1				New	Same	2	0.600 (0.015)	0.878 (0.017)		0.732 (0.046)	0.268 (0.046)		0.992	3.1	3.2	12
PFND1					Different	2	0.491 (0.027)	0.802 (0.018)		0.401 (0.051)	0.599 (0.051)		1.000	1.5	1.6	6
PFOS2			2	Old	Same	2	0.630 (0.035)	0.933 (0.012)		0.290 (0.054)	0.710 (0.054)		0.983	6.8	9.0	6
PFOD2					Different	2	0.568 (0.025)	0.942 (0.009)		0.376 (0.058)	0.624 (0.058)		0.936	21.1	16.8	12
PFNS2				New	Same	2	0.593 (0.022)	0.901 (0.013)		0.495 (0.060)	0.505 (0.060)		0.990	8.4	8.7	12
PFND2					Different	2	0.594 (0.033)	0.875 (0.016)		0.361 (0.057)	0.639 (0.057)		0.997	4.7	4.9	6

Note: Standard errors are presented in parentheses below each model estimate.

Table 6. Penn State Male Model Estimates and Fit Indices

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PMOS1		Male	1	Old	Same	2	0.807 (0.014)	0.981 (0.004)		0.365 (0.030)	0.635 (0.300)		1.000	0.7	0.9	6
PMOD1					Different	3	0.242 (0.047)	0.723 (0.012)	0.954 (0.004)	0.021 (0.009)	0.362 (0.030)	0.617 (0.030)	0.957	22.8	17.8	10
						2	0.66 (0.014)	0.94 (0.004)		0.30 (0.029)	0.70 (0.029)		0.853	1.2E+05	53.80	12
PMNS1				New	Same	2	0.655 (0.010)	0.927 (0.006)		0.545 (0.031)	0.455 (0.031)		0.952	17.0	16.5	12
PMND1					Different	3	0.336 (0.037)	0.704 (0.014)	0.924 (0.008)	0.071 (0.016)	0.430 (0.031)	0.499 (0.031)	0.995	2.0	2.0	4
						2	0.542 (0.019)	0.885 (0.008)		0.285 (0.028)	0.715 (0.028)		0.939	20.2	10.6	6
PMOS2			2	Old	Same	2	0.540 (0.029)	0.960 (0.005)		0.169 (0.026)	0.831 (0.026)		1.000	18.4	17.7	6
PMOD2					Different	3	0.457 (0.024)	0.869 (0.010)	0.976 (0.004)	0.139 (0.024)	0.367 (0.034)	0.494 (0.035)	1.000	9.5	10.2	10
						2	0.492 (0.023)	0.937 (0.005)		0.163 (0.026)	0.837 (0.026)		0.952	28.3	27.3	12
PMNS2				New	Same	2	0.560 (0.018)	0.907 (0.006)		0.260 (0.031)	0.741 (0.031)		0.970	20.5	23.7	12
PMND2					Different	2	0.687 (0.017)	0.931 (0.008)		0.392 (0.035)	0.608 (0.035)		1.000	2.9	3.4	6

Note: Standard errors are presented in parentheses below each model estimate.

Table 7. Cooper-Union Female Model Estimates and Fit Indices

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CFOS1	Cooper-Union	Female	1	Old	Same	2	0.596 (0.108)	0.872 (0.022)		0.083 (0.052)	0.917 (0.052)		0.999	3.6	3.8	6
CFOD1					Different	2	0.592 (0.046)	0.900 (0.017)		0.270 (0.084)	0.730 (0.084)		0.878	19.6	12.3	12
CFNS1				New	Same	2	0.650 (0.026)	0.868 (0.037)		0.804 (0.075)	0.196 (0.075)		1.000	6.8	7.4	12
CFND1					Different	2	0.562 (0.040)	0.927 (0.028)		0.621 (0.092)	0.380 (0.092)		0.969	1.6	1.8	6
CFOS2			2	Old	Same	2	0.755 (0.050)	0.905 (0.022)		0.300 (0.088)	0.701 (0.088)		0.997	1.6	1.8	6
CFOD2					Different	2	0.602 (0.047)	0.913 (0.016)		0.260 (0.084)	0.740 (0.084)		0.902	4.2	4.1	12
CFNS2				New	Same	2	0.606 (0.029)	0.924 (0.024)		0.700 (0.088)	0.300 (0.088)		0.981	6.4	7.2	12
CFND2					Different	2	0.495 (0.046)	0.887 (0.028)		0.489 (0.096)	0.511 (0.096)		0.970	0.7	1.0	6

Note: Standard errors are presented in parentheses below each model estimate.

Table 8. Cooper-Union Male Model Estimates and Fit Indices

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CMOS1		Male	1	Old	Same	2	0.594 (0.054)	0.935 (0.010)		0.122 (0.038)	0.878 (0.038)		0.963	4.7	5.8	6
CMOD1					Different	3	0.274 (0.081)	0.762 (0.019)	0.965 (0.008)	0.027 (0.018)	0.443 (0.057)	0.530 (0.058)	1.000	4.2	5.3	10
						2	0.709 (0.021)	0.956 (0.008)		0.404 (0.057)	0.596 (0.057)		0.850	559.7	19.1	12
CMNS1				New	Same	3	0.139 (0.089)	0.646 (0.022)	0.926 (0.010)	0.014 (0.013)	0.430 (0.057)	0.556 (0.057)	0.959	7.4	7.9	10
						2	0.612 (0.023)	0.917 (0.011)		0.401 (0.057)	0.599 (0.057)		0.894	190.4	14.9	12
CMND1					Different	2	0.487 (0.047)	0.895 (0.013)		0.170 (0.043)	0.830 (0.043)		0.936	11.5	11.1	6
CMOS2			2	Old	Same	2	0.702 (0.051)	0.941 (0.009)		0.114 (0.035)	0.887 (0.035)		0.942	6.1	6.9	6
CMOD2					Different	3	0.378 (0.113)	0.797 (0.019)	0.970 (0.006)	0.015 (0.014)	0.356 (0.054)	0.629 (0.054)	1.000	6.5	7.6	10
						2	0.759 (0.022)	0.965 (0.006)		0.318 (0.052)	0.682 (0.052)		0.916	79.2	12.8	12
CMNS2				New	Same	3	0.338 (0.070)	0.698 (0.019)	0.953 (0.009)	0.038 (0.021)	0.467 (0.056)	0.495 (0.056)	0.999	5.1	5.6	10
						2	0.654 (0.020)	0.947 (0.009)		0.464 (0.056)	0.536 (0.056)		0.908	76.0	12.5	12
CMND2					Different	2	0.670 (0.029)	0.942 (0.011)		0.374 (0.054)	0.626 (0.054)		1.000	0.4	0.7	6

Note: Standard errors are presented in parentheses below each model estimate.

Table 9. Individual Cell Contributions to Overall Chi-Squared Values.

Outcome	Expected Value	Observed Value	Cell χ^2
0	0.000	0	0.000
1	0.000	0	0.000
2	0.000	0	0.000
3	0.000	0	0.000
4	0.000	0	0.000
5	0.000	0	0.000
6	0.000	0	0.000
7	0.001	0	0.001
8	0.006	0	0.006
9	0.029	1	33.108
10	0.109	0	0.109
11	0.356	2	7.601
12	0.985	4	9.227
13	2.325	2	0.045
14	4.670	6	0.379
15	7.959	5	1.100
16	11.448	7	1.728
17	13.795	7	3.347
18	13.868	9	1.709
19	12.083	22	8.139
20	11.788	16	1.505
21	19.934	26	1.846
22	42.430	39	0.277
23	65.989	47	5.464
24	50.227	65	4.345

Model $\chi^2 = 79.938$, df = 21

Table 10. Male Vs. Female Achievement Test Comparisons.

Variable	Female Mean	Male Mean	T	p	K-S
SATM	590.020 (16.50) 50	593.611 (12.78) 157	-0.172	0.864	p > .05
SATV	490.420 (15.21) 50	478.962 (10.56) 157	0.556	0.579	p > .05
BM	15.400 (0.52) 50	15.548 (0.23) 157	-0.260	0.796	p > .05
M40	16.280 (0.80) 50	16.892 (0.37) 157	-0.767	0.444	p > .05
M110	17.400 (1.07) 50	18.032 (0.59) 157	-0.524	0.601	p > .05
M140	15.660 (1.03) 50	17.474 (0.73) 156	-1.280	0.202	p > .05
CHM	21.250 (1.94) 48	18.193 (0.99) 150	1.485	0.139	p > .05
ENG	29.176 (2.57) 28	29.252 (1.06) 103	-0.030	0.976	p > .05
HSGPA	3.488 (0.09) 50	3.228 (0.07) 157	2.341	0.021	p > .05
COLGPA	2.520 (0.19) 50	2.533 (0.10) 157	1.440	0.152	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 11. Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for Old-Same Items at Time 1.

Variable	Low		T	p	K-S	High		T	p	K-S
	Female Mean	Male Mean				Female Mean	Male Mean			
SATM	601.250 (16.41) 8	559.000 (44.48) 10	0.891	0.391	p > .05	589.049 (19.89) 41	594.245 (13.74) 142	-0.188	0.851	p > .05
SATV	533.750 (32.23) 8	511.000 (38.25) 10	0.440	0.666	p > .05	480.756 (17.24) 41	475.613 (11.31) 142	0.224	0.823	p > .05
BM	16.000 (0.89) 8	14.800 (1.85) 10	0.585	0.569	p > .05	15.268 (0.61) 41	15.592 (0.22) 142	-0.498	0.621	p > .05
M40	17.375 (1.76) 8	16.500 (2.40) 10	0.280	0.783	p > .05	16.195 (0.91) 41	16.993 (0.36) 142	-0.813	0.420	p > .05
M110	17.250 (2.27) 8	16.500 (2.40) 10	0.222	0.827	p > .05	17.585 (1.23) 41	18.127 (0.62) 142	-0.406	0.685	p > .05
M140	14.875 (1.63) 8	18.200 (4.46) 10	-0.700	0.498	p > .05	16.073 (1.19) 41	17.390 (0.73) 141	-0.875	0.383	p > .05
CHM	25.125 (4.49) 8	22.500 (5.40) 8	0.360	0.724	p > .05	20.231 (2.16) 39	17.818 (1.01) 137	1.088	0.278	p > .05
ENG	39.667 (6.77) 3	19.833 (6.83) 6	1.813	0.113	p > .05	27.920 (2.69) 25	29.532 (1.05) 94	-0.657	0.512	p > .05
HSGPA	3.803 (0.08) 8	2.879 (0.39) 10	2.309	0.044	p > .05	3.440 (0.10) 41	3.243 (0.07) 142	4.426	0.156	p > .05
COLGPA	3.119 (0.20) 8	2.220 (0.41) 10	1.945	0.070	p > .05	2.379 (0.23) 41	2.538 (0.10) 142	-0.711	0.478	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 12. Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for Old-Different Items at Time 2.

Variable	Low		T	p	K-S	High		T	p	K-S
	Female Mean	Male Mean				Female Mean	Male Mean			
SATM	592.143 (21.64) 14	607.892 (22.67) 37	-0.503	0.618	p > .05	620.048 (14.75) 21	611.787 (18.81) 75	0.346	0.731	p > .05
SATV	470.000 (17.13) 14	481.946 (19.19) 37	-0.464	0.645	p > .05	530.048 (16.45) 21	494.987 (16.37) 75	1.511	0.136	p > .05
BM	14.786 (1.11) 14	15.243 (0.05) 37	-0.409	0.684	p > .05	16.381 (0.41) 21	16.373 (0.26) 75	0.014	0.989	p > .05
M40	16.071 (1.42) 14	16.973 (0.69) 37	-0.635	0.528	p > .05	18.048 (0.87) 21	17.667 (0.47) 75	0.380	0.705	p > .05
M110	15.429 (1.92) 14	18.946 (1.11) 37	-1.634	0.109	p > .05	20.476 (1.19) 21	19.387 (0.68) 75	0.759	0.450	p > .05
M140	16.143 (1.67) 14	18.838 (1.54) 37	-0.991	0.326	p > .05	17.333 (1.57) 21	19.347 (1.06) 75	-0.929	0.355	p > .05
CHM	20.917 (3.50) 12	19.206 (1.99) 34	0.434	0.667	p > .05	23.286 (2.72) 21	18.514 (1.34) 72	1.659	0.101	p > .05
ENG	24.222 (5.08) 9	27.808 (2.38) 26	-0.716	0.479	p > .05	35.727 (3.34) 11	31.350 (1.64) 50	1.134	0.262	p > .05
HSGPA	3.571 (0.07) 14	3.348 (0.11) 37	1.644	0.107	p > .05	3.700 (0.07) 21	3.313 (0.10) 75	3.121	0.003	p > .05
COLGPA	2.945 (0.12) 14	2.769 (0.16) 37	0.870	0.389	p > .05	2.488 (0.33) 21	2.777 (0.12) 75	-0.824	0.418	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 13. Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for New-Different Items at Time 1.

Variable	Low		T	p	K-S	High		T	p	K-S
	Female Mean	Male Mean				Female Mean	Male Mean			
SATM	554.440 (28.03) 25	566.735 (15.78) 49	-0.413	0.681	p > .05	629.167 (15.04) 24	606.735 (17.78) 102	0.963	0.338	p > .05
SATV	470.840 (23.55) 25	465.510 (10.39) 49	0.207	0.837	p > .05	508.750 (19.69) 24	485.363 (15.31) 102	0.938	0.353	p > .05
BM	14.480 (0.94) 25	14.184 (0.45) 49	0.287	0.776	p > .05	16.333 (0.41) 24	16.294 (0.24) 102	0.074	0.942	p > .05
M40	15.160 (1.35) 25	14.857 (0.56) 49	0.207	0.837	p > .05	17.667 (0.93) 24	18.098 (0.43) 102	-0.443	0.659	p > .05
M110	16.520 (0.17) 25	14.857 (1.00) 49	0.904	0.369	p > .05	18.583 (1.39) 24	19.647 (0.70) 102	-0.669	0.505	p > .05
M140	14.800 (1.28) 25	14.551 (1.08) 49	0.141	0.888	p > .05	17.000 (1.62) 24	18.833 (0.94) 102	-0.877	0.382	p > .05
CHM	23.375 (3.05) 24	14.830 (1.47) 47	2.527	0.016	p < .05*	18.652 (2.45) 23	19.722 (1.29) 97	-0.369	0.713	p > .05
ENG	25.000 (3.85) 12	27.429 (1.41) 35	-0.593	0.563	p > .05	32.313 (3.34) 16	29.769 (1.47) 65	0.750	0.456	p > .05
HSGPA	3.496 (0.16) 25	3.177 (0.10) 49	1.728	0.088	p > .05	3.503 (0.08) 24	3.242 (0.09) 102	2.172	0.033	p > .05
COLGPA	2.932 (0.20) 25	2.541 (0.14) 49	1.614	0.111	p > .05	2.049 (0.32) 24	2.531 (0.13) 102	-1.570	0.119	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 14. Within Component Comparison of Males' and Females' Performance on Achievement Tests Based on Posterior Probabilities for New-Same Items at Time 2.

Variable	Low		T	p	K-S	High		T	p	K-S
	Female Mean	Male Mean				Female Mean	Male Mean			
SATM	592.278 (18.66) 18	603.167 (22.41) 48	-0.373	0.710	p > .05	626.470 (15.62) 17	616.000 (19.36) 64	0.421	0.675	p > .05
SATV	488.389 (18.77) 18	483.375 (18.82) 48	0.189	0.851	p > .05	524.706 (16.94) 17	496.156 (17.09) 64	1.187	0.241	p > .05
BM	14.722 (0.88) 18	15.417 (0.43) 48	-0.782	0.437	p > .05	16.824 (0.39) 17	16.737 (0.29) 64	0.642	0.523	p > .05
M40	16.111 (1.20) 18	16.896 (0.52) 48	-0.702	0.485	p > .05	18.471 (0.91) 17	17.844 (0.56) 64	0.532	0.596	p > .05
M110	16.722 (1.54) 18	19.021 (1.00) 48	-1.218	0.228	p > .05	20.294 (1.53) 17	19.406 (0.70) 64	0.567	0.573	p > .05
M140	17.611 (1.30) 18	18.229 (1.18) 48	-0.295	0.769	p > .05	16.059 (1.94) 17	19.891 (1.24) 64	-1.474	0.145	p > .05
CHM	22.125 (3.23) 16	19.196 (1.83) 46	0.804	0.425	p > .05	22.706 (2.88) 17	18.383 (1.37) 60	1.445	0.153	p < .05*
ENG	27.364 (4.93) 11	30.735 (1.90) 34	-0.776	0.442	p > .05	34.444 (3.37) 9	29.690 (1.93) 42	1.066	0.292	p > .05
HSGPA	3.616 (0.04) 18	3.270 (0.13) 48	2.297	0.025	p > .05	3.682 (0.08) 17	3.366 (0.09) 64	2.609	0.012	p > .05
COLGPA	2.926 (0.20) 18	2.782 (0.14) 48	0.543	0.589	p > .05	2.401 (0.36) 17	2.768 (0.13) 64	-0.963	0.347	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 15. Within Component Comparison of Males' and Females' Performance on Achievement Tests
Based on Posterior Probabilities for Old Items at Time 1.

Variable	Low					Middle					High				
	Female Mean	Male Mean	T	p	K-S	Female Mean	Male Mean	T	p	K-S	Female Mean	Male Mean	T	p	K-S
SATM	528.273 (59.52) 11	463.571 (20.41) 14	0.806	0.428	p > .05	600.000 (13.91) 22	564.044 (28.73) 45	1.126	0.265	p > .05	621.875 (22.84) 16	624.785 (12.94) 93	-0.089	0.929	p > .05
SATV	429.182 (46.54) 11	415.000 (44.43) 14	0.218	0.829	p > .05	518.636 (17.76) 22	447.156 (23.53) 45	2.425	0.018	p > .05	490.625 (22.54) 16	502.312 (11.32) 93	-0.405	0.686	p > .05
BM	13.727 (1.45) 11	12.785 (1.57) 14	0.429	0.672	p > .05	15.682 (0.81) 22	15.178 (0.35) 45	0.574	0.571	p > .05	16.125 (0.58) 16	16.129 (0.24) 93	-0.007	0.995	p > .05
M40	14.455 (2.21) 11	14.214 (2.13) 14	0.078	0.939	p > .05	17.000 (1.16) 22	16.067 (0.56) 45	0.726	0.473	p > .05	16.875 (1.19) 16	17.806 (0.42) 93	-0.827	0.410	p > .05
M110	15.636 (3.01) 11	12.500 (2.50) 14	0.809	0.427	p > .05	17.364 (1.44) 22	16.956 (0.87) 45	0.255	0.800	p > .05	19.063 (1.78) 16	19.366 (0.76) 93	-0.153	0.879	p > .05
M140	15.182 (1.68) 11	14.308 (3.82) 13	0.210	0.837	p > .05	16.318 (1.54) 22	17.400 (1.34) 45	-0.492	0.624	p > .05	15.750 (2.11) 16	17.903 (0.87) 93	-0.949	0.345	p > .05
CHM	21.000 (3.71) 10	14.417 (3.42) 12	1.303	0.208	p > .05	20.095 (3.10) 21	17.698 (1.71) 43	0.734	0.466	p > .05	22.375 (3.59) 16	18.744 (1.31) 90	1.049	0.297	p > .05
ENG	26.667 (6.23) 6	20.600 (4.28) 10	0.829	0.421	p > .05	27.929 (3.74) 14	28.355 (1.76) 31	-0.118	0.907	p > .05	33.250 (4.42) 8	30.678 (1.33) 59	0.651	0.517	p > .05
HSGPA	3.270 (0.34) 11	2.479 (0.33) 14	1.651	0.112	p < .05*	3.602 (0.08) 22	3.136 (0.14) 45	2.905	0.005	p > .05	3.516 (0.11) 16	3.370 (0.07) 93	0.869	0.387	p > .05
COLGPA	2.564 (0.40) 11	2.111 (0.34) 14	0.870	0.393	p > .05	2.863 (0.23) 22	2.263 (0.19) 45	1.894	0.063	p > .05	1.955 (0.41) 16	2.702 (0.12) 93	-1.748	0.098	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test.

SATM = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 16. Within Component Comparison of Males' and Females' Performance on Achievement Tests
Based on Posterior Probabilities for New Items at Time 2.

Variable	Low		T	p	K-S	Middle		T	p	K-S	High		T	p	K-S
	Female Mean	Male Mean				Female Mean	Male Mean				Female Mean	Male Mean			
SATM	598.462 (23.28) 13	597.000 (30.61) 26	0.038	0.970	p > .05	600.111 (29.86) 9	605.525 (25.16) 40	-0.098	0.922	p > .05	625.384 (13.57) 13	622.457 (22.42) 46	0.112	0.911	p > .05
SATV	473.846 (17.08) 13	475.077 (24.91) 26	-0.041	0.968	p > .05	533.444 (31.16) 9	489.525 (21.58) 40	0.915	0.365	p > .05	519.231 (19.53) 13	500.500 (20.16) 46	0.667	0.509	p > .05
BM	14.769 (1.20) 13	17.620 (0.70) 26	-0.148	0.883	p > .05	16.444 (0.65) 9	16.150 (0.35) 40	-0.877	0.385	p > .05	16.923 (0.43) 13	16.457 (0.35) 46	0.675	0.503	p > .05
M40	16.385 (1.56) 13	16.615 (0.84) 26	-0.143	0.887	p > .05	17.222 (1.33) 9	17.725 (0.61) 40	-0.351	0.727	p > .05	18.154 (1.10) 13	17.652 (0.63) 46	0.381	0.705	p > .05
M110	16.692 (2.02) 13	18.077 (1.44) 26	-0.857	0.345	p > .05	20.556 (1.70) 9	19.225 (0.93) 40	0.629	0.532	p > .05	19.769 (1.74) 13	19.913 (0.85) 46	-0.078	0.938	p > .05
M140	17.000 (1.44) 13	18.615 (1.55) 26	-0.665	0.510	p > .05	20.000 (1.66) 24	18.825 (1.69) 40	0.496	0.624	p > .05	14.538 (2.37) 13	19.804 (1.27) 46	1.954	0.056	p > .05
CHM	24.727 (4.07) 11	18.208 (2.52) 24	1.409	0.168	p > .05	17.778 (4.68) 9	18.649 (1.77) 37	-0.205	0.839	p > .05	23.692 (2.64) 13	19.089 (1.73) 45	1.307	0.197	p > .05
ENG	23.286 (6.53) 7	27.952 (2.40) 21	-0.839	0.409	p > .05	33.571 (3.71) 7	30.200 (2.80) 25	0.593	0.557	p > .05	35.500 (5.14) 6	31.667 (1.91) 30	0.794	0.433	p > .05
HSGPA	3.595 (0.08) 13	3.277 (0.15) 26	1.845	0.074	p < .05*	3.751 (0.09) 9	3.267 (0.16) 40	2.682	0.010	p < .05*	3.329 (0.10) 13	3.402 (0.10) 46	1.628	0.112	p < .05*
COLGPA	3.107 (0.14) 13	2.642 (0.22) 26	1.814	0.078	p > .05	2.481 (0.49) 9	2.894 (0.13) 40	-0.812	0.438	p > .05	2.366 (0.41) 13	2.745 (0.16) 46	-1.015	0.314	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 17. Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for Old-different Items at Time 1.

Variable	Low Mean	High Mean	T	p	K-S
SATM	555.781 (19.52) 73	615.921 (12.26) 126	-2.745	0.007	p < .05*
SATV	463.589 (16.25) 73	492.905 (10.71) 126	-1.565	0.119	p > .05
BM	14.863 (0.41) 73	16.063 (0.21) 126	-2.580	0.011	p > .05
M40	15.685 (0.61) 73	17.690 (0.38) 126	-2.799	0.006	p > .05
M110	16.082 (0.85) 73	19.159 (0.64) 126	-2.894	0.004	p > .05
M140	15.521 (1.01) 73	17.944 (0.77) 126	-1.907	0.058	p > .05
CHM	17.507 (1.39) 71	19.642 (1.18) 120	-1.440	0.254	p > .05
ENG	28.857 (1.52) 49	29.462 (1.30) 78	-0.297	0.767	p > .05
HSGPA	3.186 (0.11) 73	3.351 (0.06) 126	-1.308	0.193	p < .05*
COLGPA	2.431 (0.15) 73	2.579 (0.11) 126	-0.808	0.420	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 18. Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for New-same Items at Time 1.

Variable	Low Mean	High Mean	T	p	K-S
SATM	567.492 (13.06) 126	632.453 (17.75) 75	-2.983	0.003	p < .05*
SATV	462.651 (10.82) 126	511.120 (15.41) 75	-2.636	0.009	p < .05*
BM	14.960 (0.29) 126	16.413 (0.32) 75	-3.242	0.001	p < .05*
M40	15.976 (0.43) 126	18.240 (0.53) 75	-3.269	0.001	p < .05*
M110	16.421 (0.65) 126	20.387 (0.80) 75	-3.774	0.001	p < .05*
M140	16.008 (0.75) 126	18.851 (1.03) 74	-2.260	0.025	p > .05
CHM	17.417 (1.09) 120	21.125 (1.54) 72	-20.136	0.046	p > .05
ENG	28.386 (1.22) 83	30.133 (1.77) 45	-0.828	0.409	p > .05
HSGPA	3.245 (0.08) 126	3.359 (0.08) 75	-1.027	0.306	p < .05*
COLGPA	2.436 (0.11) 126	2.643 (0.15) 75	-1.137	0.257	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 19. Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for Old-same Items at Time 2.

Variable	Low Mean	High Mean	T	p	K-S
SATM	619.438 (23.05) 32	607.522 (13.25) 115	0.427	0.670	p > .05
SATV	486.000 (20.58) 32	496.652 (11.60) 115	-0.434	0.665	p > .05
BM	15.844 (0.39) 32	15.965 (0.27) 115	-0.519	0.827	p > .05
M40	17.875 (0.61) 32	17.261 (0.41) 115	0.728	0.468	p > .05
M110	19.219 (1.20) 32	19.008 (0.57) 115	0.167	0.867	p > .05
M140	19.188 (1.30) 32	18.470 (0.85) 115	0.411	0.682	p > .05
CHM	19.667 (2.43) 30	19.596 (1.07) 109	0.029	0.947	p > .05
ENG	28.391 (2.30) 23	30.822 (1.48) 73	-0.829	0.409	p > .05
HSGPA	3.357 (0.13) 32	3.414 (0.07) 115	-0.390	0.697	p < .05*
COLGPA	2.783 (0.18) 32	2.741 (0.10) 115	0.196	0.843	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 20. Comparison of High- and Low-Performing Components' Achievement Test Scores Based on Posterior Probabilities for New-different Items at Time 2.

Variable	Low Mean	High Mean	T	p	K-S
SATM	606.152 (19.28) 46	611.921 (14.29) 101	-0.232	0.817	p > .05
SATV	487.239 (16.13) 46	497.654 (12.76) 101	-0.473	0.637	p > .05
BM	15.087 (0.54) 46	16.327 (0.21) 101	-2.122	0.038	p > .05
M40	16.804 (0.71) 46	17.663 (0.39) 101	-1.147	0.253	p > .05
M110	17.696 (1.01) 46	19.673 (0.59) 101	-1.789	0.076	p > .05
M140	18.435 (1.34) 46	18.713 (0.85) 101	-0.179	0.858	p > .05
CHM	19.073 (1.80) 41	19.837 (1.19) 98	-0.352	0.726	p > .05
ENG	27.364 (2.29) 33	31.746 (1.46) 63	-1.682	0.096	p > .05
HSGPA	3.371 (0.09) 46	3.416 (0.08) 101	-0.337	0.737	p < .05*
COLGPA	2.831 (0.13) 46	2.712 (0.11) 101	0.633	0.528	p > .05

Note: For Each Variable, Standard Errors (in Parentheses) and Sample Sizes are Presented Below Mean Values. The K-S column provides p-values for the Kolmogorov-Smirnov test. **SATM** = Scholastic Achievement Test - Math Portion, **SATV** = Scholastic Achievement Test - Verbal Portion, **BM** = Penn State Placement Test for Basic Math Skills, **M40-M140** = Penn State Placement Tests of Increasing Difficulty for Math Skills, **CHM** = Penn State Placement Test for Chemistry, **ENG** = Penn State Placement Test For English Skills, **HSGPA** = Cumulative High School Grade Point Average, **COLGPA** = Cumulative Penn State Grade Point Average.

Table 21. Joint Two Component Frequency Tables and Model Estimates for All Subjects on Same and Different, Old and New Items at Times 1 and 2.

A

Old Same Items
Time 2

Old
Same
Items
Time 1

	"Low"	"High"
"Low"	0.059 0.033 (0.014) 0.506 0.300 (0.127) 9	0.058 0.076 (0.014) 0.494 0.700 (0.127) 21
"High"	0.119 0.116 (0.014) 0.137 0.131 (0.016) 32	0.763 0.775 (0.014) 0.863 0.869 (0.016) 213

Raw Score Correlation
r = 0.205 (p <.001) ^A

Component Score Correlation
r = 0.148 (p = .014) ^A

Model Correlation
 $\hat{\rho}$ = 0.102 ^A

χ^2 (78) = 95.3 , p = n.s.
 L^2 (78) = 100.0 , p = n.s.

C

New Same Items
Time 1

Old
Same
Items
Time 1

	"Low"	"High"
"Low"	0.16 0.104 (0.014) 1.000 0.922 (0.107) 47	0.000 0.009 (0.014) 0.000 0.078 (0.107) 4
"High"	0.431 0.500 (0.014) 0.513 0.563 (0.016) 227	0.408 0.388 (0.014) 0.487 0.437 (0.016) 176

Raw Score Correlation
r = 0.450 (p <.001) ^A

Component Score Correlation
r = 0.231 (p = .001) ^B

Model Correlation
 $\hat{\rho}$ = 0.404 ^A

χ^2 (132) = 130.2 , p = n.s.
 L^2 (132) = 157.7 , p = n.s.

B

Old Different Items
Time 1

Old
Same
Items
Time 1

	"Low"	"High"
"Low"	0.152 0.084 (0.011) 1.000 0.745 (0.081) 38	0.005 0.029 (0.011) 0.000 0.255 (0.081) 13
"High"	0.177 0.296 (0.011) 0.218 0.334 (0.012) 134	0.666 0.591 (0.011) 0.782 0.666 (0.012) 267

Raw Score Correlation
r = 0.367 (p <.001) ^A

Component Score Correlation
r = 0.268 (p = .001) ^A

Model Correlation
 $\hat{\rho}$ = 0.395 ^A

χ^2 (132) = 193.0 , p < .05
 L^2 (132) = 182.2 , p < .05

D

Old Different Items
Time 2

Old
Same
Items
Time 2

	"Low"	"High"
"Low"	0.180 0.117 (0.006) 1.000 0.830 (0.036) 44	0.002 0.024 (0.006) 0.000 0.170 (0.036) 9
"High"	0.042 0.180 (0.006) 0.056 0.210 (0.008) 68	0.776 0.679 (0.006) 0.944 0.790 (0.008) 256

Raw Score Correlation
r = 0.641 (p <.001) ^A

Component Score Correlation
r = 0.405 (p = .001) ^B

Model Correlation
 $\hat{\rho}$ = 0.801 ^C

χ^2 (132) = 375.8 , p < .05
 L^2 (132) = 210.5 , p < .05

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 21 Cont'd.

		New Same Items Time 2				
		"Low"		"High"		
Old Same Items Time 2	"Low"	0.207 (0.017)	0.135	0.000 (0.017)	0.005	Raw Score Correlation r = 0.662 (p <.001) ^A
		0.630 (0.029)	0.962	0.370 (0.029)	0.038	
		51		2	Component Score Correlation r = 0.406 (p = .001) ^B	
	0.207 (0.017)	0.326	0.586 (0.017)	0.533		
	"High"	0.162 (0.034)	0.380	0.838 (0.034)	0.620	Model Correlation $\hat{\rho} = 0.711$ ^A
		123		201		

$$\chi^2 (78) = 95.3, p < .05$$

$$L^2 (132) = 207.2, p < .05$$

		New Different Items				
		Time 1				
Old Diff. Items Time 1	"Low"	"Low"		"High"		Raw Score Correlation r = 0.641 (p <.001) ^A
		0.320 (0.010)	0.232	0.005 (0.010)	0.148	
		0.862 (0.030)	0.610	0.138 (0.030)	0.390	
		105		67		
	"High"					Component Score Correlation r = 0.405 (p = .001) ^B
		0.170 (0.010)	0.108	0.505 (0.010)	0.511	
		0.034 (0.023)	0.175	0.966 (0.023)	0.825	
		49		231		
Model Correlation $\hat{\rho} = 0.801$ ^C						

$$\chi^2 (132) = 513.7, p < .05$$

$$L^2 (132) = 250.8, p < .05$$

		Old Different Items Time 2				
		"Low"		"High"		
Old Diff. Items Time 1	"Low"	0.139 (0.016)	0.153	0.150 (0.016)	0.201	Raw Score Correlation r = 0.268 (p < .001) ^A
		0.829 (0.002)	0.433	0.171 (0.002)	0.567	
		42		55		
	"High"	0.104 (0.016)	0.146	0.607 (0.016)	0.500	Component Score Correlation r = 0.216 (p = .001) ^A
		0.022 (0.003)	0.226	0.978 (0.003)	0.774	
		40		137		
					Model Correlation $\hat{\rho}$ = 0.169 ^A	

$$\chi^2 (222) = 211.1, p = \text{n.s.}$$

$$L^2 (222) = 219.1, p = \text{n.s.}$$

		New Different Items Time 2				
		"Low"		"High"		
Old Diff. Items Time 2	"Low"	0.246 (0.015)	0.204	0.000 (0.015)	0.093	Raw Score Correlation r = 0.623 (p < .001) ^A
		0.696 (0.041)	0.688	0.304 (0.041)	0.313	
		77		35	Component Score Correlation r = 0.593 (p = .001) ^A	
		0.105 (0.015)	0.074	0.649 (0.015)		0.629
	"High"	0.134 (0.039)	0.106	0.866 (0.039)	0.894	Model Correlation $\hat{\rho} = 0.884$ ^B
			28		237	

$$\chi^2 (132) = 192.5, p < .05$$

$$L^2 (132) = 168.6, p < .05$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 21 Cont'd.

		New Same Items Time 2				
		"Low"		"High"		
New Same Items Time 1	"Low"	0.335 (0.016)	0.324	0.196 (0.016)	0.225	Raw Score Correlation r = 0.401 (p <.001) ^A
		1.000 (0.101)	0.589	0.000 (0.101)	0.411	
		89		62	Component Score Correlation r = 0.357 (p = .001) ^A	
	"High"	0.078 (0.016)	0.105	0.391 (0.016)		0.345
		0.271 (0.021)	0.234	0.729 (0.021)		0.766
			29			95
Model Correlation $\hat{\rho} = 0.257$ ^A						

$$\chi^2 (78) = 95.3, p = n.s.$$

$$L^2 (222) = 175.4, p = n.s.$$

		New Different Items Time 2				
		"Low"		"High"		
New Same Items Time 2	"Low"	0.372 (0.013)	0.231	0.059 (0.013)	0.231	Raw Score Correlation r = 0.621 (p <.001) ^A
		0.984 (0.031)	0.500	0.016 (0.031)	0.500	
		87		87	Component Score Correlation	
	"High"	0.020 (0.013)	0.048	0.549 (0.013)	0.491	r = 0.457 (p = .001) ^A
		0.252 (0.015)	0.089	0.748 (0.015)	0.911	
			18		185	Model Correlation $\hat{\rho} = 0.542$ ^A

$$\chi^2 (132) = 117.6, p = n.s.$$

$$L^2 (132) = 130.3, p = n.s.$$

<div>J</div>		New Different Items				
		Time 1				
New Same Items Time 1	"Low"	"Low"		"High"		Raw Score Correlation r = 0.471 (p < .001) ^{A B}
		0.505	0.291	0.111	0.313	
		(0.001)		(0.001)		
		0.480	0.482	0.520	0.518	
	(0.055)		(0.055)			
	132		142		Component Score Correlation r = 0.364 (p = .001) ^A	
	0.000	0.051	0.384	0.344		
	(0.001)		(0.001)			
0.147	0.128	0.853	0.872			
(0.022)		(0.022)				
23		156		Model Correlation $\hat{\rho} = 0.589$ ^B		

$$\chi^2 (132) = 169.2, p < .05$$

$$L^2 (132) = 184.1, p < .05$$

L

New Different Items

Time 2

		"Low"		"High"	
New Diff. Items Time 1	"Low"	0.332	0.164	0.141	0.145
		(0.020)		(0.020)	
		1.000	0.529	0.000	0.471
		(0.064)		(0.064)	
		45		40	Raw Score Correlation r = 0.427 (p < .001) ^A
	"High"	0.076	0.127	0.451	0.564
		(0.020)		(0.020)	
0.143		0.184	0.857	0.816	
(0.020)			(0.020)		
	35		155	Component Score Correlation r = 0.351 (p = .001) ^A	
				Model Correlation $\hat{\rho}$ = 0.288 ^A	

$$\chi^2 (78) = 111.8, p < .05$$

$$L^2 (78) = 120.4, p < .05$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 22. Joint Two Component Frequency Tables and Model Estimates for Males on Same and Different, Old and New Items at Times 1 and 2.

A		B	
Old Same Items Time 2		Old Different Items Time 1	
Old Same Items Time 1	"Low"	"High"	
	0.026 0.020 (0.014)	0.053 0.064 (0.014)	Raw Score Correlation $r = 0.113$ ($p < .059$) ^A
	0.316 0.235 (0.190)	0.684 0.765 (0.190)	
	4	13	Component Score Correlation $r = 0.092$ ($p = .193$) ^A
Old Same Items Time 1	"Low"	"High"	
	0.128 0.113 (0.014)	0.792 0.804 (0.014)	Model Correlation $\hat{\rho} = 0.040$ ^A
	0.141 0.123 (0.015)	0.859 0.877 (0.015)	
	23	164	
χ^2 (78) = 95.3, $p < .05$		χ^2 (132) = 113.6, $p = \text{n.s.}$	
L^2 (78) = 115.2, $p < .05$		L^2 (132) = 161.6, $p < .05$	
C		D	
New Same Items Time 1		Old Different Items Time 2	
Old Same Items Time 1	"Low"	"High"	
	0.124 0.084 (0.017)	0.000 0.009 (0.017)	Raw Score Correlation $r = 0.426$ ($p < .001$) ^A
	1.000 0.903 (0.169)	0.000 0.097 (0.169)	
	28	3	Component Score Correlation $r = 0.237$ ($p = .001$) ^B
Old Same Items Time 1	"Low"	"High"	
	0.386 0.450 (0.017)	0.490 0.456 (0.017)	Model Correlation $\hat{\rho} = 0.366$ ^{A B}
	0.442 0.497 (0.190)	0.558 0.503 (0.190)	
	150	152	
χ^2 (132) = 90.1, $p = \text{n.s.}$		χ^2 (132) = 514.2, $p < .05$	
L^2 (132) = 128.5, $p = \text{n.s.}$		L^2 (132) = 206.8, $p < .05$	
Old Same Items Time 2	"Low"	"High"	
	0.148 0.100 (0.007)	0.004 0.021 (0.007)	Raw Score Correlation $r = 0.772$ ($p < .001$) ^A
	1.000 0.824 (0.045)	0.000 0.176 (0.045)	
	28	6	Component Score Correlation $r = 0.477$ ($p = .001$) ^B
Old Same Items Time 2	"Low"	"High"	
	0.021 0.161 (0.007)	0.828 0.718 (0.007)	Model Correlation $\hat{\rho} = 0.787$ ^A
	0.027 0.183 (0.008)	0.973 0.817 (0.008)	
	45	201	

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 22. Cont'd.

E		New Same Items Time 2				
		"Low"		"High"		
Old Same Items Time 2	"Low"	0.170 (0.018)	0.114	0.000 (0.018)	0.007	Raw Score Correlation r = 0.662 (p <.001) ^A
		0.553 (0.043)	0.941	0.447 (0.043)	0.059	
		32		2	Component Score Correlation r = 0.398 (p = .001) ^B	
	"High"	0.184 (0.018)	0.300	0.646 (0.018)	0.579	Model Correlation $\hat{\rho}$ = 0.638 ^A
		0.186 (0.035)	0.341	0.814 (0.035)	0.659	
		84		162		

$$\chi^2 (78) = 95.3, p < .05$$

$$L^2 (132) = 202.4, p < .05$$

		New Different Items			
		Time 1			
		"Low"		"High"	
		0.306 (0.001)	0.196	0.001 (0.001)	0.142
Old Diff. Items Time 1	"Low"	0.847 (0.039)	0.580	0.153 (0.039)	0.420
		65		47	
	"High"	0.000 (0.001)	0.084	0.693 (0.001)	0.578
		0.026 (0.023)	0.127	0.974 (0.023)	0.873
		28		192	

Raw Score Correlation
r = 0.653 (p <.001)^A

Component Score Correlation
r = 0.477 (p = .001)^B

Model Correlation
 $\hat{\rho} = 1.000$ ^C

$$\chi^2 (132) = 344.1, p < .05$$

$$L^2 (132) = 207.8, p < .05$$

		Old Different Items Time 2				
		"Low"		"High"		
Old Diff. Items Time 1	"Low"	0.091 (0.017)	0.118	0.126 (0.017)	0.176	Raw Score Correlation r = 0.165 (p <.001) ^A
		0.783 (0.002)	0.400	0.217 (0.002)	0.600	
		24		36		
	"High"	0.106 (0.017)	0.147	0.676 (0.017)	0.559	Component Score Correlation r = 0.198 (p = .005) ^A
		0.014 (0.002)	0.208	0.986 (0.002)	0.792	
		30		114		

Model Correlation
 $\hat{\rho} = 0.101$ ^A

$$\chi^2 (222) = 139.1, p = \text{n.s.}$$

$$L^2 (222) = 165.6, p = \text{n.s.}$$

		New Different Items Time 2				
		"Low"		"High"		
Old Diff. Items Time 2	"Low"	0.193 (0.017)	0.171	0.000 (0.017)	0.089	Raw Score Correlation r = 0.623 (p <.001) ^A
		0.656 (0.061)	0.658	0.344 (0.061)	0.342	
		48		25		
	"High"	0.010 (0.017)	0.075	0.707 (0.017)	0.664	Component Score Correlation r = 0.593 (p = .001) ^A
		0.161 (0.039)	0.101	0.839 (0.039)	0.899	
		21		186		

Model Correlation
 $\hat{\rho} = 0.884$ ^B

$$\chi^2 (132) = 219.3, p < .05$$

$$L^2 (132) = 156.6, p = \text{n.s.}$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 22. Cont'd.

		New Same Items Time 2				
		"Low"		"High"		
New Same Items Time 1	"Low"	0.244 (0.019)	0.250	0.197 (0.019)	0.221	Raw Score Correlation r = 0.336 (p <.001) ^A
		1.000 (0.122)	0.531	0.000 (0.122)	0.469	
		51		45	Component Score Correlation r = 0.289 (p = .001) ^A	
	"High"	0.106 (0.019)	0.132	0.454 (0.019)		0.397
		0.229 (0.021)	0.250	0.771 (0.021)	0.750	Model Correlation $\hat{\rho}$ = 0.210 ^A
			27		81	

$$\chi^2 (78) = 95.3, p = n.s.$$

$$L^2 (222) = 153.8, p = n.s.$$

		New Different Items Time 2				
		"Low"		"High"		
New Same Items Time 2	"Low"	0.308 (0.014)	0.204	0.053 (0.014)	0.211	Raw Score Correlation r = 0.649 (p <.001) ^A
		1.000 (0.005)	0.491	0.000 (0.005)	0.509	
		57		59	Component Score Correlation r = 0.478 (p = .001) ^B	
	"High"	0.020 (0.014)	0.043	0.619 (0.014)		0.543
		0.025 (0.002)	0.073	0.975 (0.002)		0.927
			12			152

$$\chi^2 (132) = 117.5, p = n.s.$$

$$L^2 (132) = 131.0, p = n.s.$$

J

		New Different Items				
		Time 1				
		"Low"		"High"		
New Same Items Time 1	"Low"	0.410 (0.001)	0.235	0.118 (0.001)	0.301	Raw Score Correlation $r = 0.468$ ($p < .001$) ^{A B}
		0.417 (0.082)	0.438	0.583 (0.082)	0.562	
		78		100	Component Score Correlation $r = 0.378$ ($p = .001$) ^B	
	"High"	0.000 (0.001)	0.045	0.472 (0.001)	0.419	Model Correlation $\hat{\rho} = 0.612$ ^A
		0.137 (0.022)	0.097	0.863 (0.022)	0.903	
			15		139	

$$\chi^2 (132) = 162.4, p < .05$$

$$L^2 (132) = 171.4, p < .05$$

<div>L</div> <

$$\chi^2 (78) = 72.7, p = n.s.$$

$$L^2 (78) = 92.1, p = n.s.$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 23. Joint Two Component Frequency Tables and Model Estimates for Females on Same and Different, Old and New Items at Times 1 and 2.

A

Old Same Items
Time 2

B

Old Different Items
Time 1

Old
Same
Items
Time 1

Old
Same
Items
Time 1

"Low"	"High"
0.152 0.070 (0.037)	0.083 0.113 (0.037)
0.644 0.385 (0.153)	0.356 0.615 (0.153)
5	8
0.098 0.127 (0.037)	0.667 0.690 (0.037)
0.126 0.155 (0.048)	0.874 0.845 (0.048)
9	49

Raw Score Correlation
 $r = 0.350$ ($p < .003$)^A

Component Score Correlation
 $r = 0.223$ ($p = .062$)^A

Model Correlation
 $\hat{\rho} = 0.291$ ^A

"Low"	"High"
0.238 0.125 (0.026)	0.025 0.042 (0.026)
1.000 0.750 (0.110)	0.000 0.250 (0.110)
15	5
0.256 0.375 (0.026)	0.481 0.458 (0.026)
0.352 0.450 (0.034)	0.648 0.550 (0.034)
45	55

Raw Score Correlation
 $r = 0.250$ ($p < .006$)^A

Component Score Correlation
 $r = 0.224$ ($p = .014$)^A

Model Correlation
 $\hat{\rho} = 0.155$ ^A

χ^2 (78) = 95.3 , $p < .05$
 L^2 (78) = 115.2 , $p < .05$

χ^2 (132) = 113.6 , $p = \text{n.s.}$
 L^2 (132) = 161.6 , $p < .05$

C

New Same Items
Time 1

D

Old Different Items
Time 2

Old
Same
Items
Time 1

Old
Same
Items
Time 2

"Low"	"High"
0.260 0.157 (0.031)	0.000 0.008 (0.031)
1.000 0.950 (0.134)	0.000 0.050 (0.134)
19	1
0.560 0.636 (0.031)	0.180 0.198 (0.031)
0.748 0.762 (0.041)	0.252 0.238 (0.041)
77	24

Raw Score Correlation
 $r = 0.440$ ($p < .001$)^A

Component Score Correlation
 $r = 0.172$ ($p = .059$)^B

Model Correlation
 $\hat{\rho} = 0.366$ ^{A B}

"Low"	"High"
0.265 0.165 (0.024)	0.001 0.031 (0.024)
1.000 0.842 (0.089)	0.000 0.158 (0.089)
16	3
0.099 0.237 (0.024)	0.635 0.567 (0.024)
0.136 0.295 (0.032)	0.864 0.705 (0.032)
23	55

Raw Score Correlation
 $r = 0.648$ ($p < .001$)^A

Component Score Correlation
 $r = 0.443$ ($p = .001$)^A

Model Correlation
 $\hat{\rho} = 0.616$ ^A

χ^2 (132) = 90.1 , $p = \text{n.s.}$
 L^2 (132) = 128.5 , $p = \text{n.s.}$

χ^2 (132) = 514.2 , $p < .05$
 L^2 (132) = 206.8 , $p < .05$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 23. Cont'd.

		New Same Items Time 2				
		"Low"		"High"		
Old Same Items Time 2	"Low"	0.313 (0.046)	0.196	0.000 (0.046)	0.000	Raw Score Correlation r = 0.641 (p < .001) ^{A B}
		0.752 (0.051)	1.000	0.248 (0.051)	0.000	
		19		0	Component Score Correlation r = 0.405 (p = .001) ^A	
	"High"	0.272 (0.046)	0.402	0.415 (0.046)		0.402
		0.000 (0.208)	0.500	1.000 (0.208)	0.500	Model Correlation $\hat{\rho} = 0.801$ ^B
		39		39		

$$\chi^2 (78) = 95.3, p < .05$$

$$L^2 (132) = 202.4, p < .05$$

		New Different Items				
		Time 1				
Old Diff. Items Time 1	"Low"	"Low"		"High"		Raw Score Correlation r = 0.561 (p <.001) ^A
		0.475 (0.019)	0.333	0.011 (0.019)	0.167	
		0.897 (0.054)	0.667	0.103 (0.054)	0.333	
	"High"	40		20		Component Score Correlation r = 0.317 (p = .001) ^A
		0.276 (0.019)	0.175	0.238 (0.019)	0.325	
		0.043 (0.087)	0.350	0.957 (0.087)	0.650	
	21		39		Model Correlation $\hat{\rho} = 0.448$ ^A	

$$\chi^2 (132) = 344.7, p < .05$$

$$L^2 (132) = 201.8, p < .05$$

		Old Different Items Time 2				
		"Low"		"High"		
Old Diff. Items Time 1	"Low"	0.276 (0.035)	0.257	0.231 (0.035)	0.271	Raw Score Correlation r = 0.360 (p <.002) ^A
		0.884 (0.030)	0.486	0.116 (0.030)	0.514	
		18		19	Component Score Correlation r = 0.187 (p = .121) ^A	
	"High"	0.096 (0.035)	0.143	0.397 (0.035)		0.329
		0.197 (0.146)	0.303	0.803 (0.146)	0.697	Model Correlation $\hat{\rho} = 0.205$ ^A
		10		23		

$$\chi^2 (222) = 139.1, p = \text{n.s.}$$

$$L^2 (222) = 165.6, p < .05$$

		New Different Items Time 2				
		"Low"		"High"		
Old Diff. Items Time 2	"Low"	0.395 (0.038)	0.299	0.000 (0.038)	0.103	Raw Score Correlation r = 0.610 (p < .001) ^A
		0.746 (0.063)	0.744	0.254 (0.063)	0.256	
		29		10		
	"High"	0.120 (0.038)	0.072	0.485 (0.038)	0.526	Component Score Correlation r = 0.632 (p = .001) ^A
		0.000 (0.209)	0.121	1.000 (0.209)	0.879	
		7		51		

Model Correlation
 $\hat{\rho} = 0.930$ ^B

$$\chi^2 (132) = 219.3, p < .05$$

$$L^2 (132) = 156.6, p = \text{n.s.}$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 23. Cont'd.

		New Same Items Time 2				
		"Low"		"High"		
New Same Items Time 1	"Low"	0.598 (0.041)	0.535	0.193 (0.041)	0.239	Raw Score Correlation r = 0.401 (p <.001) ^A
		1.000 (0.175)	0.691	0.000 (0.175)	0.309	
		38		17		
	"High"	0.000 (0.041)	0.028	0.209 (0.041)	0.197	Component Score Correlation r = 0.357 (p = .001) ^A
		0.403 (0.063)	0.125	0.597 (0.063)	0.875	
		2		14		
						Model Correlation $\hat{\rho} = 0.257$ ^A

$$\chi^2 (78) = 95.3, p = n.s.$$

$$L^2 (222) = 165.2, p = n.s.$$

		New Different Items Time 2				
		"Low"		"High"		
New Same Items Time 2	"Low"	0.564 (0.033)	0.309	0.074 (0.033)	0.289	Raw Score Correlation r = 0.517 (p <.001) A
		0.980 (0.040)	0.517	0.020 (0.040)	0.483	
		30		28		
	"High"	0.006 (0.033)	0.062	0.356 (0.033)	0.340	Component Score Correlation r = 0.369 (p = .001) A
		0.537 (0.037)	0.154	0.463 (0.037)	0.846	
		6		33		

$$\chi^2 (132) = 117.4, p = n.s.$$

$$L^2 (132) = 131.0, p = n.s.$$

J

		New Different Items Time 1				
		"Low"		"High"		
New Same Items Time 1	"Low"	0.74 (0.025)	0.446	0.102 (0.025)	0.347	Raw Score Correlation $r = 0.354$ (p < .001) ^A
		0.544 (0.069)	0.563	0.456 (0.069)	0.438	
		54		42		
	"High"	0.027 (0.025)	0.066	0.131 (0.025)	0.140	Component Score Correlation $r = 0.196$ (p = .031) ^A
		0.196 (0.070)	0.320	0.804 (0.070)	0.680	
		8		17		
						Model Correlation $\hat{\rho} = 0.351$ ^A

$$\chi^2 (132) = 163.0, p < .05$$

$$L^2 (132) = 170.9, p < .05$$

<div>L</div> 	
--	--

$$\chi^2 (78) = 72.7, p = n.s.$$

$$L^2 (78) = 92.1, p = n.s.$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 24. Joint Two Component Frequency Tables and Model Estimates for Penn State Subjects on Same and Different, Old and New Items at Times 1 and 2.

A	Old Same Items				B	Old Different Items				
	Time 2					Time 1				
	"Low"		"High"			"Low"		"High"		
	0.048	0.026	0.060	0.083		0.15	0.089	0.002	0.026	
	(0.016)		(0.016)			(0.012)		(0.012)		
	0.440	0.240	0.560	0.760		1.000	0.775	0.000	0.225	
	(0.153)		(0.153)			(0.096)		(0.096)		
	6		19			31		9		
	0.148	0.140	0.744	0.750		0.184	0.295	0.664	0.590	
	(0.016)		(0.016)			(0.012)		(0.012)		
Old Same Items Time 1	"Low"	0.167	0.158	0.833	0.842	"High"	0.225	0.333	0.775	0.667
		(0.018)		(0.018)			(0.014)		(0.014)	
	32		171		103		206			
	Raw Score Correlation r = 0.167 (p < .012) ^A				Raw Score Correlation r = 0.374 (p < .001) ^A					
	Component Score Correlation r = 0.069 (p = .299) ^A				Component Score Correlation r = 0.289 (p = .001) ^A					
	Model Correlation $\hat{\rho}$ = 0.071 ^A				Model Correlation $\hat{\rho}$ = 0.438 ^A					
	χ^2 (78) = 94.7, p = n.s.				χ^2 (132) = 132, p = n.s.					
	L^2 (78) = 96.3, p = n.s.				L^2 (132) = 157, p = n.s.					
	C	New Same Items				D	Old Different Items			
		Time 1					Time 2			
"Low"		"High"		"Low"			"High"			
0.154		0.105	0.000	0.009	0.196		0.141	0.000	0.022	
(0.015)			(0.015)		(0.013)			(0.013)		
1.000		0.925	0.000	0.075	1.000		0.864	0.000	0.136	
(0.120)			(0.120)		(0.070)			(0.070)		
37			3		38			6		
0.450		0.507	0.396	0.379	0.035		0.159	0.769	0.678	
(0.015)			(0.015)		(0.013)			(0.013)		
Old Same Items Time 1	"Low"	0.531	0.572	0.469	0.428	"High"	0.045	0.190	0.955	0.810
		(0.017)		(0.017)			(0.017)		(0.017)	
	178		133		43		183			
	Raw Score Correlation r = 0.426 (p < .001) ^A				Raw Score Correlation r = 0.792 (p < .001) ^A					
	Component Score Correlation r = 0.230 (p = .001) ^B				Component Score Correlation r = 0.543 (p = .001) ^B					
	Model Correlation $\hat{\rho}$ = 0.401 ^A				Model Correlation $\hat{\rho}$ = 0.790 ^A					
	χ^2 (132) = 115, p = n.s.				χ^2 (132) = 429, p = n.s.					
	L^2 (132) = 133, p = n.s.				L^2 (132) = 200, p = n.s.					

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 24 Cont'd.

E

New Same Items
Time 2

Old Same Items Time 2	"Low"	"Low"		"High"		Raw Score Correlation r = 0.205 (p <.001) ^A
		0.217 (0.021)	0.033	0.000 (0.021)	0.076	
	"High"	1.000 (0.107)	0.300	0.000 (0.107)	0.700	Component Score Correlation r = 0.148 (p = .014) ^B
		9		21		
	"Low"	0.160 (0.021)	0.116	0.623 (0.021)	0.775	Model Correlation $\hat{\rho} = 0.795$ ^A
		0.213 (0.026)	0.131	0.787 (0.026)	0.869	
	32		213			

χ^2 (132) = 307 , p = n.s.
 L^2 (132) = 187 , p = n.s.

		New Different Items				
		Time 1				
Old Diff. Items Time 1	"Low"	"Low"		"High"		Raw Score Correlation r = 0.205 (p <.001) ^A
		0.328 (0.011)	0.033	0.007 (0.011)	0.076	
	0.985 (0.034)	0.300	0.015 (0.034)	0.700	Component Score Correlation r = 0.148 (p = .014) ^B	
	9		21			
	"High"	0.182 (0.011)	0.116	0.483 (0.011)	0.775	Model Correlation $\hat{\rho} = 0.682$ ^A
		0.274 (0.017)	0.131	0.726 (0.017)	0.869	
		32		213		

χ^2 (132) = 248 , p = n.s.
 I^2 (132) = 196 , p = n.s.

F

Old Different Items
Time 2

Old Diff. Items Time 1		"Low"		"High"		Raw Score Correlation $r = 0.205$ ($p < .001$) ^A
		0.126 0.033 (0.018)	0.159 0.076 (0.018)			
		0.443 0.300 (0.062)	0.557 0.700 (0.062)			
		9	21			
Old Diff. Items Time 1		"Low"		"High"		Component Score Correlation $r = 0.148$ ($p = .014$) ^A Model Correlation $\hat{\rho} = 0.158$ ^A
		0.12 0.116 (0.018)	0.594 0.775 (0.018)			
		0.169 0.131 (0.024)	0.831 0.869 (0.024)			
		32	213			

χ^2 (222) = 165 , p = n.s.
 L^2 (222) = 193 , p = n.s.

H

New Different Items

Time 2

		"Low"		"High"		
Old Diff. Items Time 2	"Low"	0.251	0.033	0.000	0.076	Raw Score Correlation $r = 0.205$ ($p < .001$) ^A
		(0.018)		(0.018)		
		1.000	0.300	0.000	0.700	Component Score Correlation $r = 0.148$ ($p = .014$) ^A
		(0.072)		(0.072)		
		9		21		
		0.083	0.116	0.666	0.775	
"High"	(0.018)		(0.018)		Model Correlation $\hat{\rho} = 0.957$ ^B	
	0.111	0.131	0.889	0.869		
	(0.024)		(0.024)			
	32		213			

χ^2 (132) = 205 , $p = \text{n.s.}$
 L^2 (132) = 155 , $p = \text{n.s.}$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 24 Cont'd.

I

New Same Items

Time 2

		"Low"		"High"		
New Same Items Time 1	"Low"	0.306	0.033	0.225	0.076	Raw Score Correlation r = 0.205 (p < .001) ^A
		(0.018)		(0.018)		
		0.576	0.300	0.424	0.700	
		9		21	Component Score Correlation r = 0.148 (p = .014) ^A	
	0.090	0.116	0.379	0.775		
	(0.018)		(0.018)			
	"High"	0.188	0.131	0.812	0.869	Model Correlation $\hat{\rho} = 0.207$ ^A
(0.039)			(0.039)			
		32		213		

χ^2 (222) = 185

, p = n.s.

L^2 (222) = 190

, p = n.s.

J

New Different Items
Time 1

		"Low"		"High"		
New Same Items Time 1	"Low"	0.517	0.033	0.097	0.076	Raw Score Correlation $r = 0.205$ (p <.001) ^A
		(0.015)		(0.015)		
		0.843	0.300	0.157	0.700	
		(0.024)		(0.024)		Component Score Correlation $r = 0.148$ (p = .014) ^A
	9		21			
	0.030	0.116	0.356	0.775		
	"High"	(0.015)		(0.015)		Model Correlation $\hat{\rho} = 0.454$ ^A
0.080		0.131	0.920	0.869		
(0.038)			(0.038)			
		32		213		

χ^2 (132) = 138 , p = n.s.
 L^2 (132) = 158 , p = n.s.

K

New Different Items

Time 2

		"Low"		"High"		
New Same Items Time 2	"Low"	0.354	0.033	0.053	0.076	Raw Score Correlation $r = 0.205$ (p < .001) ^A
		(0.016)		(0.016)		
		0.871	0.300	0.129	0.700	Component Score Correlation $r = 0.148$ (p = .014) ^A
		(0.040)		(0.040)		
	9		21			
	"High"	0.033	0.116	0.560	0.775	Model Correlation $\hat{\rho} = 0.518$ ^A
		(0.016)		(0.016)		
		0.056	0.131	0.944	0.869	
		(0.027)		(0.027)		
		32		213		

χ^2 (132) = 141

$, p = n.s.$

I^2 (132) = 134

$, p = n.s.$

L

New Different Items
Time 2

		"Low"		"High"		
New Diff. Items Time 1	"Low"	0.305	0.033	0.173	0.076	Raw Score Correlation $r = 0.205$ (p < .001) ^A
		(0.022)		(0.022)		
		0.634	0.300	0.366	0.700	
		(0.045)		(0.045)		Component Score Correlation $r = 0.148$ (p = .014) ^A
	9		21			
	0.084	0.116	0.438	0.775		
	"High"	(0.022)		(0.022)		Model Correlation $\hat{\rho} = 0.231$ ^A
0.152		0.131	0.848	0.869		
(0.044)			(0.044)			
		32		213		

χ^2 (78) = 97.5 , p = n.s.
 L^2 (78) = 112 , p = n.s.

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 25 Cont'd.

E

		New Same Items				
		Time 1		Time 2		
Old Same Items	Time 1	"Low"		"High"		
		Raw Score Correlation	Component Score Correlation	Raw Score Correlation	Component Score Correlation	
	Time 2	"Low"		"High"		
		Raw Score Correlation	Component Score Correlation	Raw Score Correlation	Component Score Correlation	
Time 2	"Low"	0.162 (0.028)	0.033	0.000 (0.028)	0.076	Raw Score Correlation r = 0.205 (p <.001) ^A
		1.000 (0.266)	0.300	0.000 (0.266)	0.700	
	"High"	0.336 (0.028)	0.116	0.501 (0.028)	0.775	Component Score Correlation r = 0.148 (p = .014) ^B
		0.408 (0.031)	0.131	0.592 (0.031)	0.869	
		32		213		Model Correlation $\hat{\rho} = 0.536$ ^A
		χ^2 (132) = 40.6 , p = n.s. L^2 (132) = 87 , p = n.s.				

F

Old Different Items

Time 2

		"Low"		"High"		
Old Diff. Items Time 1	"Low"	0.225 (0.035)	0.033	0.093 (0.035)	0.076	Raw Score Correlation r = 0.205 (p < .001) ^A
		0.716 (0.114)	0.300	0.284 (0.114)	0.700	
	"High"	9		21		Component Score Correlation r = 0.148 (p = .014) ^A
		0.0103 (0.035)	0.116	0.671 (0.035)	0.775	
		0.024 (0.050)	0.131	0.976 (0.050)	0.869	
		32		213		
						Model Correlation $\hat{\rho} = 0.148$ ^A

χ^2 (222) = 35 , p = n.s.
 L^2 (222) = 82.1 , p = n.s.

G

New Different Items

Time 1

Old
Diff.
Items
Time 1

	"Low"	"High"	
"Low"	0.290 (0.030)	0.033	Raw Score Correlation $r = 0.205$ (p < .001) ^A
	0.970 (0.098)	0.300	
	9	21	Component Score Correlation $r = 0.148$ (p = .014) ^A
	0.137 (0.030)	0.116	
"High"	0.002 (0.030)	0.076	Model Correlation $\hat{\rho} = 0.561$ ^A
	0.030 (0.098)	0.700	
	0.814 (0.043)	0.869	
	32	213	

χ^2 (132) = 93 , p = n.s.
 L^2 (132) = 112 , p = n.s.

H

New Different Items

Time 2

Old
Diff.
Items

Time 2

		"Low"		"High"		
"Low"		0.224	0.033	0.001	0.076	Raw Score Correlation r = 0.205 (p < .001) ^A
		(0.026)		(0.026)		
		1.000	0.300	0.000	0.700	Component Score Correlation r = 0.148 (p = .014) ^A
		(0.131)		(0.131)		
	9			21		
"High"		0.164	0.116	0.611	0.775	Model Correlation $\hat{\rho} = 0.711$ ^A
		(0.026)		(0.026)		
		0.220	0.131	0.780	0.869	
		(0.033)		(0.033)		
	32			213		

χ^2 (132) = 69

, p = n.s.

L^2 (132) = 94.5

, p = n.s.

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 25 Cont'd.

I

		New Same Items Time 2				
		"Low"		"High"		
New Same Items Time 1	"Low"	0.462 (0.025)	0.033	0.071 (0.025)	0.076	Raw Score Correlation r = 0.205 (p <.001) ^A
		0.859 (0.046)	0.300	0.141 (0.046)	0.700	
		9		21		
	"High"	0.025 (0.025)	0.116	0.442 (0.025)	0.775	Component Score Correlation r = 0.148 (p = .014) ^A
		0.041 (0.055)	0.131	0.959 (0.055)	0.869	
		32		213		
					Model Correlation $\hat{\rho}$ = 0.526 ^A	
<div> χ^2 (222) = 47.5 , p = n.s. </div> <div> L^2 (222) = 84.3 , p = n.s. </div>						

J

New Different Items
Time 1

		"Low"		"High"		
New Same Items Time 1	"Low"	0.382	0.033	0.170	0.076	Raw Score Correlation $r = 0.205$ ($p < .001$) ^A
		(0.029)		(0.029)		
		0.692	0.300	0.308	0.700	
		(0.052)		(0.052)		Component Score Correlation $r = 0.148$ ($p = .014$) ^A
	9		21			
	"High"	0.047	0.116	0.400	0.775	Model Correlation $\hat{\rho} = 0.424$ ^A
(0.029)			(0.029)			
0.105		0.131	0.895	0.869		
		(0.065)		(0.065)		
		32		213		

χ^2 (132) = 70.4 , $p = \text{n.s.}$
 L^2 (132) = 91.4 , $p = \text{n.s.}$

K

		New Different Items Time 2				
		"Low"		"High"		
New Same Items Time 2	"Low"	0.412 (0.030)	0.033	0.079 (0.030)	0.076	Raw Score Correlation r = 0.205 (p <.001) ^A
		0.834 (0.061)	0.300	0.166 (0.061)	0.700	
		9		21		Component Score Correlation r = 0.148 (p = .014) ^A
	"High"	0.001 (0.030)	0.116	0.508 (0.030)	0.775	
		0.000 (0.061)	0.131	1.000 (0.061)	0.869	Model Correlation $\hat{\rho}$ = 0.498 ^A
		32		213		
χ^2 (132) = 27.1 , p = n.s. L^2 (132) = 74.2 , p = n.s.						

L

New Different Items

Time 2

		"Low"		"High"		
New Diff. Items Time 1	"Low"	0.438	0.033	0.025	0.076	Raw Score Correlation $r = 0.205$ ($p < .001$) ^A
		(0.032)		(0.032)		
		0.940	0.300	0.060	0.700	
		(0.068)		(0.068)		
		9		21		Component Score Correlation $r = 0.148$ ($p = .014$) ^A
	"High"	0.044	0.116	0.493	0.775	
(0.032)			(0.032)			
0.077		0.131	0.923	0.869		
	(0.060)		(0.060)		Model Correlation $\hat{\rho} = 0.640$ ^B	
	32		213			

χ^2 (78) = 87.2
 L^2 (78) = 65.2

$, p = \text{n.s.}$
 $, p = \text{n.s.}$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 26. Patterns of Joint Proportion Estimates (π 's).

Variable Y			
A.	"Low"	"High"	
Variable X	"Low"	X	O
	"High"	O	X

Variable Y			
B.	"Low"	"High"	
Variable X	"Low"	X	X
	"High"	O	X

Variable Y			
C.	"Low"	"High"	
Variable X	Low"	X	O
	"High"	X	X

Note: X's represent large proportions (π_{ab} -estimates), while O's represent small proportions (π_{ab} -estimates).

Table 27. Joint Two Component and Univariate Three Component Model Comparison.

Univariate Three Component Membership
Joint Two Component Membership

		Old Items at Time 1		
Old Same Items at Time 1	Old Different Items at Time 1	Low	Intermediate	High
Low	Low	30	8	0
Low	High	2	11	0
High	Low	29	107	0
High	High	0	36	231

$$\chi^2 (6) = 443.7, p < .0001$$

Univariate Three Component Membership
Joint Two Component Membership

		New Items at Time 1		
New Same Items at Time 1	New Different Items at Time 1	Low	Intermediate	High
Low	Low	120	12	0
Low	High	28	114	0
High	Low	4	20	0
High	High	0	37	119

$$\chi^2 (6) = 529.5, p < .0001$$

Table 28. Joint Three Component Frequency Tables and Model Estimates for All Subjects on Old and New Items at Times 1 and 2.

A

New Items, Time 1

		"Low"		"Intermediate"		"High"			
Old Items Time 1	"Low"	0.136	0.099	0.000	0.033	0.000	0.002	Raw Score Correlation r = 0.618 (p < 0.001) ^A	
		(0.011)		(0.015)		(0.007)			
		0.997	0.738	0.001	0.246	0.001	0.016		
			(0.077)		(0.110)		(0.053)		
			45		15		1		
	"Inter."	0.223	0.163	0.020	0.154	0.002	0.040	Component Score Correlation r = 0.498 (p < 0.001) ^A	
		(0.020)		(0.023)		(0.010)			
		0.915	0.457	0.078	0.432	0.007	0.111		
			(0.081)		(0.093)		(0.039)		
		74		70		18			
"High"	0.075	0.073	0.271	0.216	0.273	0.220	Model Correlation $\hat{\rho} = 0.585$ ^A		
	(0.016)		(0.018)		(0.006)				
	0.122	0.143	0.438	0.424	0.440	0.433			
		(0.025)		(0.029)		(0.010)			
		33		98		100			

$$\chi^2 (565) = 274.7, p = \text{n.s.}$$

$$L^2 (565) = 336.6, p = \text{n.s.}$$

B

New Items, Time 2

		"Low"		"Intermediate"		"High"		
Old Items Time 2	"Low"	0.163 (0.017)	0.119	0.000 (0.030)	0.008	0.000 (0.026)	0.000	Raw Score Correlation r = 0.731 (p < 0.001) ^A
		0.959 (0.099)	0.938	0.021 (0.174)	0.063	0.021 (0.149)	0.000	
		45		3		0		
	"Inter."	0.097 (0.021)	0.088	0.029 (0.032)	0.125	0.000 (0.028)	0.050	Component Score Correlation r = 0.631 (p < 0.001) ^A
		1.000 (0.263)	0.333	0.000 (0.403)	0.475	0.000 (0.346)	0.192	
		33		47		19		
"High"	0.028 (0.015)	0.040	0.258 (0.026)	0.225	0.425 (0.018)	0.345	Model Correlation $\hat{\rho} = 0.895$ ^B	
	0.053 (0.020)	0.065	0.362 (0.034)	0.370	0.585 (0.024)	0.565		
	15		85		130			

$$\chi^2 (565) = 905.2, p < 0.05$$

$$L^2 (565) = 433.2, p = \text{n.s.}$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 28 Cont'd.

C

Old Items, Time 2

Old Items Time 1		"Low"		"Intermediate"		"High"		Raw Score Correlation $r = 0.250$ ($p < 0.001$) ^A
		0.037 (0.015)	0.025	0.030 (0.022)	0.051	0.026 (0.019)	0.036	
		0.393 (0.147)	0.226	0.322 (0.226)	0.452	0.284 (0.187)	0.323	
		7		14		10		
	"Inter."	0.050 (0.020)	0.058	0.075 (0.030)	0.124	0.126 (0.027)	0.167	Component Score Correlation $r = 0.306$ ($p < 0.001$) ^A
		0.191 (0.084)	0.167	0.297 (0.125)	0.354	0.512 (0.115)	0.479	
		16		34		46		
	"High"	0.077 (0.016)	0.044	0.000 (0.022)	0.084	0.580 (0.022)	0.411	Model Correlation $\hat{\rho} = 0.194$ ^A
		0.119 (0.024)	0.081	0.003 (0.034)	0.155	0.878 (0.033)	0.764	
		12		23		113		

$$\chi^2 (565) = 576.9, p = \text{n.s.}$$

$$L^2 (565) = 352.1, p = \text{n.s.}$$

D

New Items, Time 2

New Items Time 1		"Low"		"Intermediate"		"High"		Raw Score Correlation $r = 0.450$ ($p < 0.001$) ^A
		0.216 (0.020)	0.149	0.117 (0.026)	0.109	0.027 (0.018)	0.036	
		0.579 (0.050)	0.506	0.325 (0.066)	0.370	0.097 (0.047)	0.123	
		41		30		10		
	"Inter."	0.101 (0.021)	0.055	0.162 (0.041)	0.196	0.101 (0.036)	0.149	Component Score Correlation $r = 0.481$ ($p < 0.001$) ^A
		0.262 (0.072)	0.136	0.476 (0.144)	0.491	0.261 (0.127)	0.373	
		15		54		41		
	"High"	0.000 (0.002)	0.036	0.000 (0.031)	0.051	0.276 (0.031)	0.218	Model Correlation $\hat{\rho} = 0.355$ ^A
		0.050 (0.008)	0.119	0.050 (0.097)	0.167	0.901 (0.096)	0.714	
		10		14		60		

$$\chi^2 (565) = 236.4, p = \text{n.s.}$$

$$L^2 (565) = 280.3, p = \text{n.s.}$$

Note: Within each panel, correlations with the same letter are non-significantly different.

Table 29. Pattern of Membership in the Non-zero Cells in the Three Component Bivariate Mixture.

		New Items, Time X		
		"Low"	"Intermediate"	"High"
Old Items Time X	"Low"	OS1 "Low" OD1 "Low"	0	0
		NS1 "Low" ND1 "Low"		
	"Inter."	OS1 "High" OD1 "Low"	0	0
		NS1 "Low" ND1 "Low"		
	"High"	OS1 "High" OD1 "High"	OS1 "High" OD1 "High"	OS1 "High" OD1 "High"
		NS1 "Low" ND1 "Low"		
			NS1 "Low" ND1 "High"	NS1 "High" ND1 "High"

		Performance at Time 2 by group		
		“Low”	“Mid”	“High”
Performance at Time 1 by Group	“High”	0	0	τ_{33}
	“Mid”	0	τ_{22}	0
	“Low”	τ_{11}	0	0

Figure 1. State Change restrictions under the hypothesis that no learning occurs.
Note That Each Row Sums to 1.

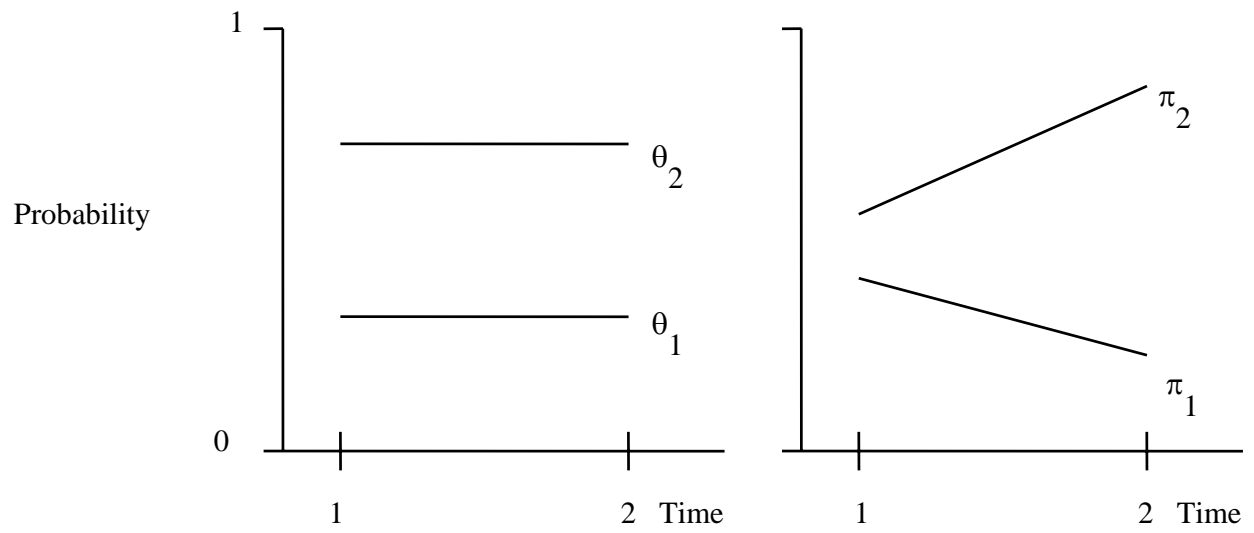


Figure 2. Hypothetical Growth Function Over Two Time Points of the Kind Predicted by Piaget and Inhelder (1956).

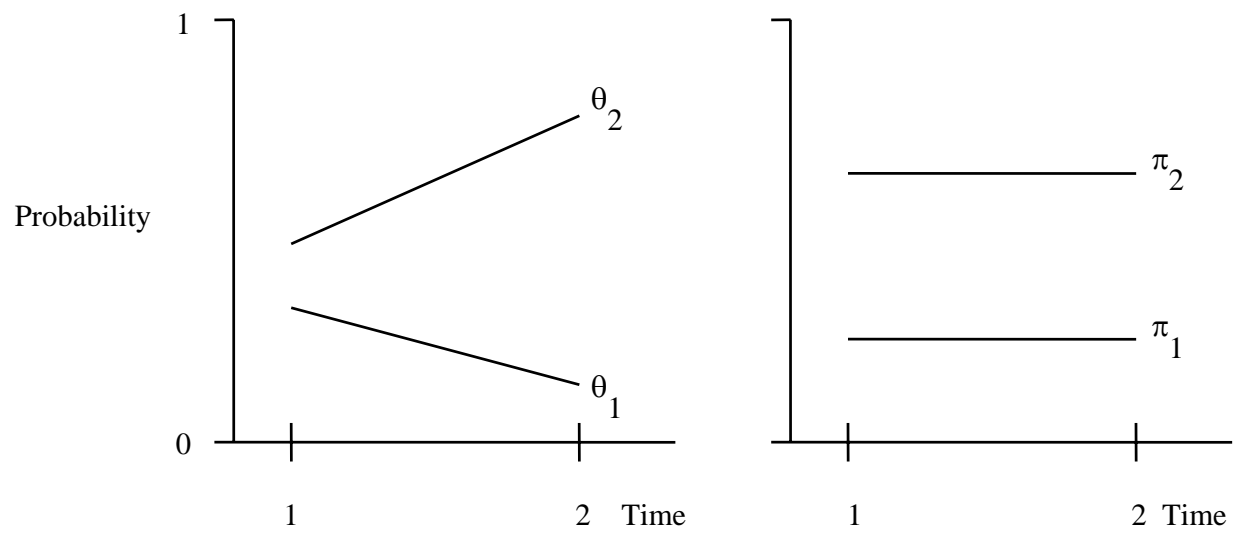


Figure 3. Hypothetical Change In Probabilities Of Success Without Component Membership Change.

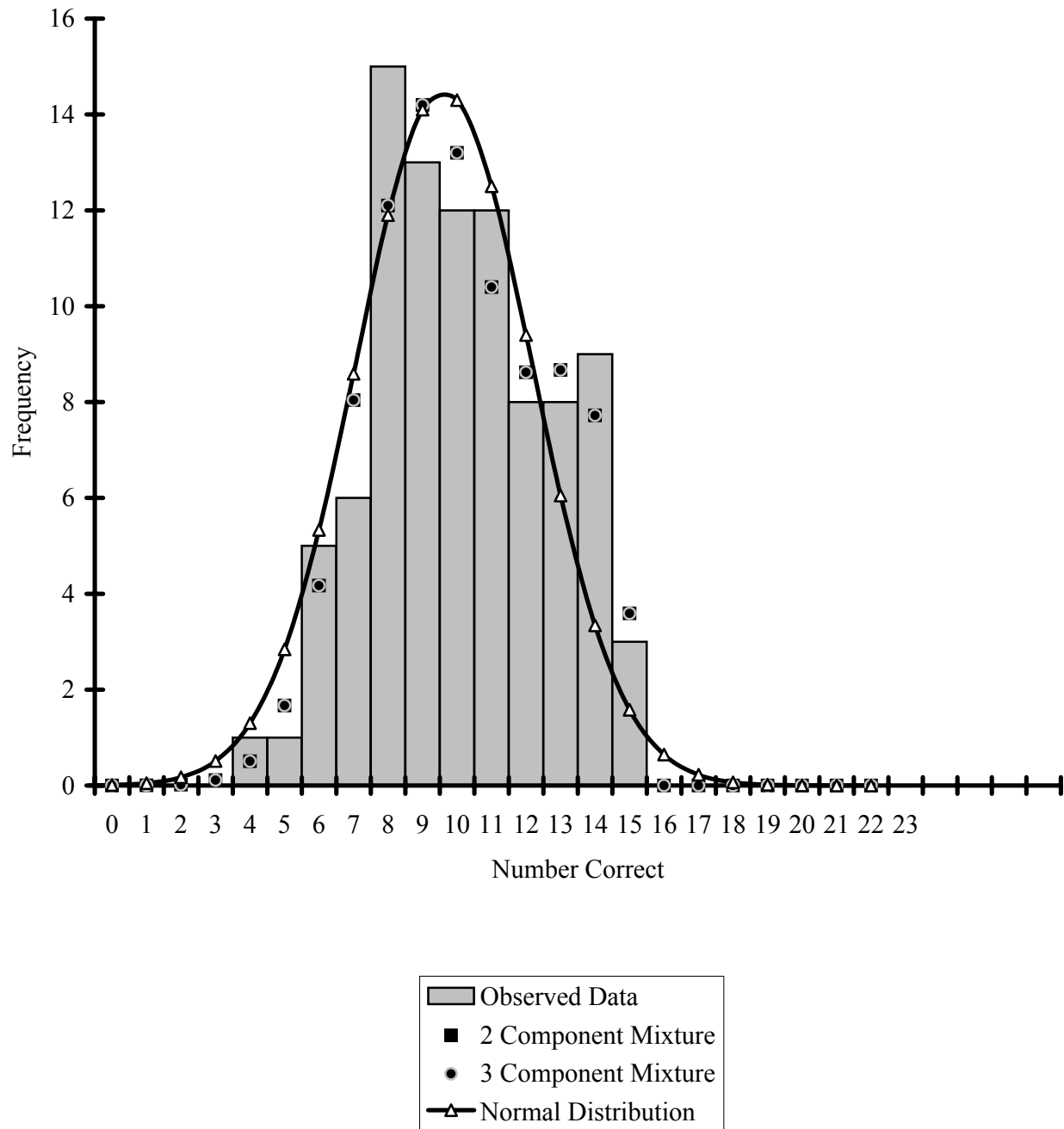


Figure 4. Penn State Females' Performance on New-Same Items at Time 1, with Normal, 2- and 3-Component Binomial Mixture Estimates. Sample size $n = 93$ for 15 items.

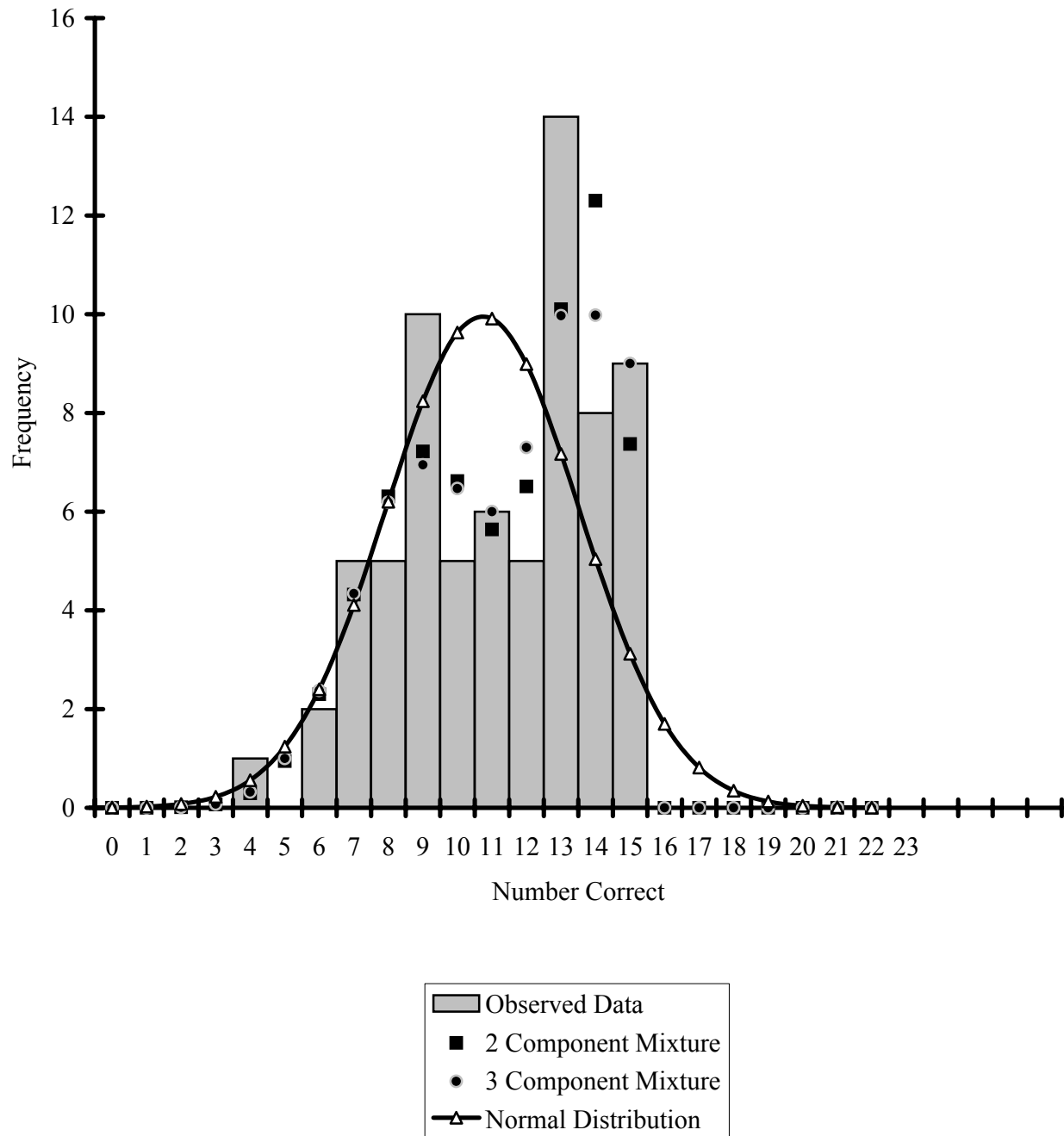


Figure 5. Penn State Females' Performance on New-Same Items at Time 2, with Normal, 2- and 3-Component Binomial Mixture Estimates. Sample size $n = 70$ for 9 items.

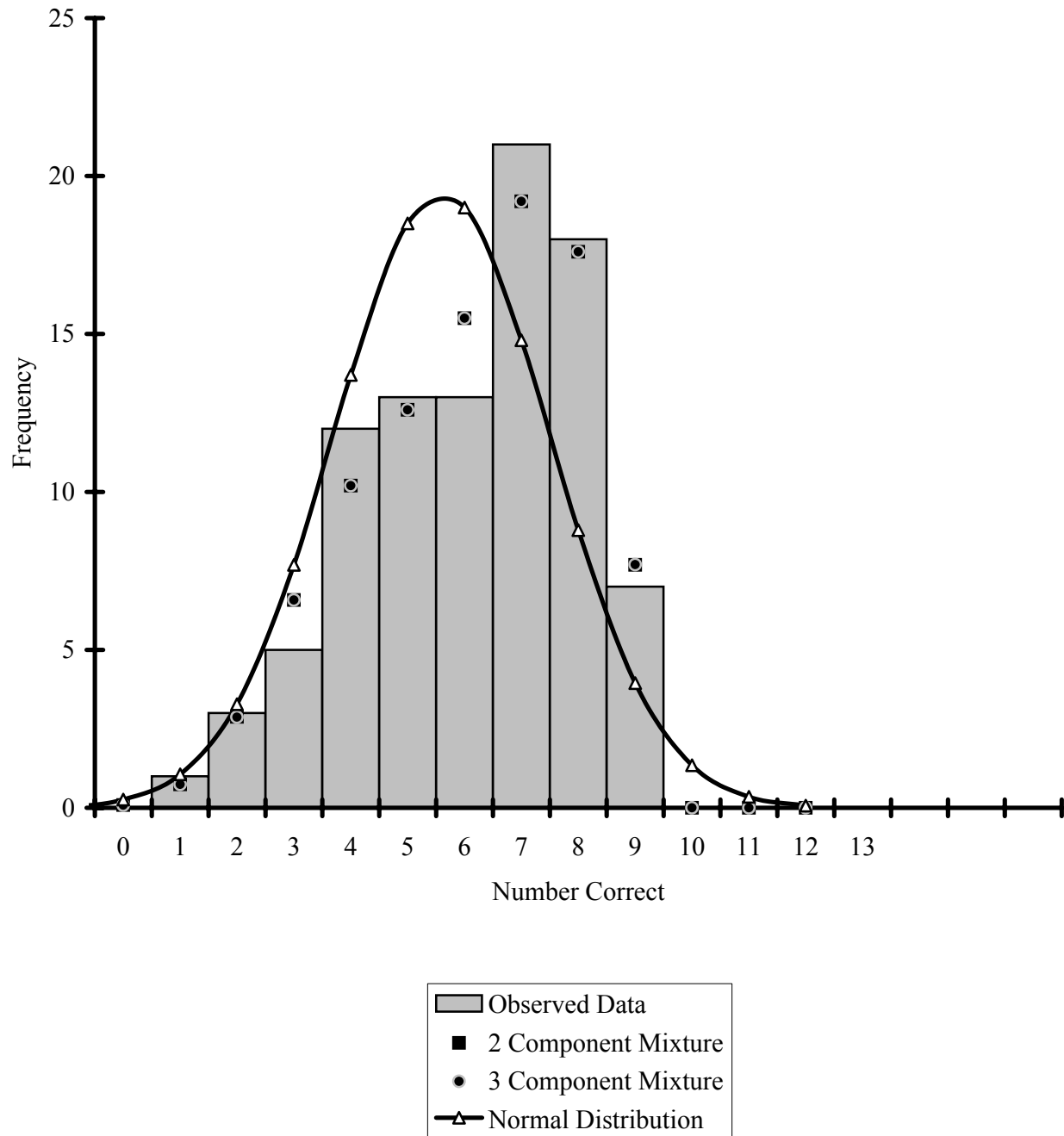


Figure 6. Penn State Females' Performance on New-Different Items at Time 1, with Normal, 2- and 3-Component Binomial Mixture Estimates. Sample size $n = 93$ for 9 items.

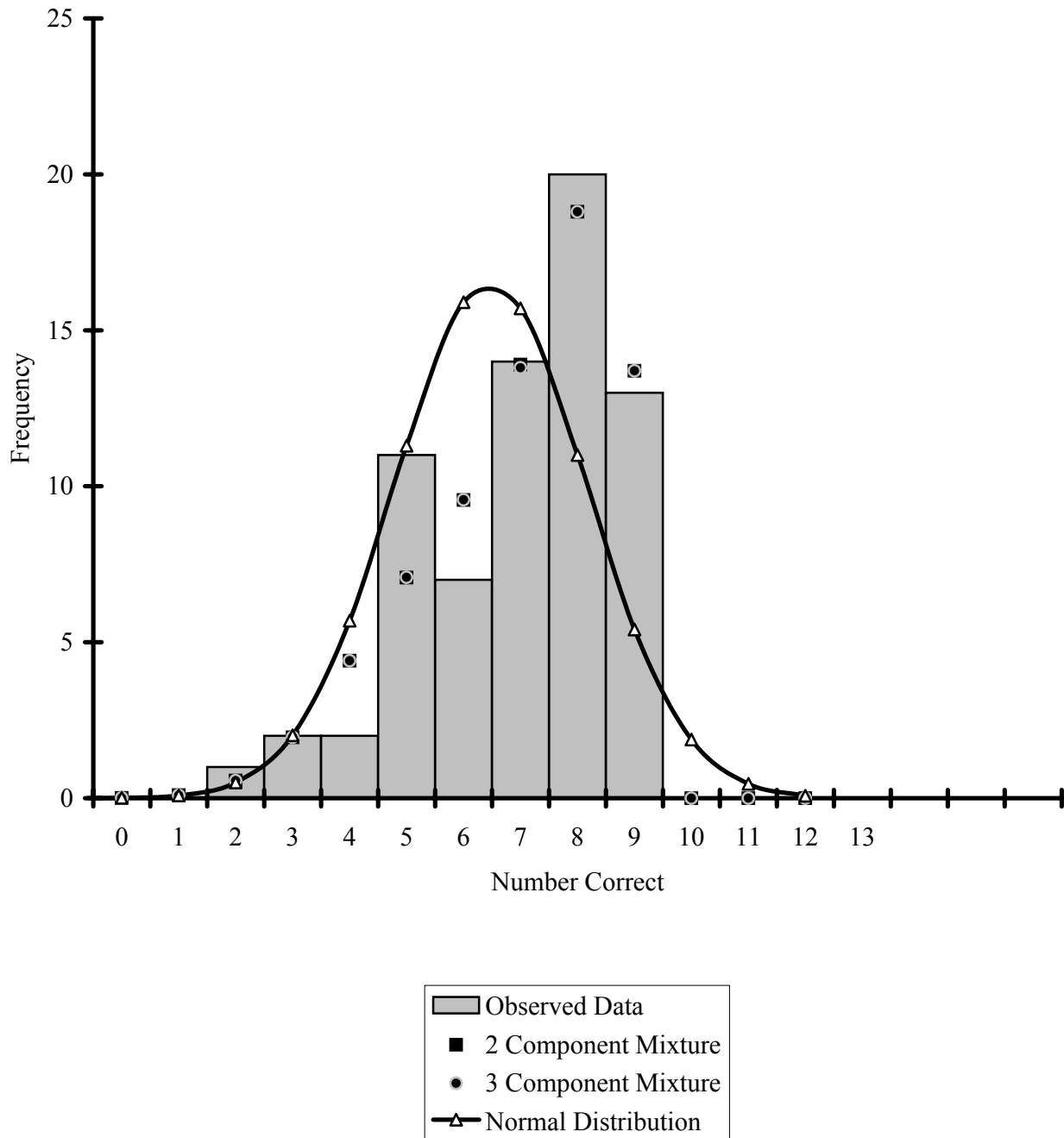


Figure 7. Penn State Females' Performance on New-Different Items at Time 2, with Normal, 2- and 3-Component Binomial Mixture Estimates. Sample size $n = 70$ for 9 items.

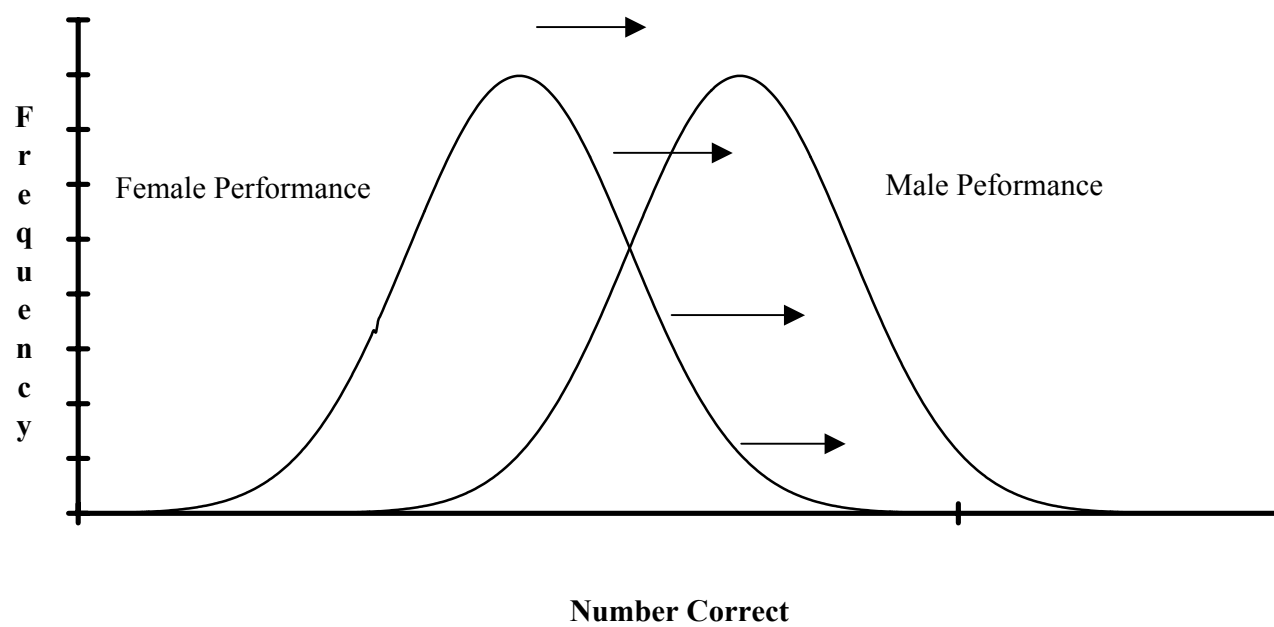


Figure 8. Additive Shift Model of Sex-Differences.

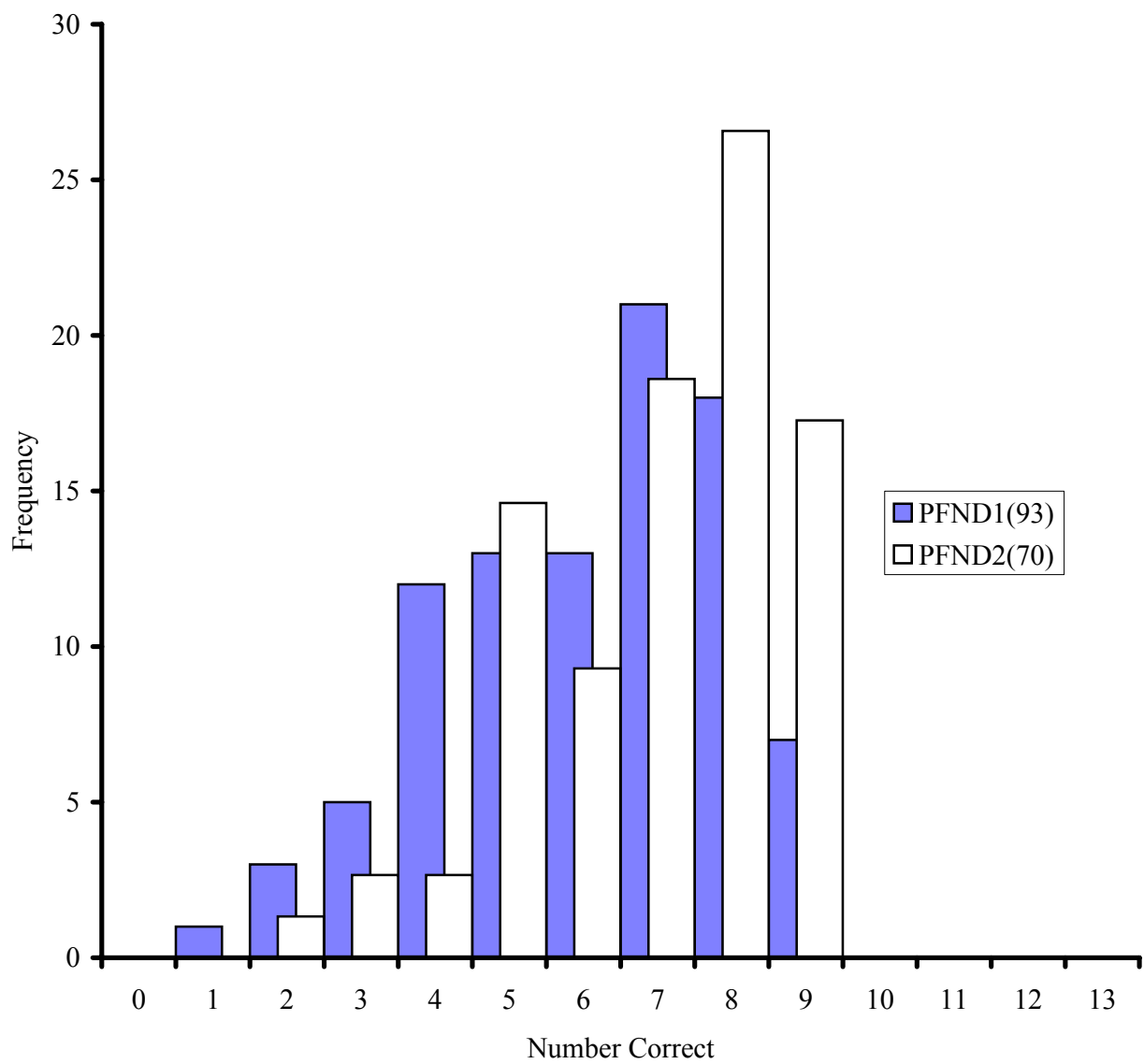


Figure 9. Additive Shift Failure in Mental Rotation Performance.

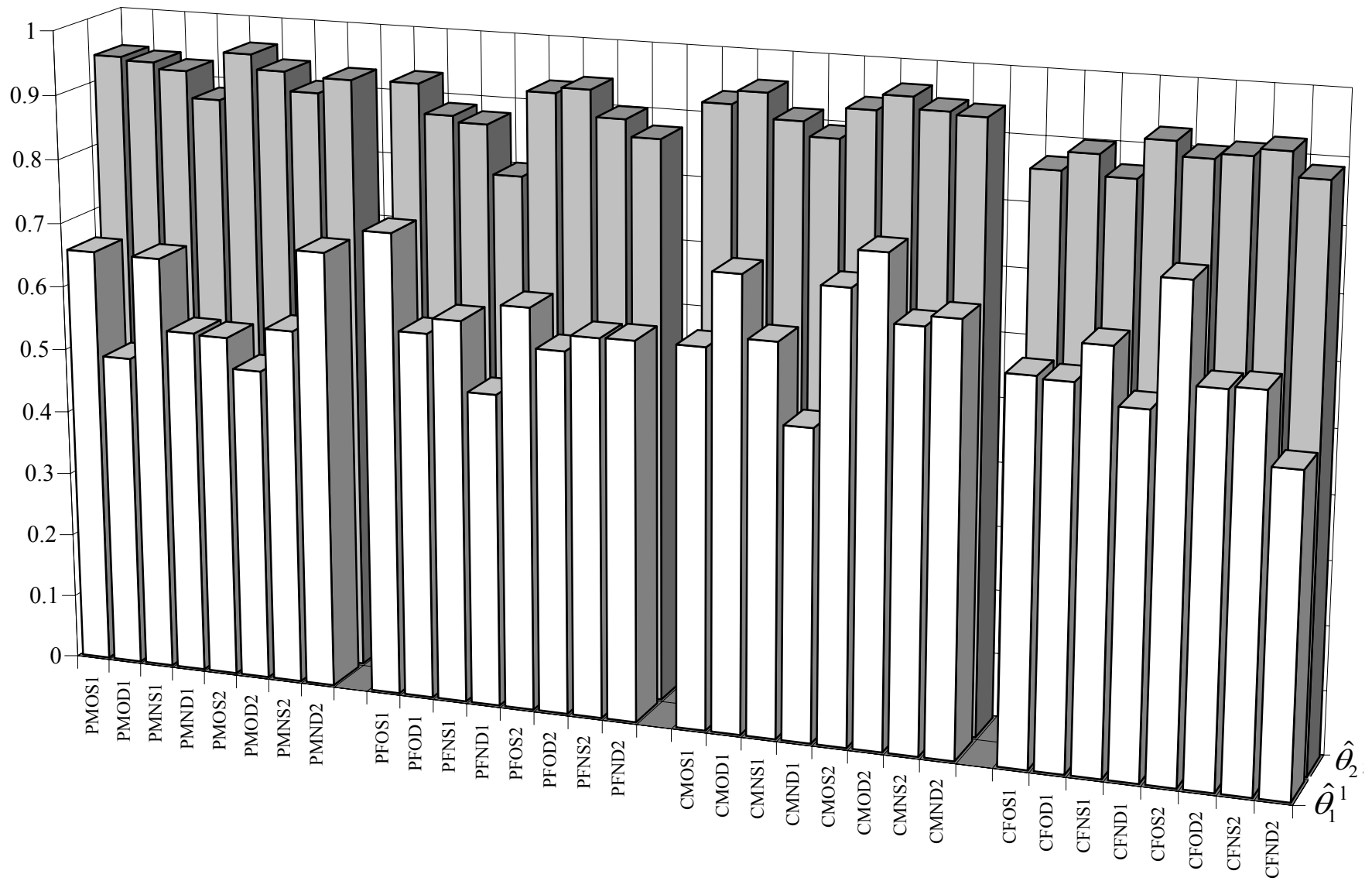


Figure 10. Probability of Success (θ) Estimates for Each of the 32 Basic-level Groups.

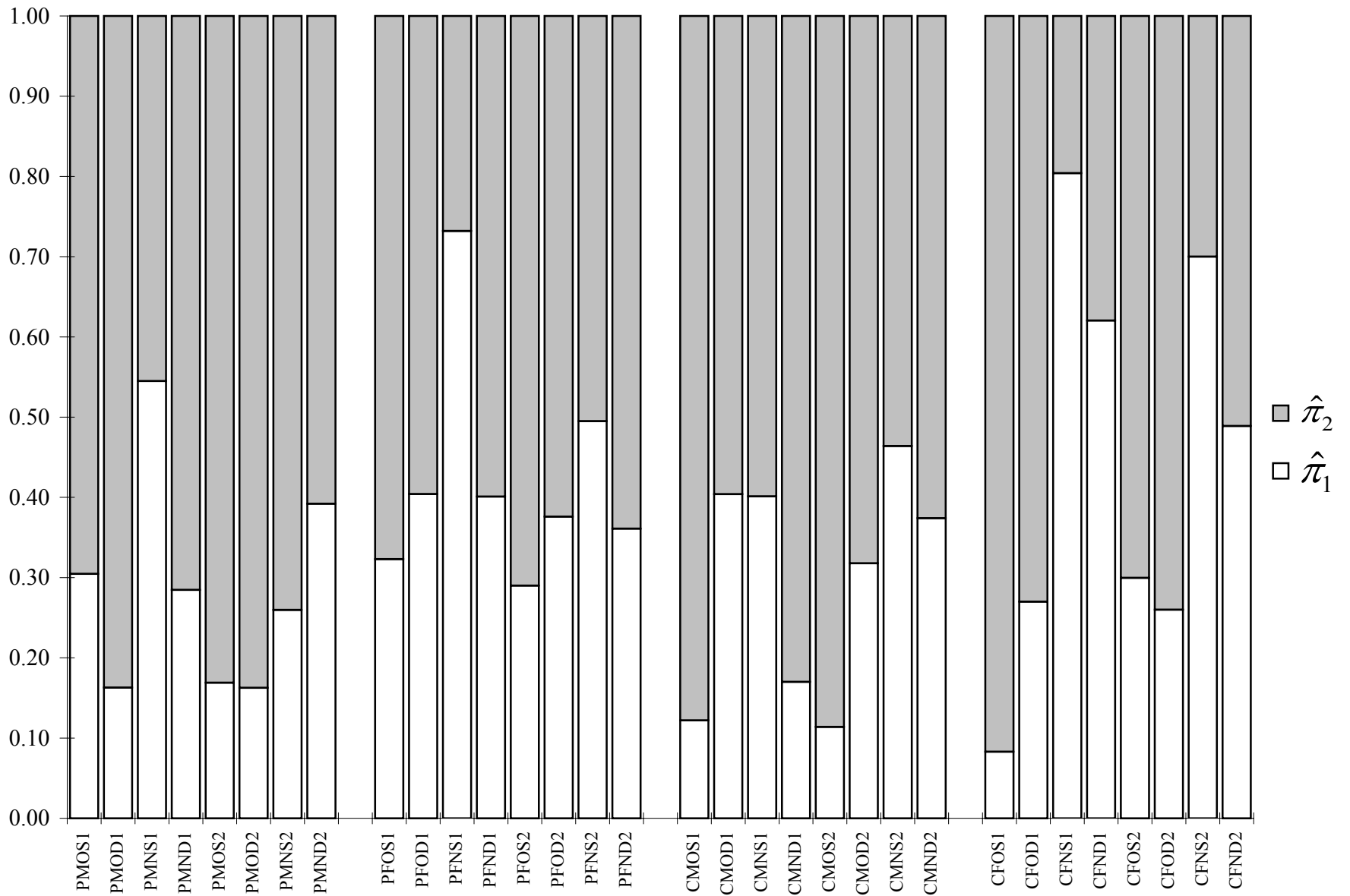


Figure 11. Proportion (π) estimates for Each of the 32 Basic-level Groups. Probabilities are the Lengths Referenced to the Ordinal Scale.

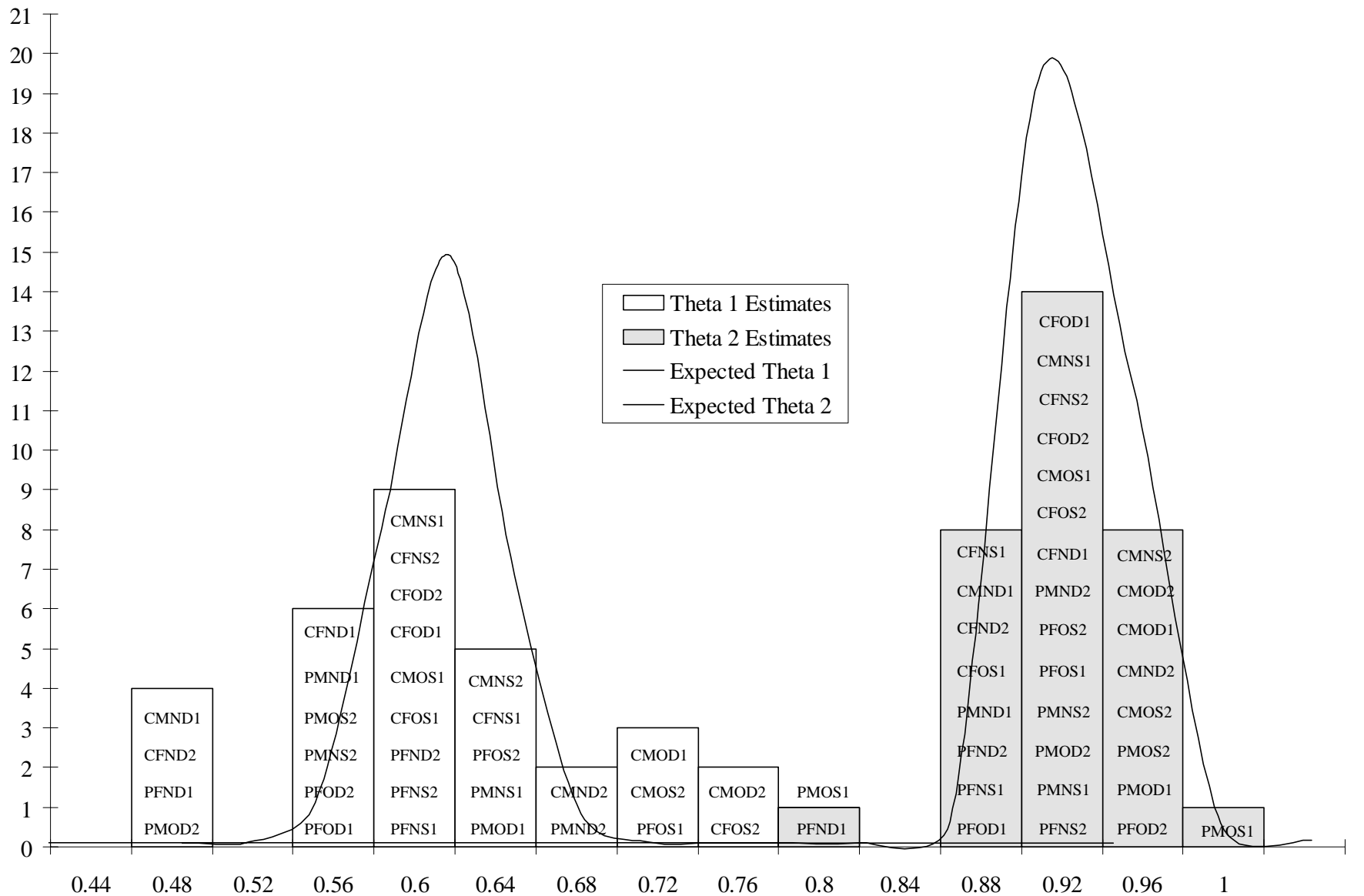


Figure 12. Expected and Observed Bimodal Distributions of Theta 1 and Theta 2 for the 32 Basic-Level Groups.

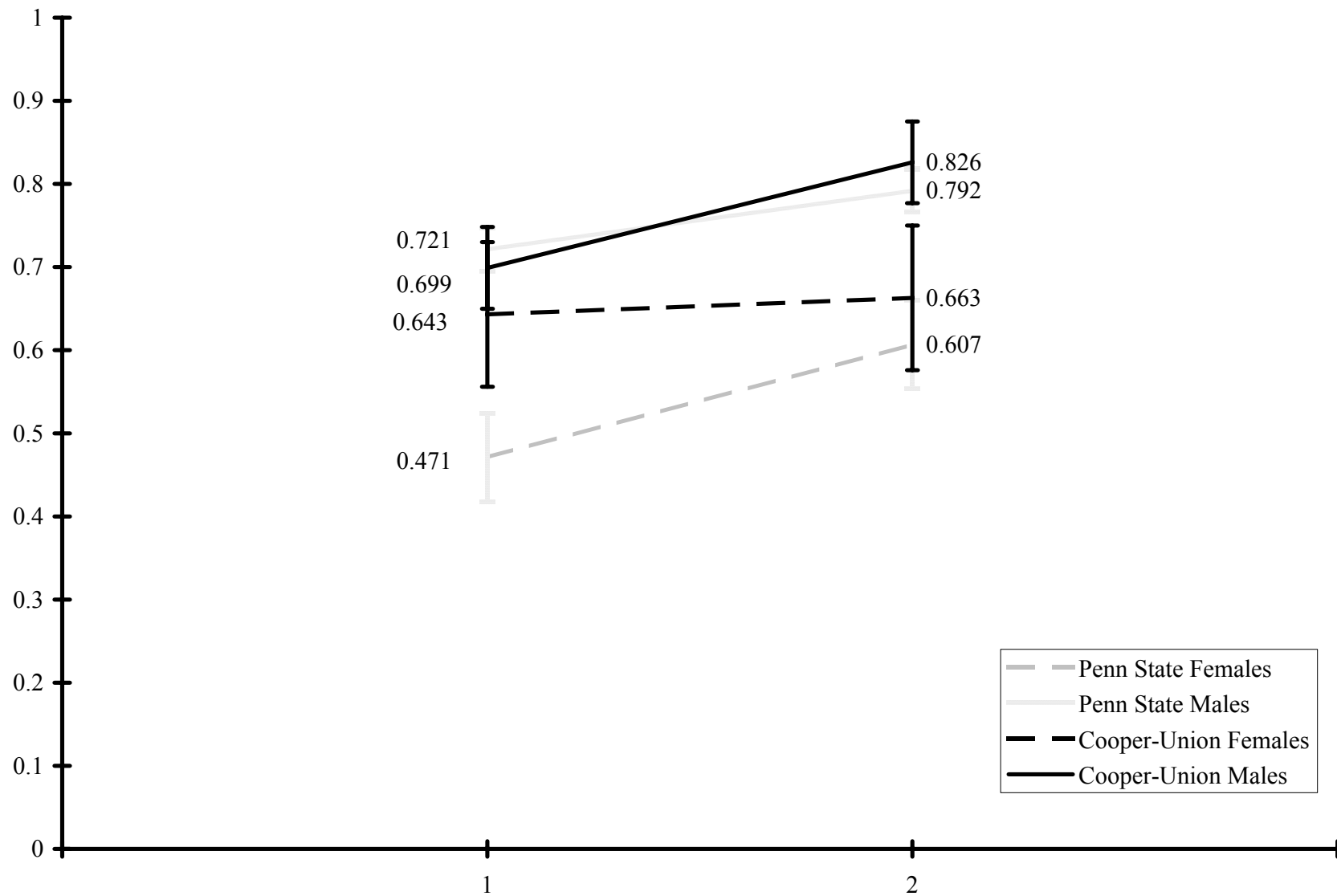


Figure 13. Old-different item "High" Component Proportion Estimates (π_2) and Standard Errors for Penn State and Cooper-Union Males and Females at Time 1 and Time 2.

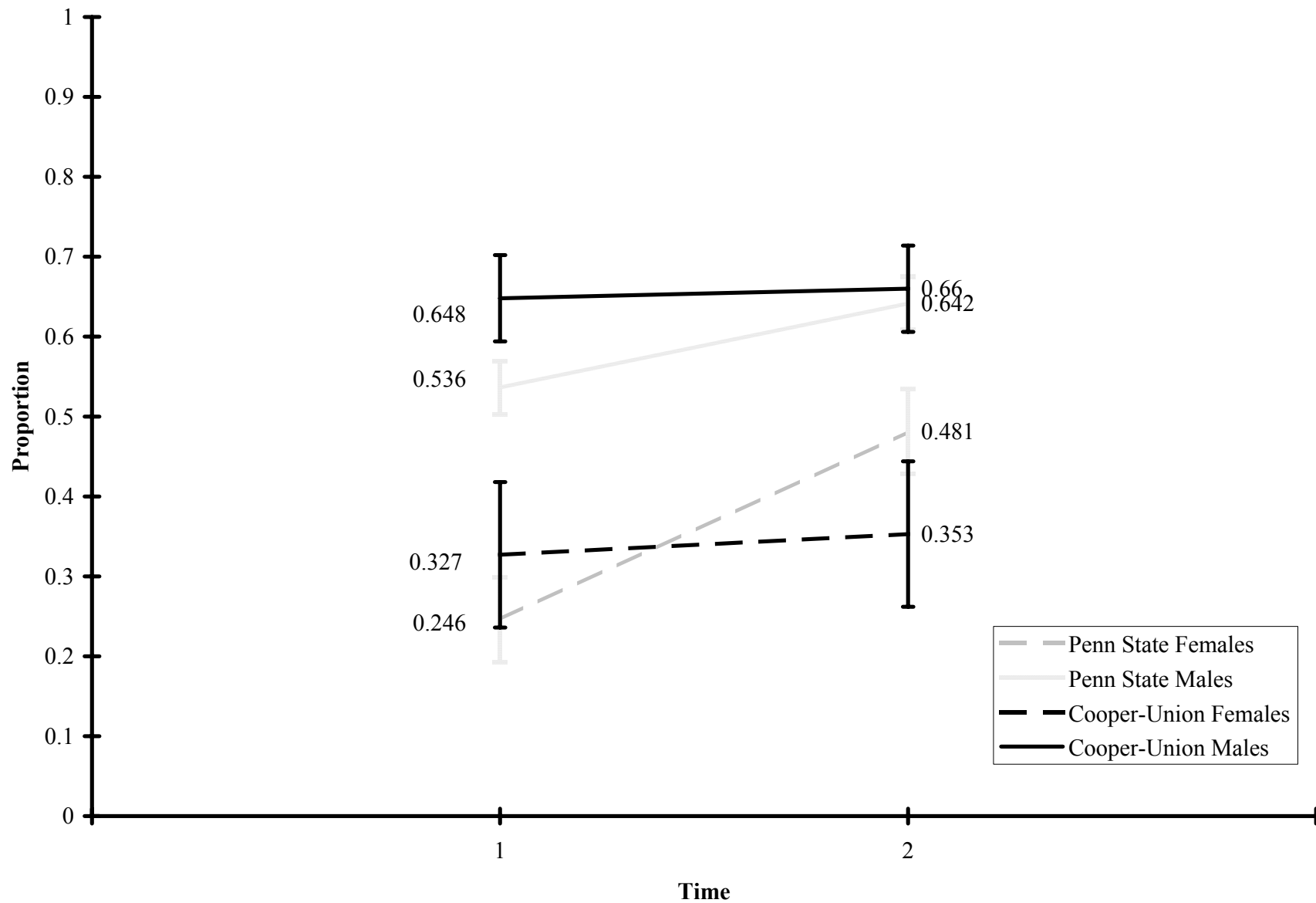


Figure 14. New-different item "High" Component Proportion Estimates (π_2) and Standard Errors for Penn State and Cooper-Union Males and Females at Time 1 and Time 2.

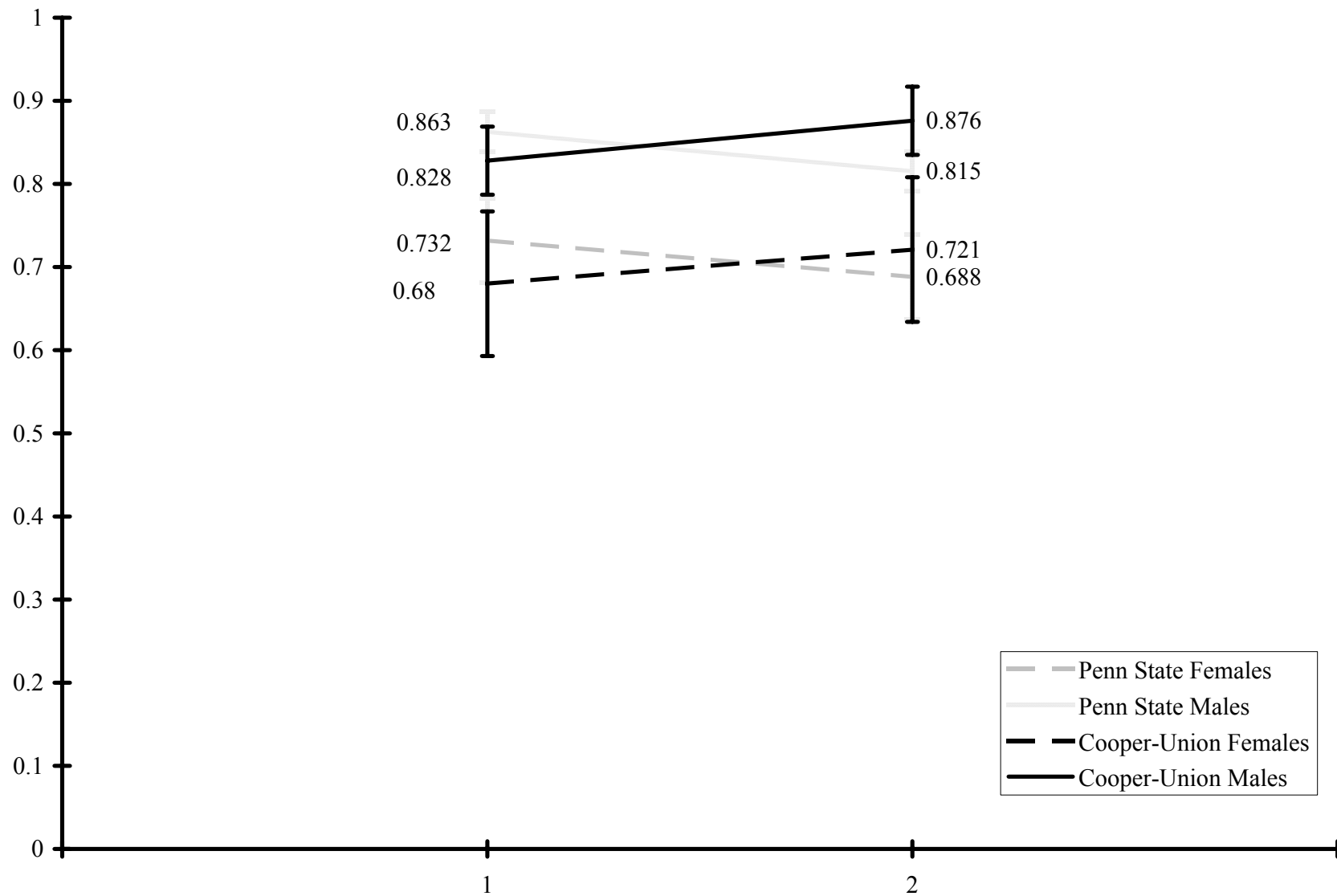


Figure 15. Old-same item "High" Component Proportion Estimates (π_2) and Standard Errors for Penn State and Cooper-Union Males and Females at Time 1 and Time 2.

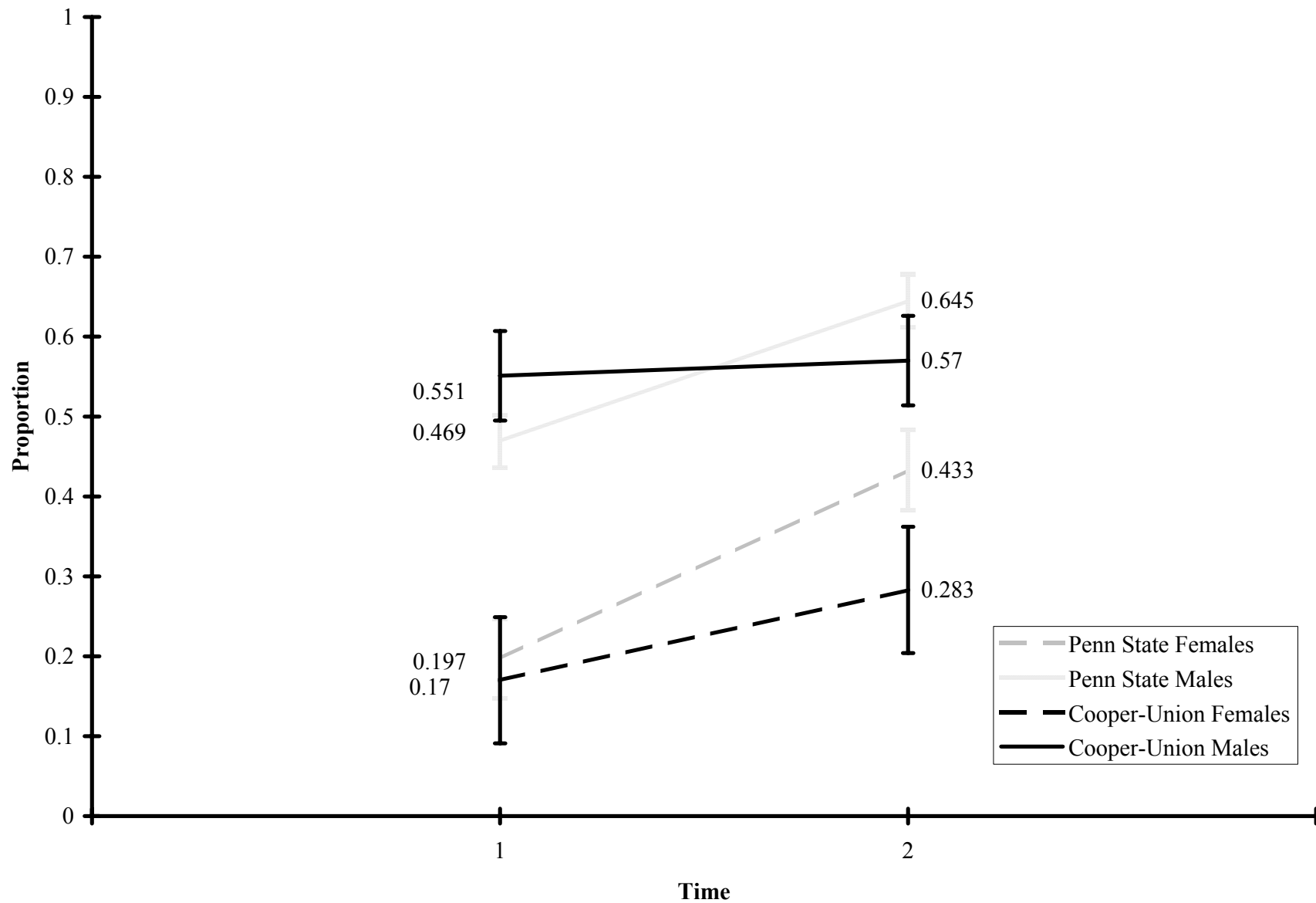


Figure 16. New-Same item "High" Component Proportion Estimates (π_2) and Standard Errors for Penn State and Cooper-Union Males and Females at Time 1 and Time 2.

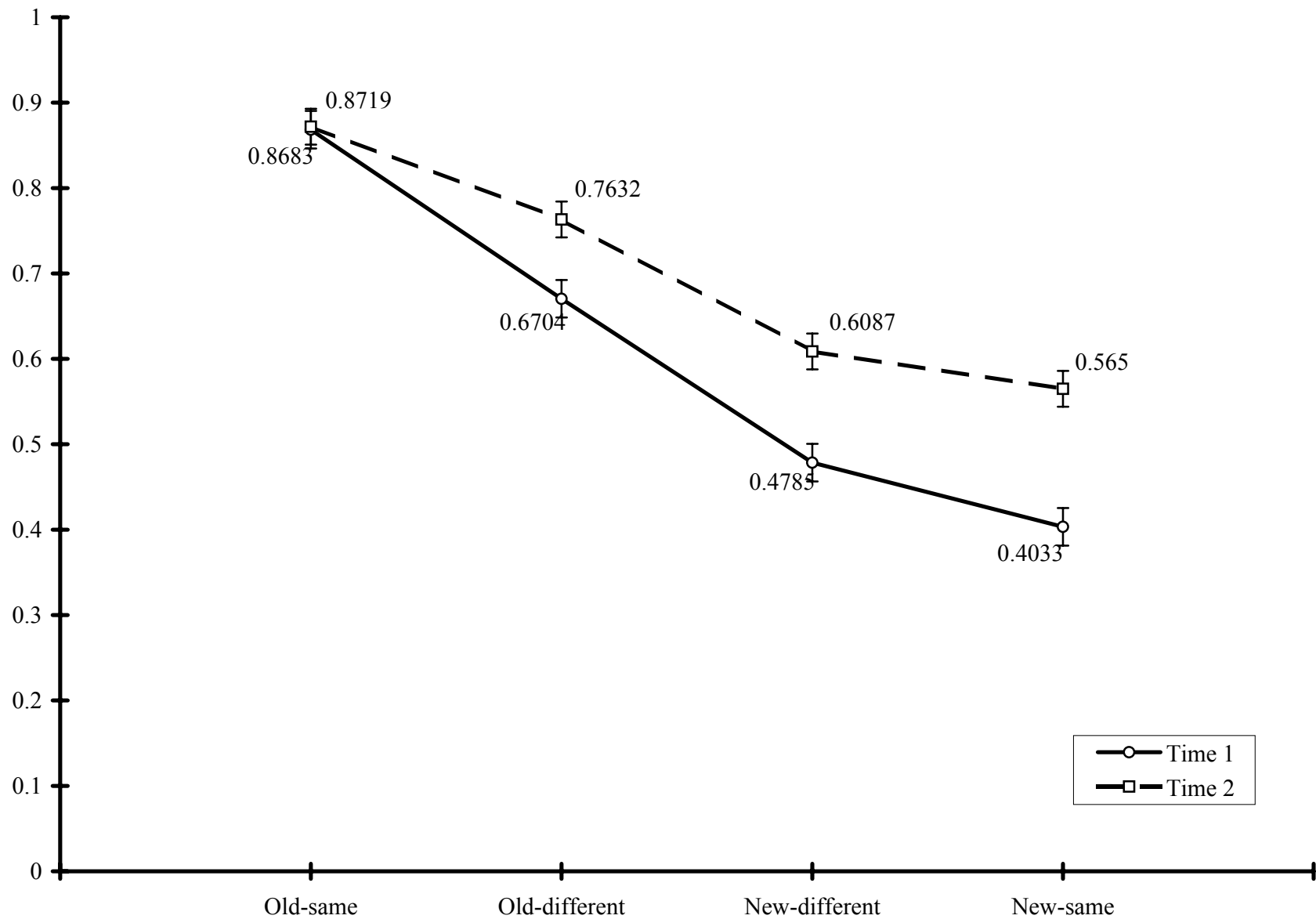


Figure 17. "High" Component Mixing Proportions (π_2) and Standard Errors for All Four Item-type by Item-status Sets for Time 1 and Time 2.

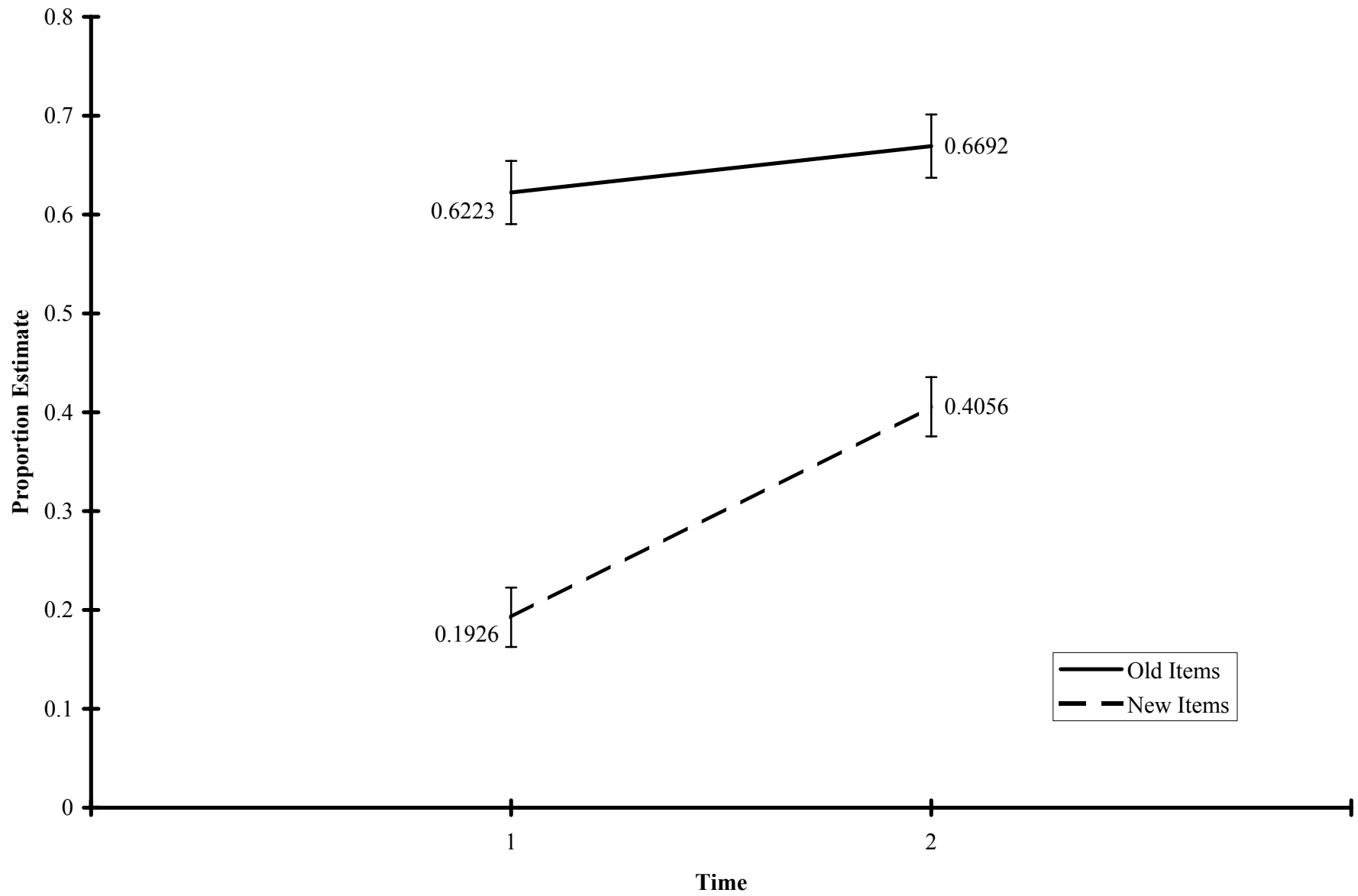


Figure 18. Female "High" Component Proportion Estimates (π_2) and Standard Errors for Old and New Items.

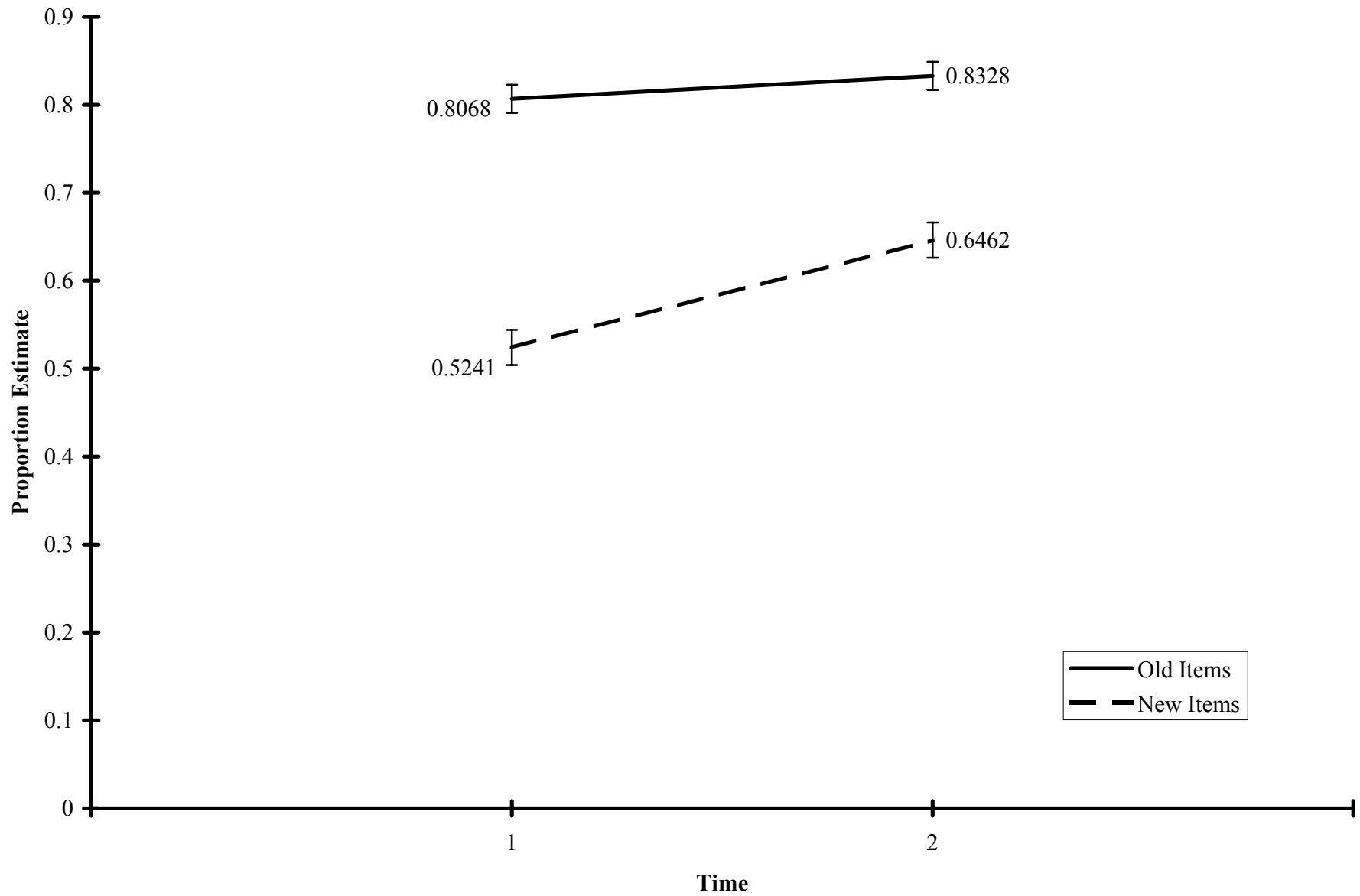


Figure 19. Male "High" Component Proportion Estimates (π_2) and Standard Errors for Old and New Items.

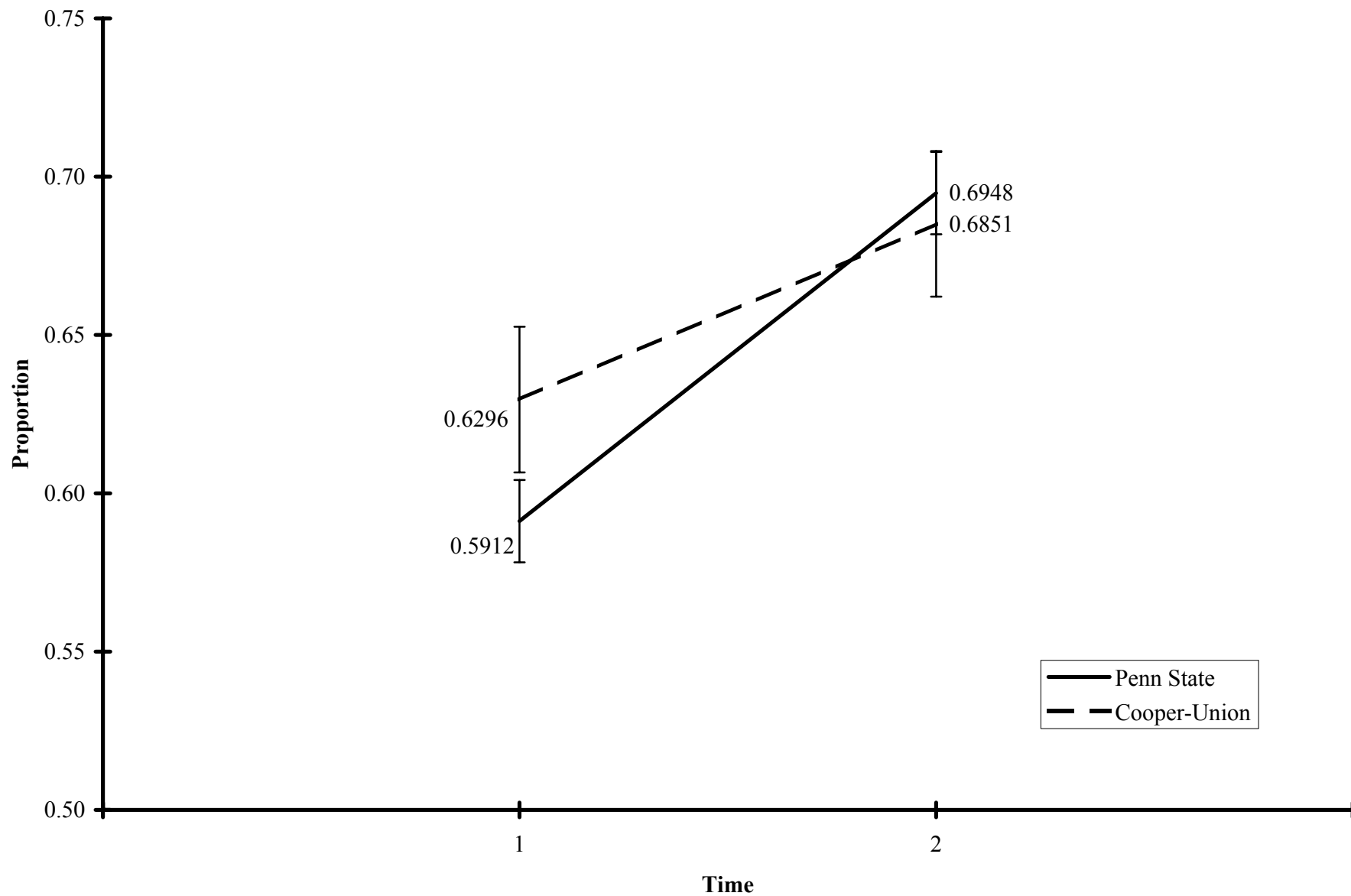


Figure 20. "High" Component Proportion Estimates (π_2) for Penn State and Cooper Union Subjects' Performance on All items at Time 1 and Time 2.

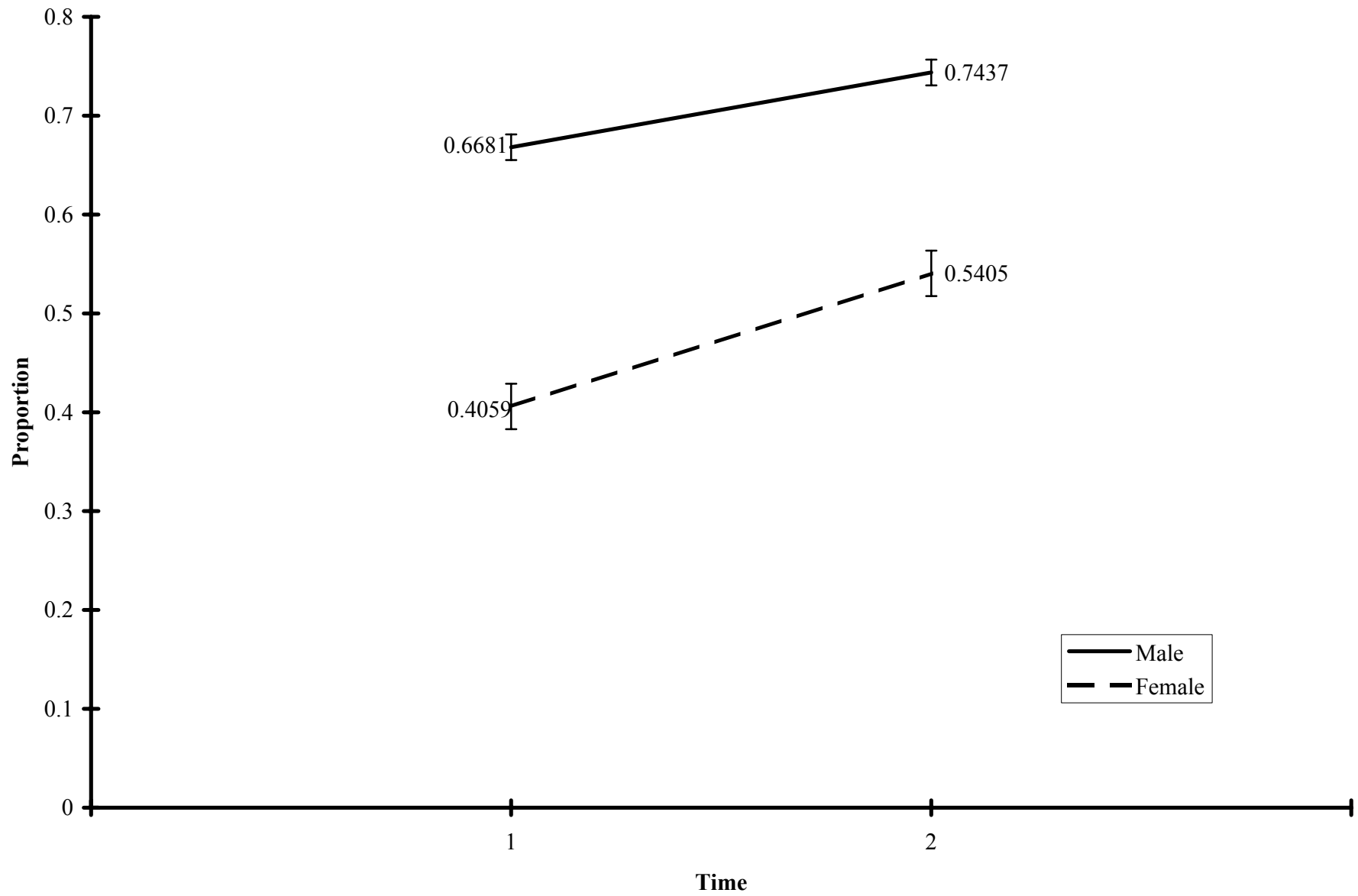


Figure 21. "High" Component Proportion Estimates (π_2) for Male and Female Subjects' Performance on All items at Time 1 and Time 2.

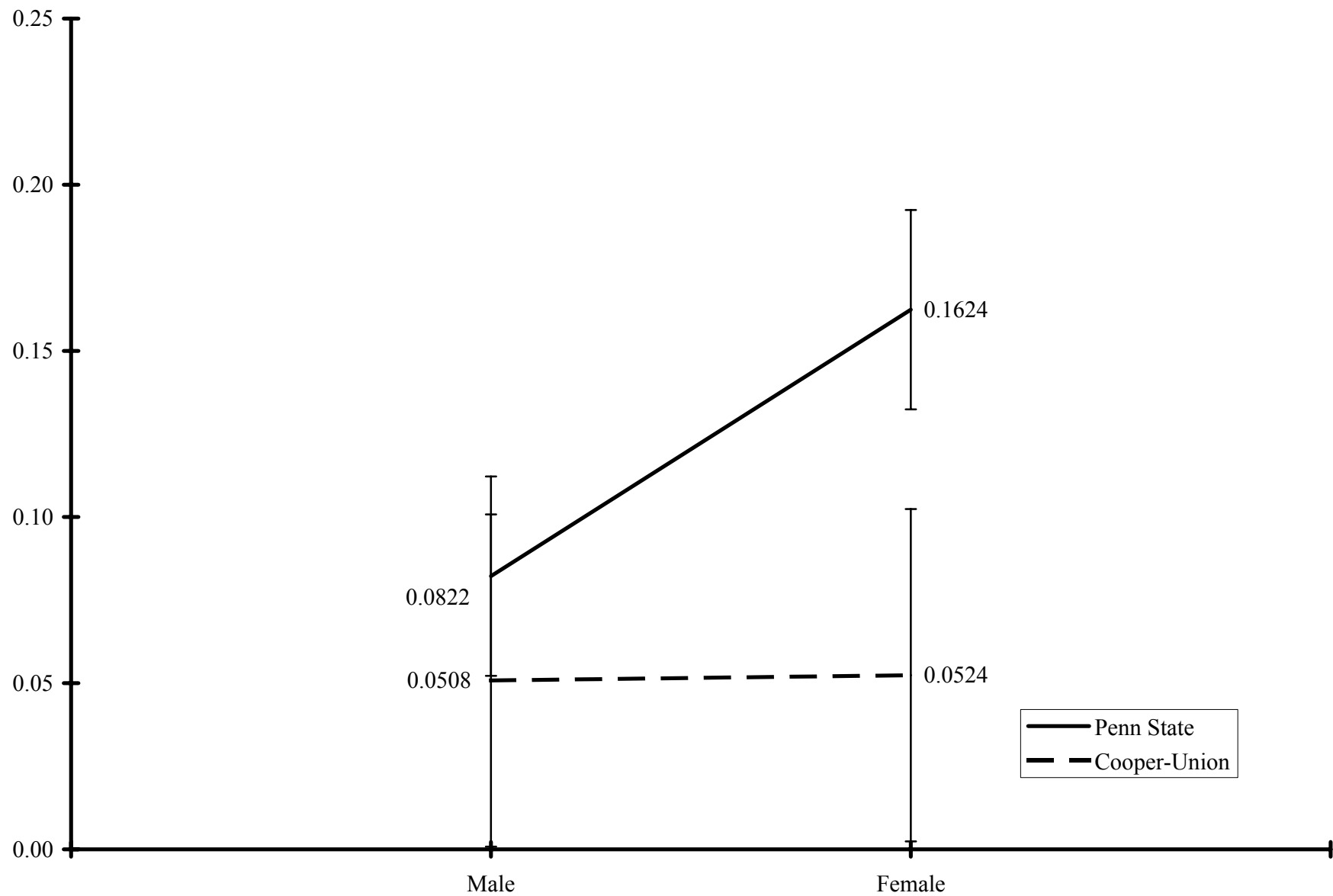


Figure 22. "High" Component Proportion Estimate Differences ($\pi_2^{(\text{Time } 1)} - \pi_2^{(\text{Time } 2)}$) for Penn State and Cooper-Union Males and Females on All items.

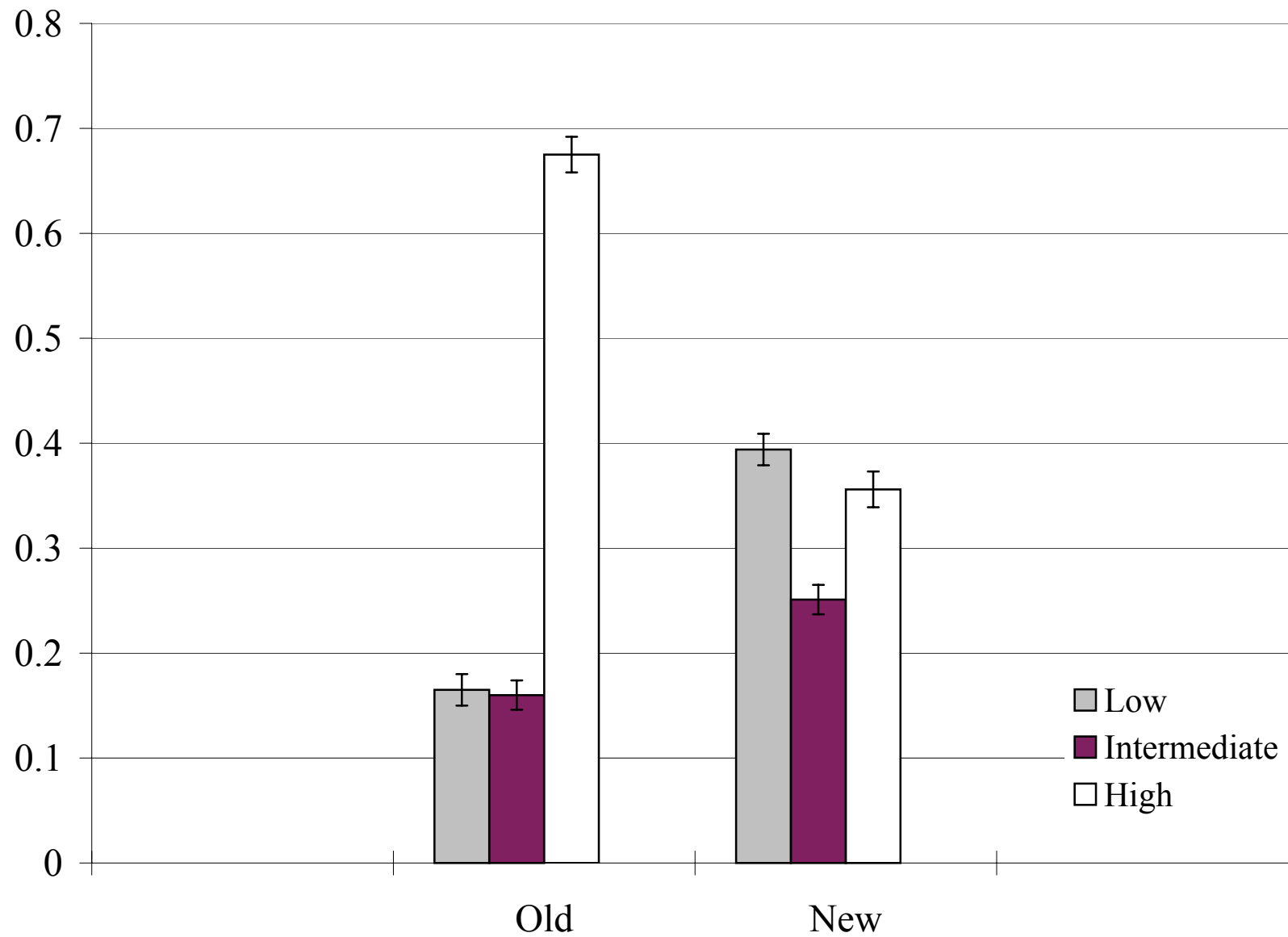


Figure 23. "Low," "Intermediate," and "High" Component Proportion Estimates (π 's) and Standard Errors for All Subjects on Old and New Items.

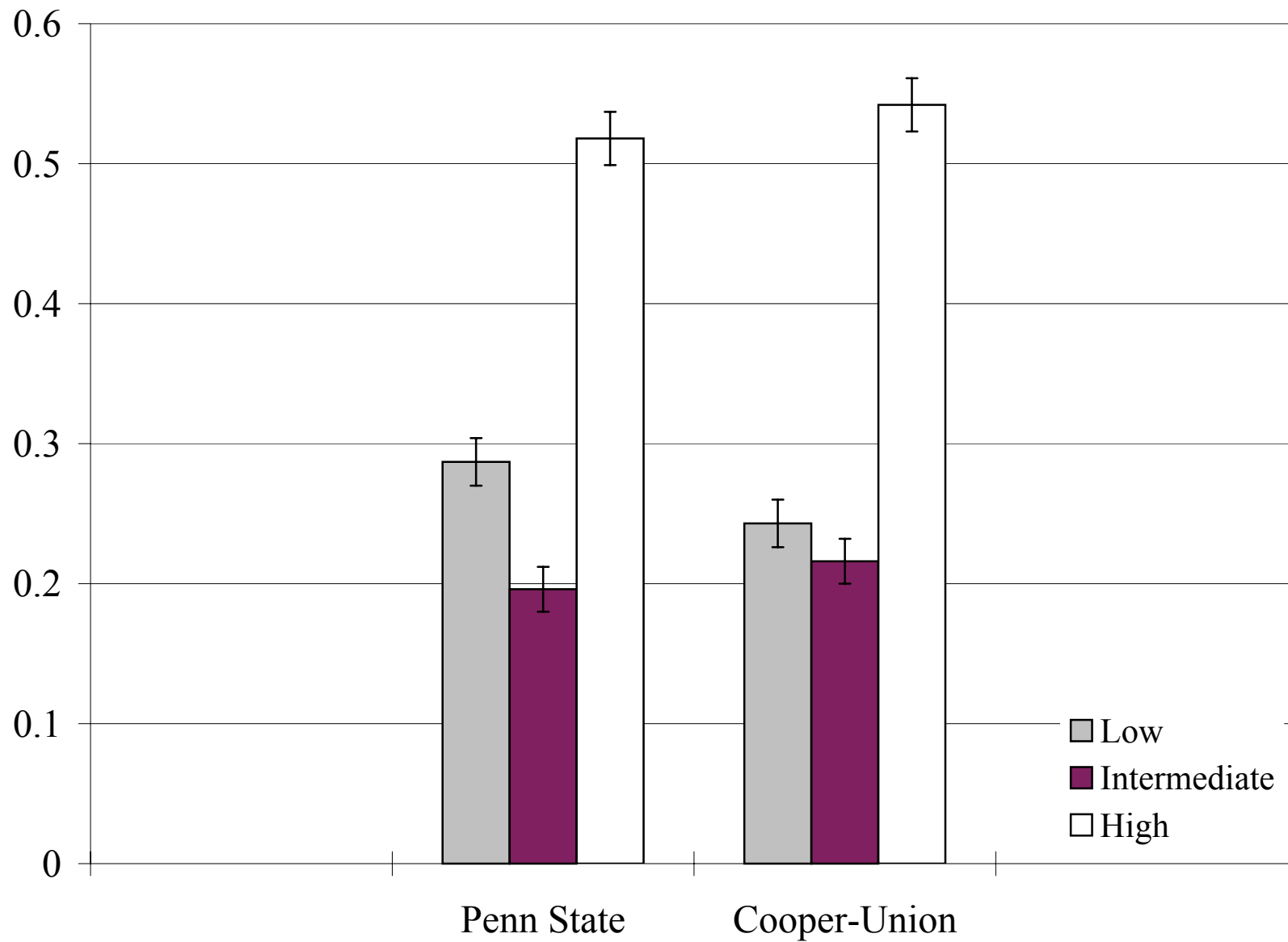


Figure 24. "Low," "Intermediate," and "High" Component Proportion Estimates (π 's) and Standard Errors for Penn State and Cooper-Union Subjects' Performance on All Items.

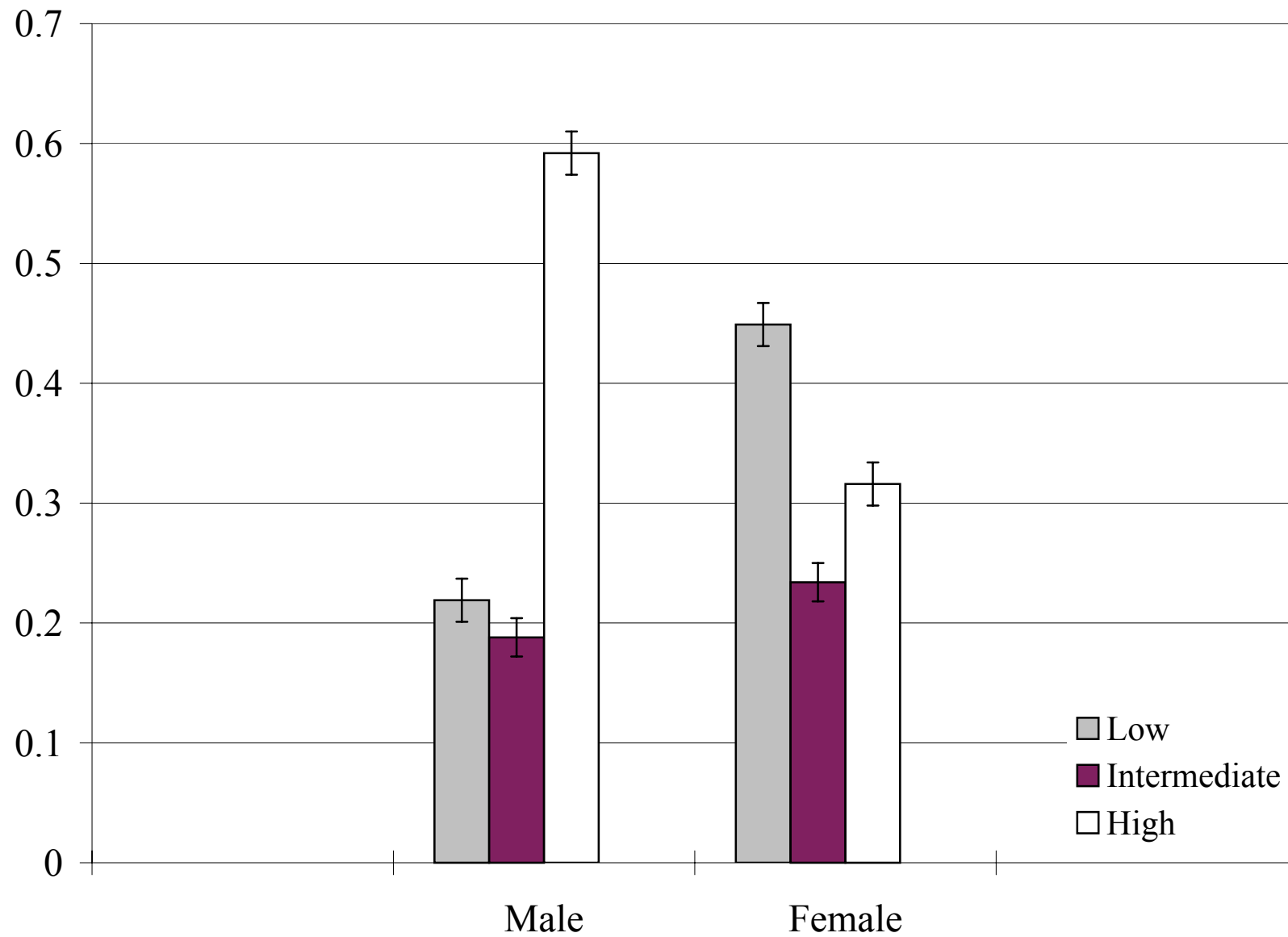


Figure 25. "Low," "Intermediate," and "High" Component Proportion Estimates (π 's) and Standard Errors for Males and Females' Performance on All Items.

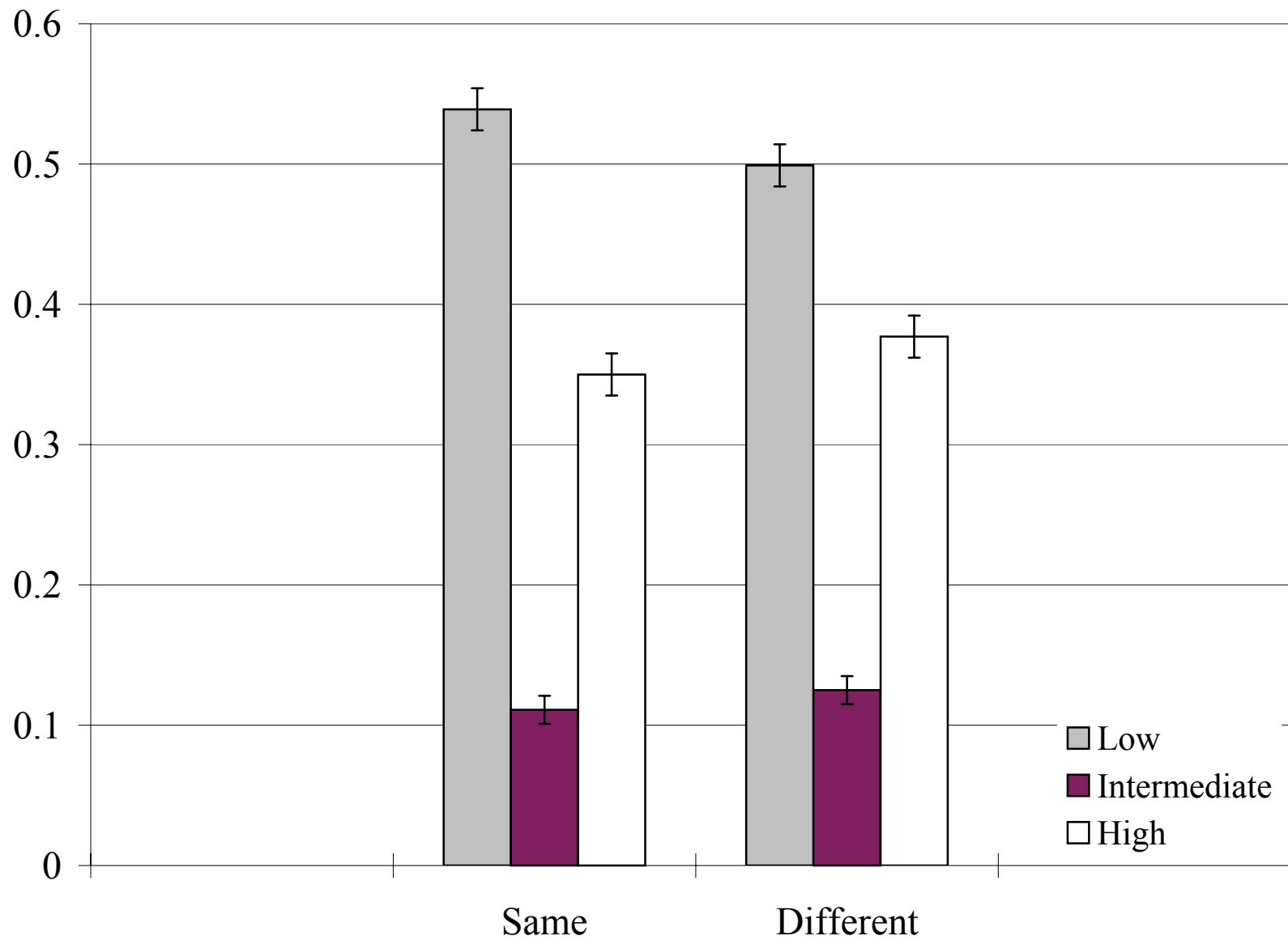


Figure 26. "Low," "Intermediate," and "High" Component Proportion Estimates (π 's) and Standard Errors for All Subjects' Performance on "Same" and "Different" Items.

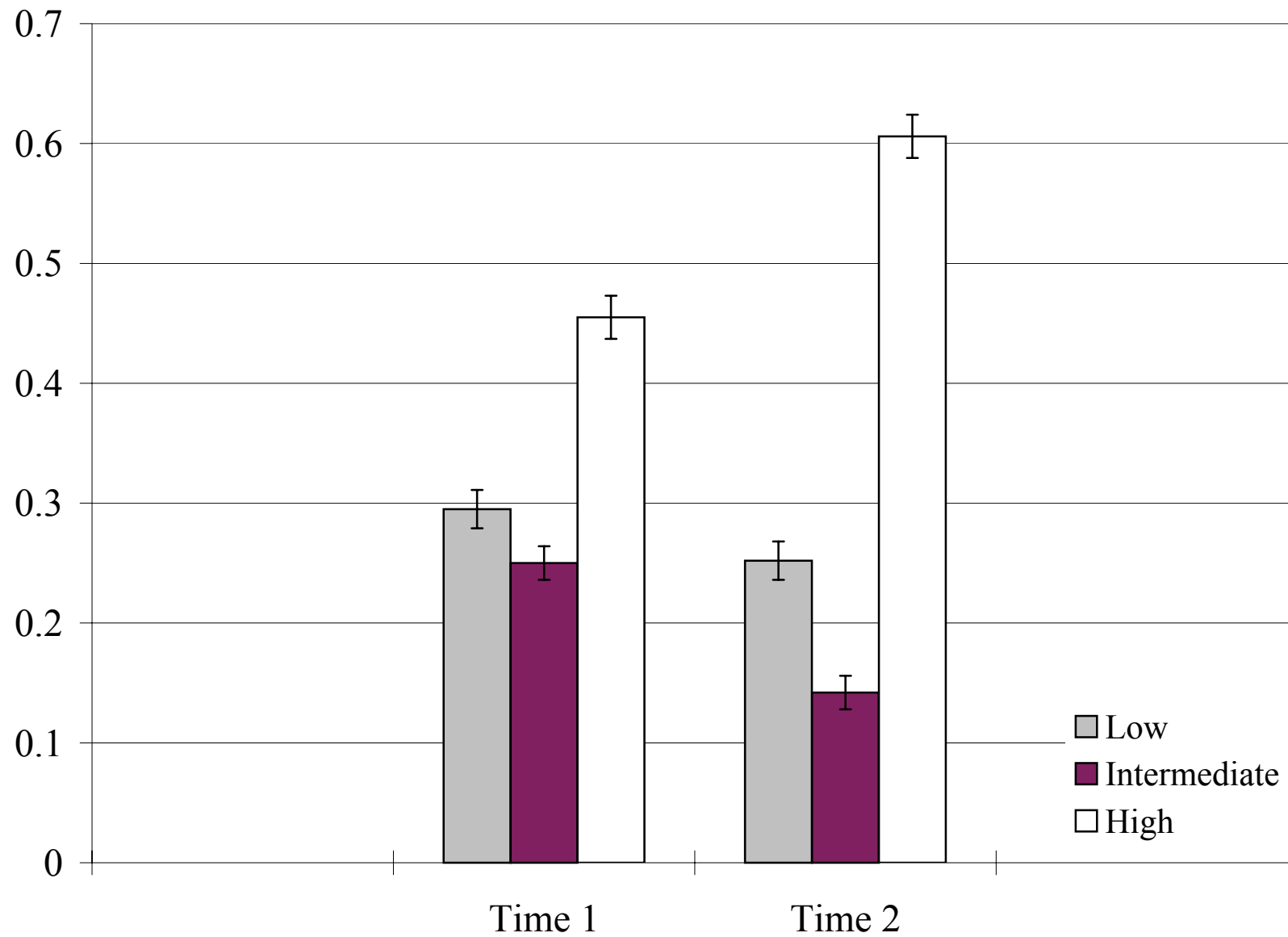


Figure 27. "Low," "Intermediate," and "High" Component Proportion Estimates (π 's) and Standard Errors for All Subjects' Performance on All Items at Time 1 and Time 2.

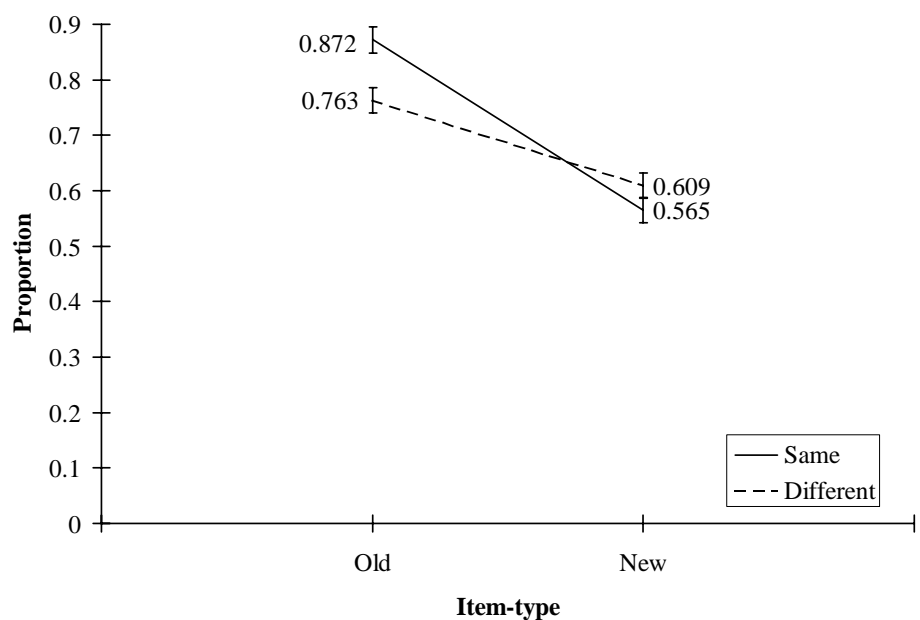
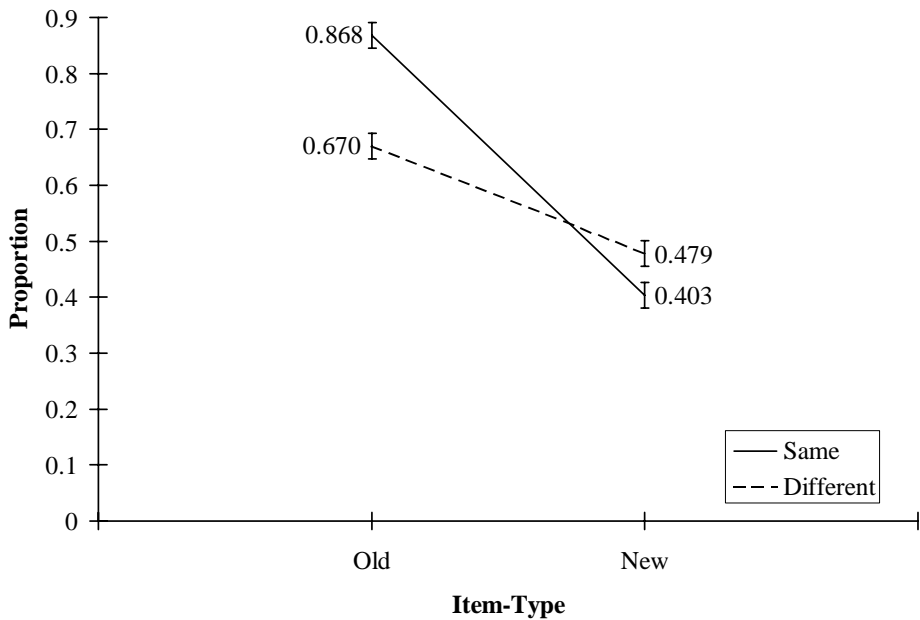
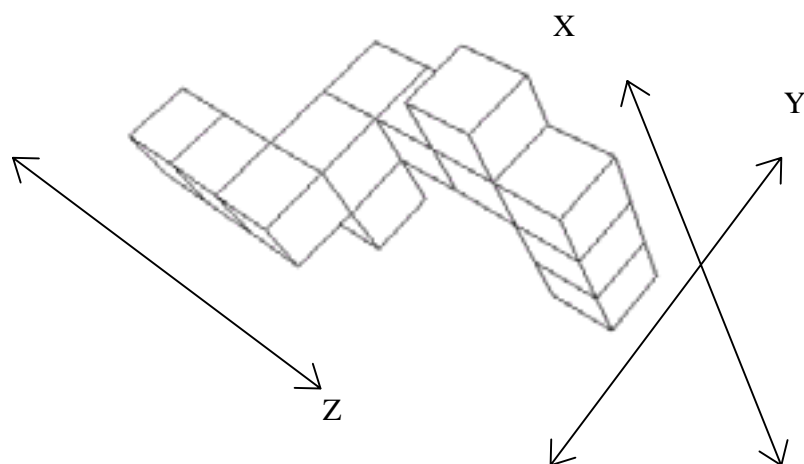
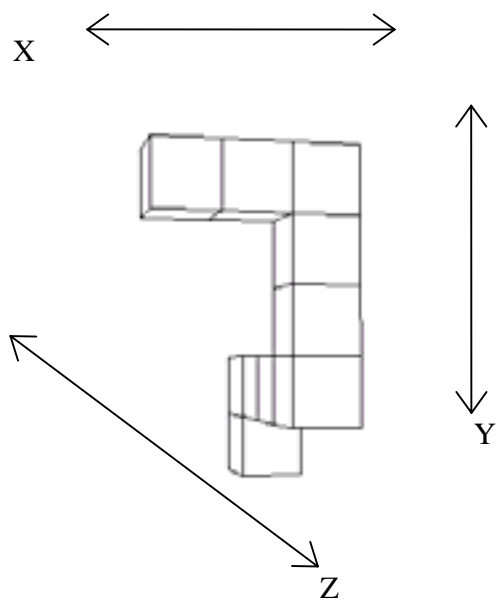


Figure 28. “High” component mixing proportion Item-type by Item-status Interactions at Time 1 and Time 2.



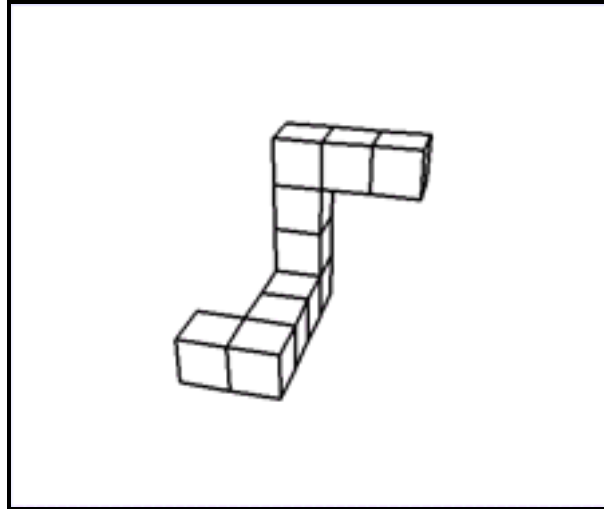
New Item, Target III



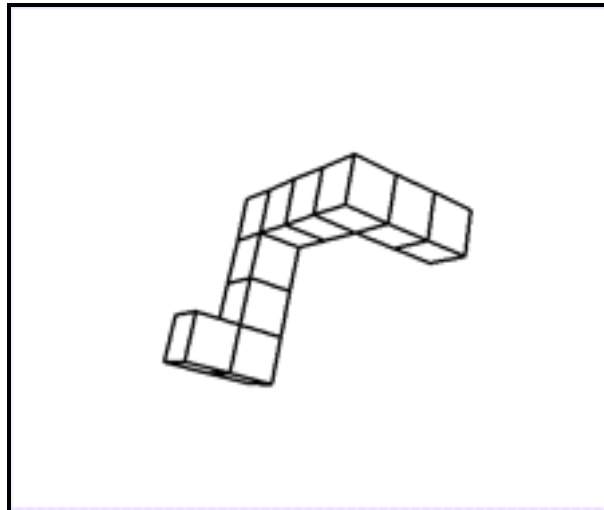
Old Item, Target IX

Figure 29. X, Y, and Z axes of Mental Rotation Objects.

Panel A. Initial
Orientation of a Mental
Rotation Object.



Panel B. $45^\circ X, -45^\circ Z$
Rotation about the
Centroid from the Initial
Orientation.



Panel C. $-45^\circ Z, 45^\circ X$
Rotation about the
Centroid from the Initial
Orientation.

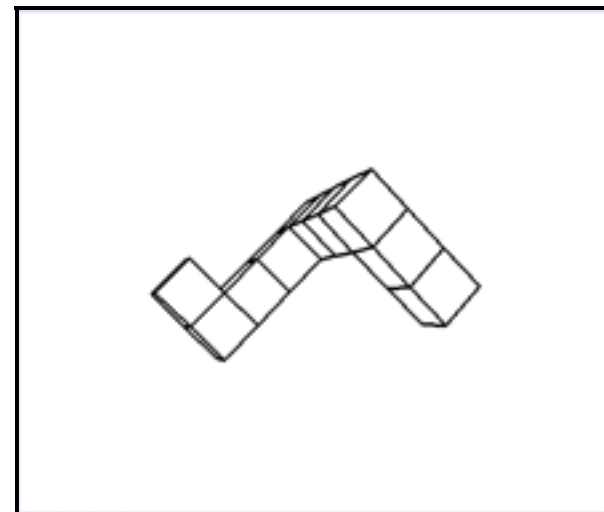


Figure 30. Non-commutative Property of Sequential Rotations About Two Axes.

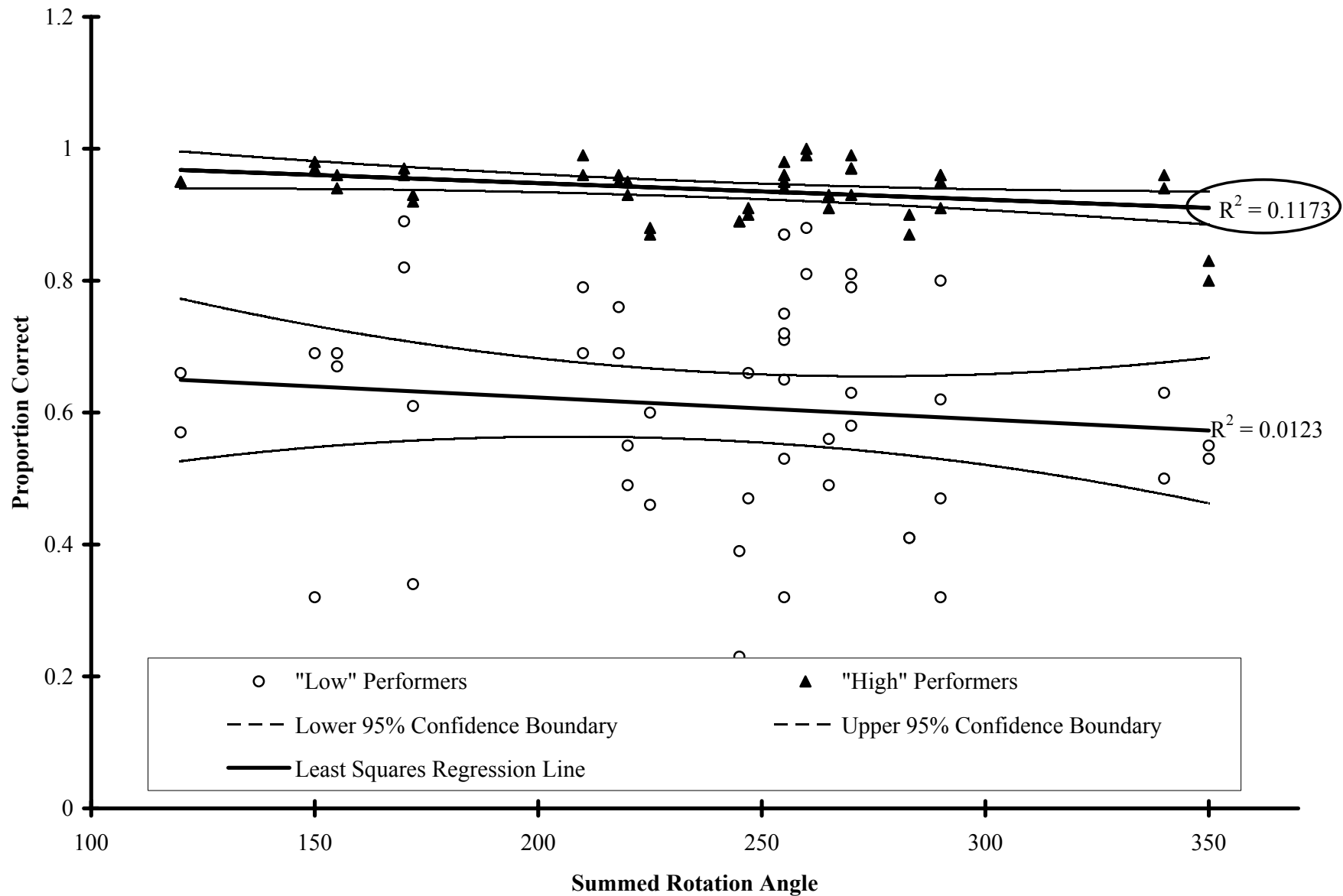


Figure 31. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Summed Angle of Rotation for "High" and "Low" Performers. Encircled r^2 's are Significantly Different than Zero.

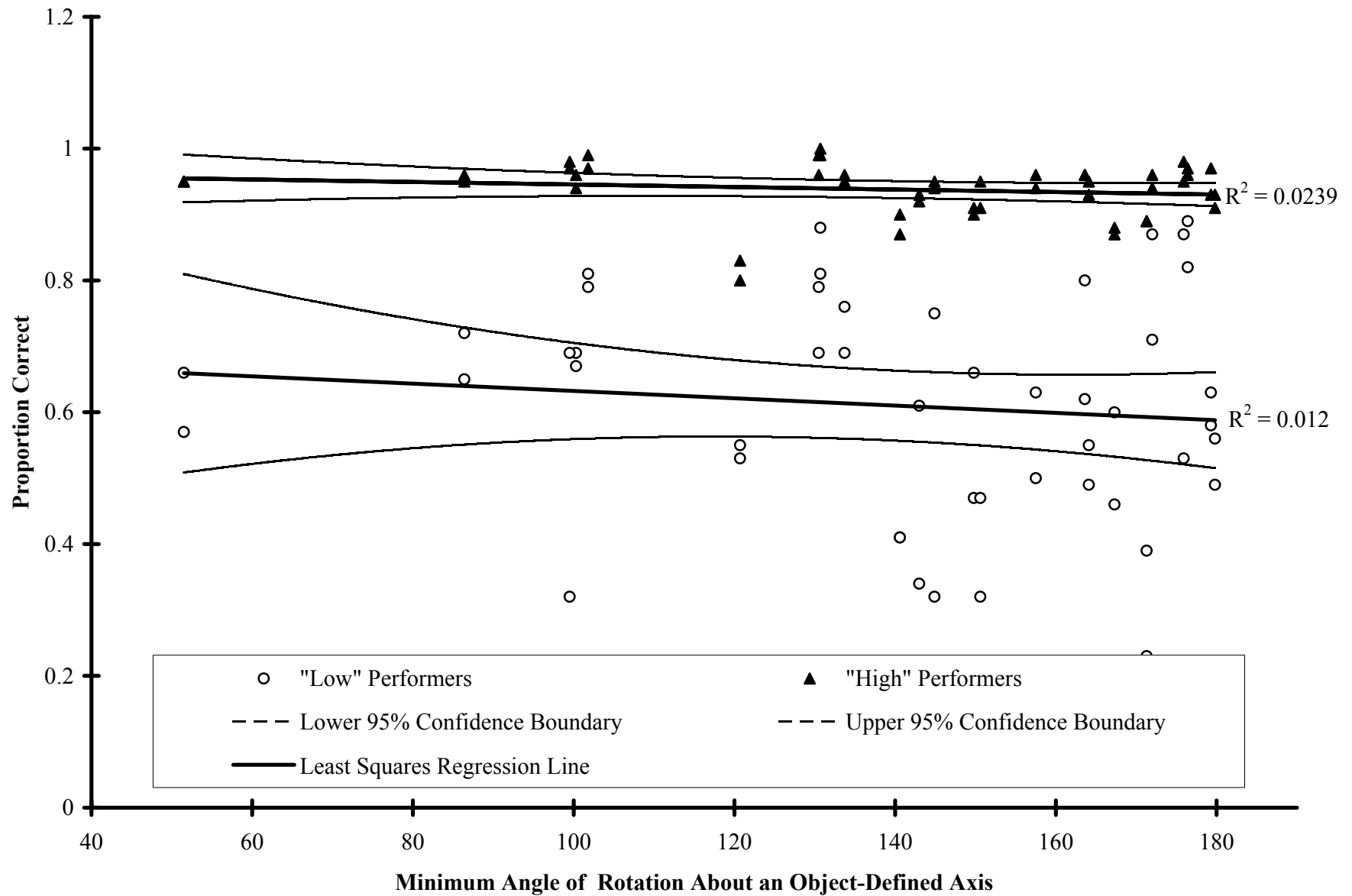


Figure 32. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Object-Defined Minimum Angle of Rotation for "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

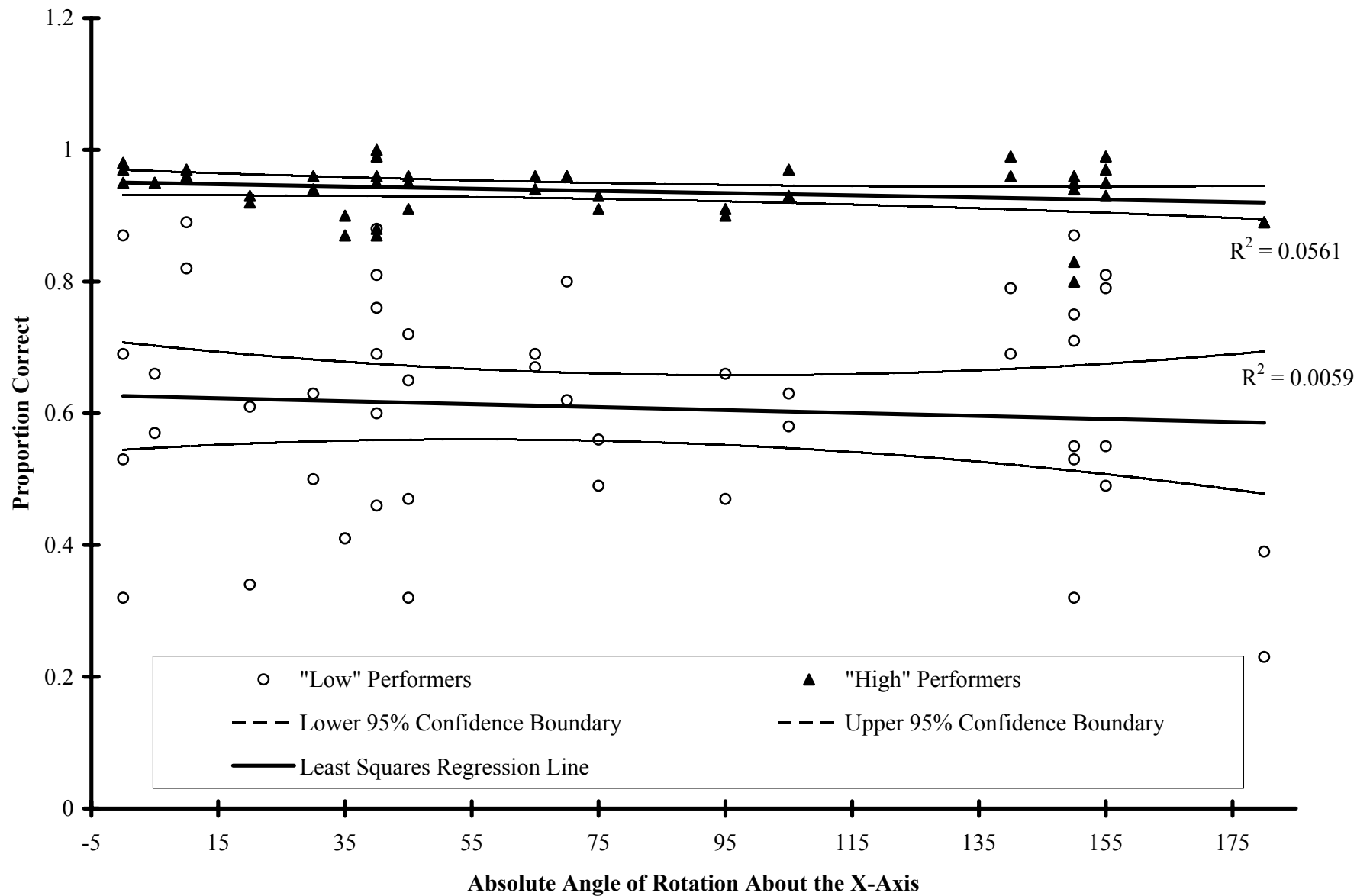


Figure 33. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation on the X-axis for "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

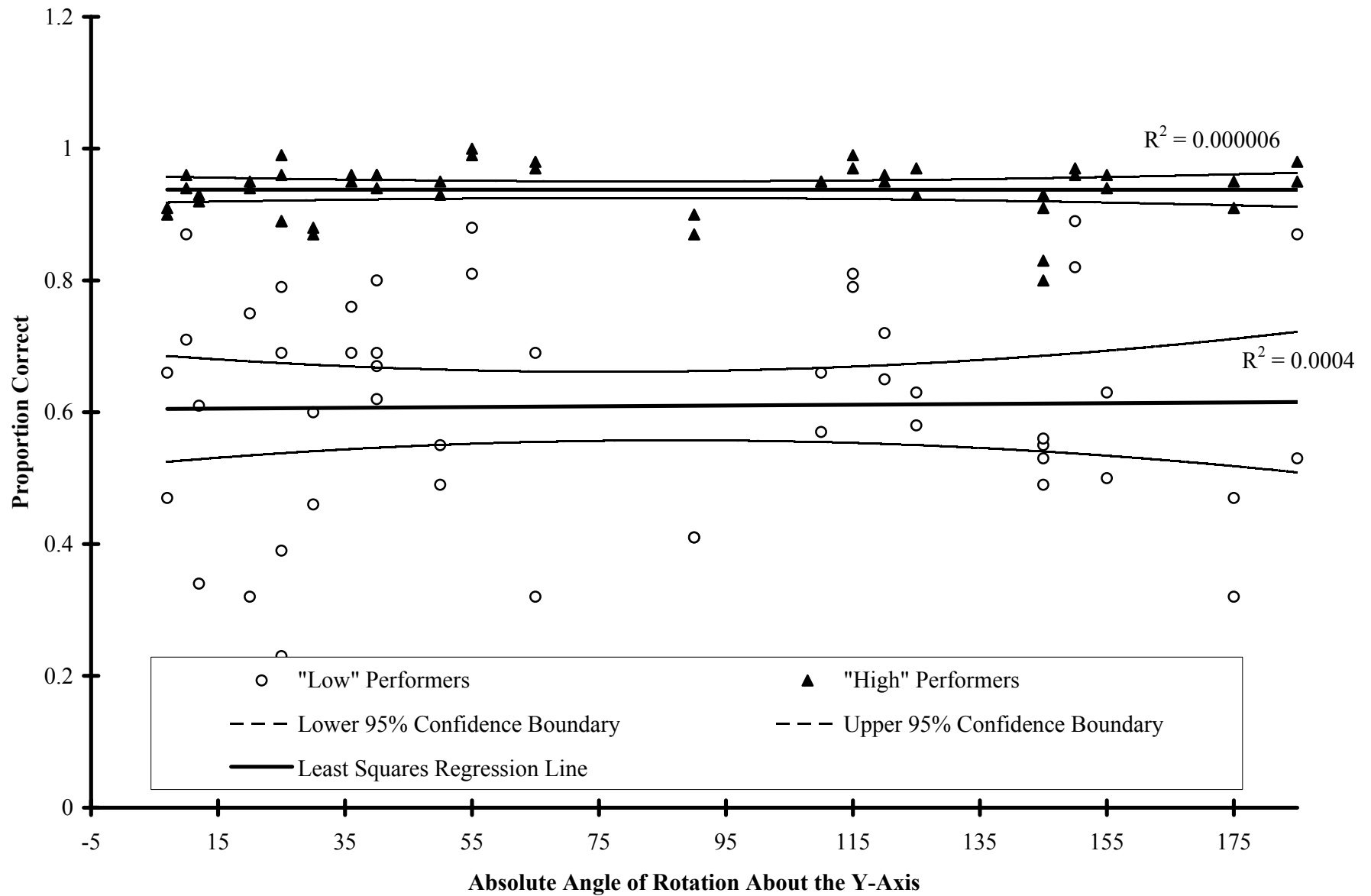


Figure 34. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation on the Y-Axis for "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

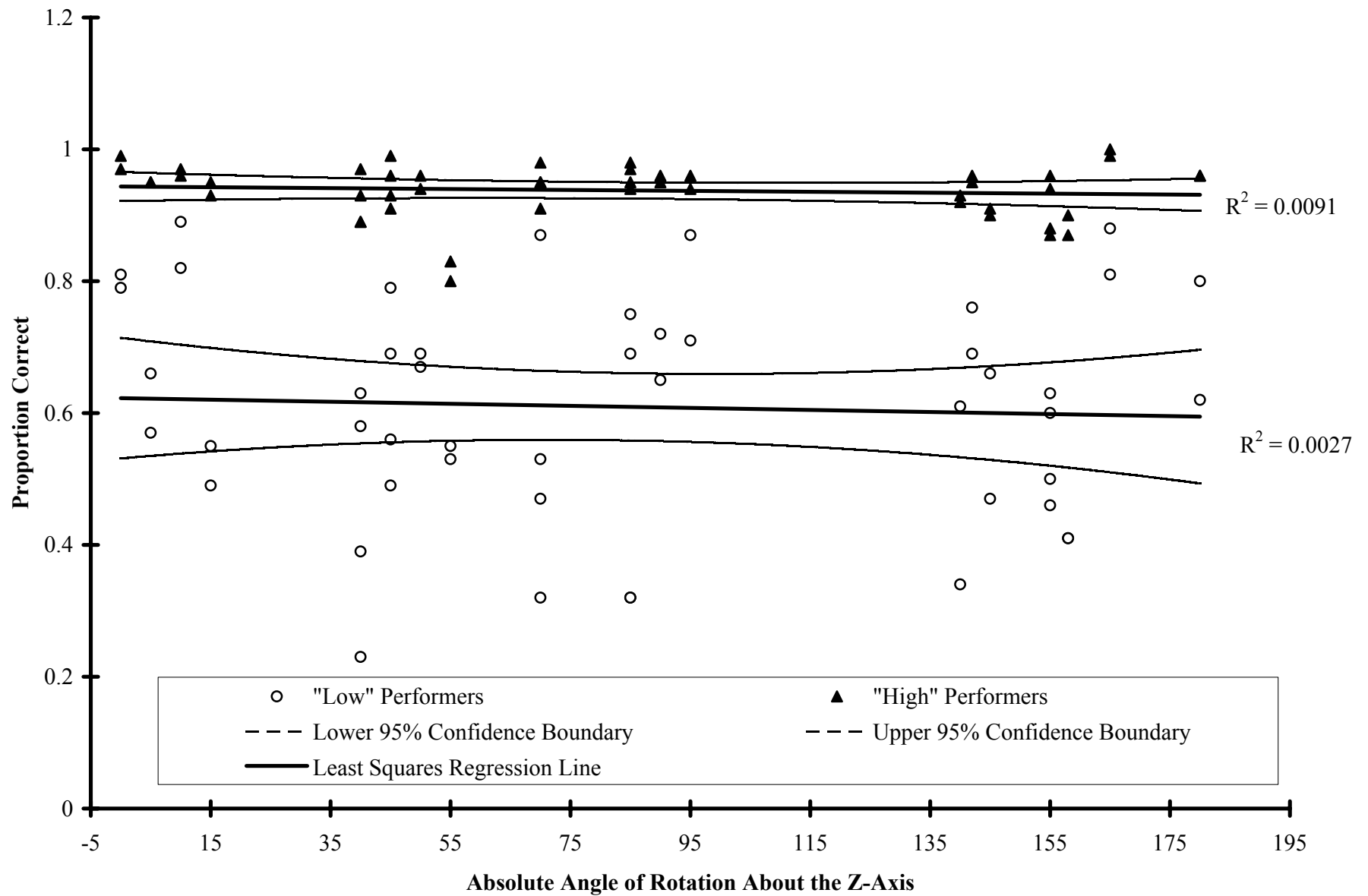


Figure 35. Scatterplot, Regression Lines, and Confidence Bands for Accuracy Vs. Angle of Rotation About the Z-axis for "High" and "Low" Performers. Encircled r^2 's are Significantly Different than Zero.

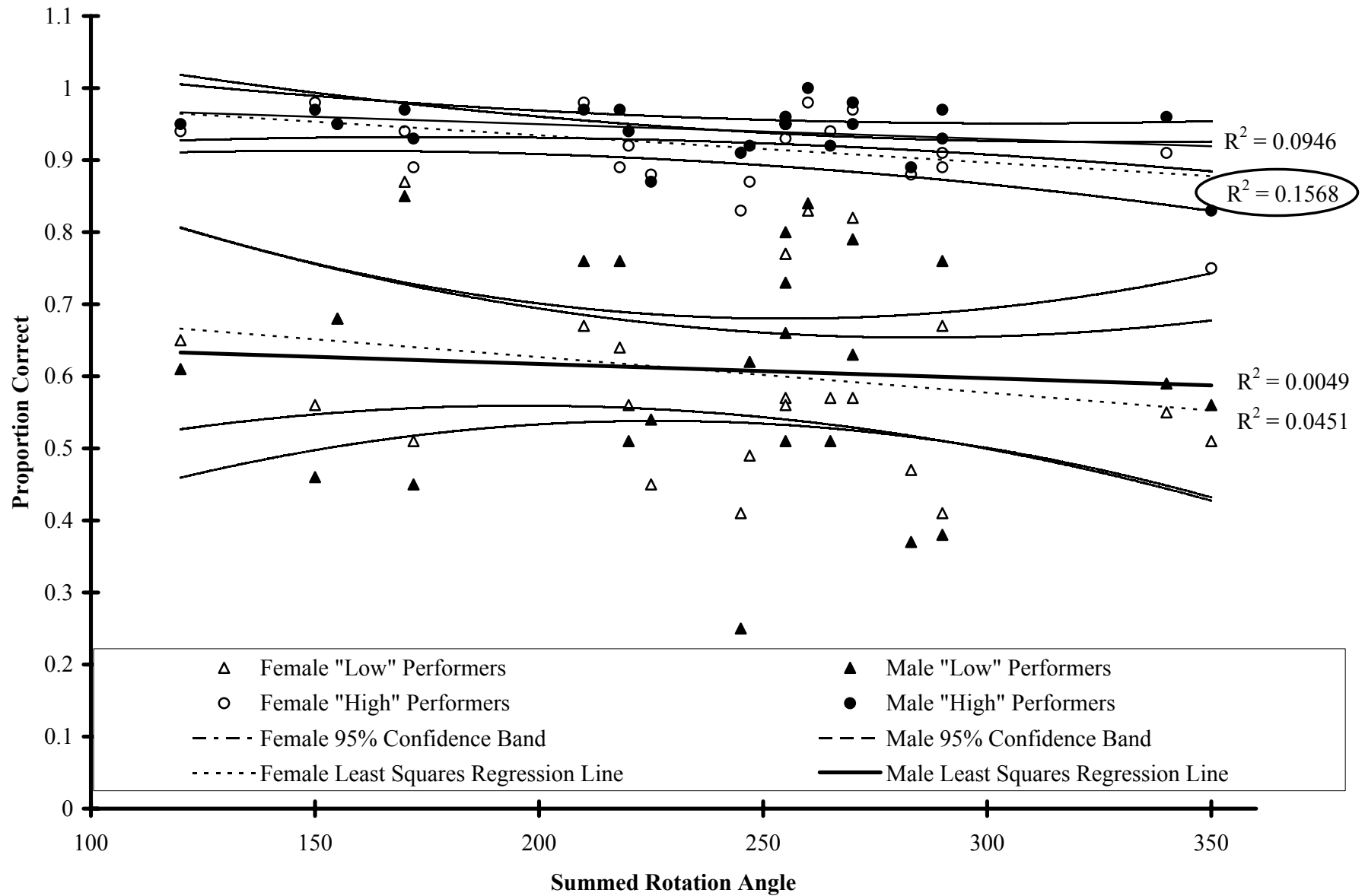


Figure 36. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Summed Angle of Rotation for Male and Female "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

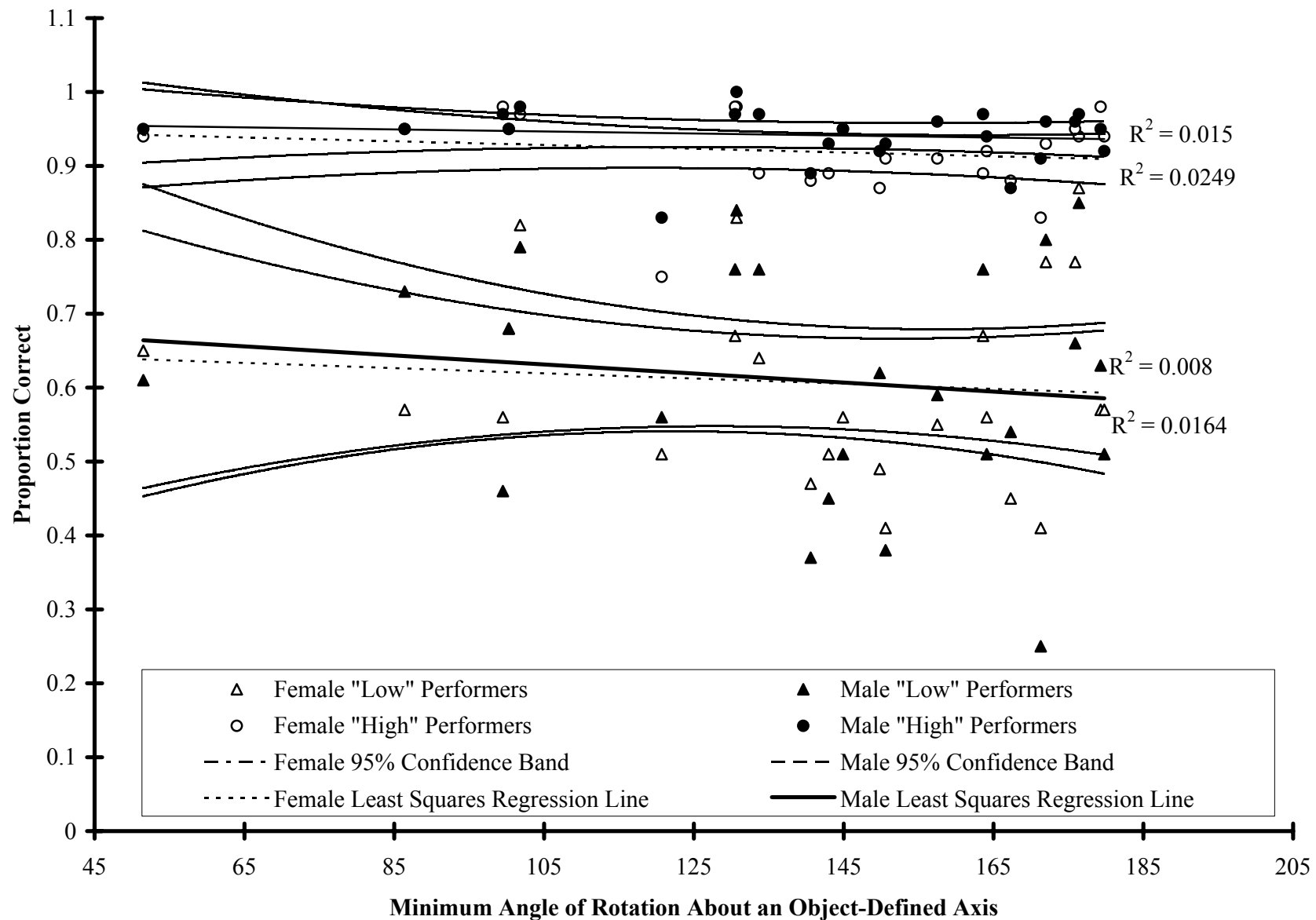


Figure 37. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Object-Defined Minimum Angle of Rotation for Male and Female "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

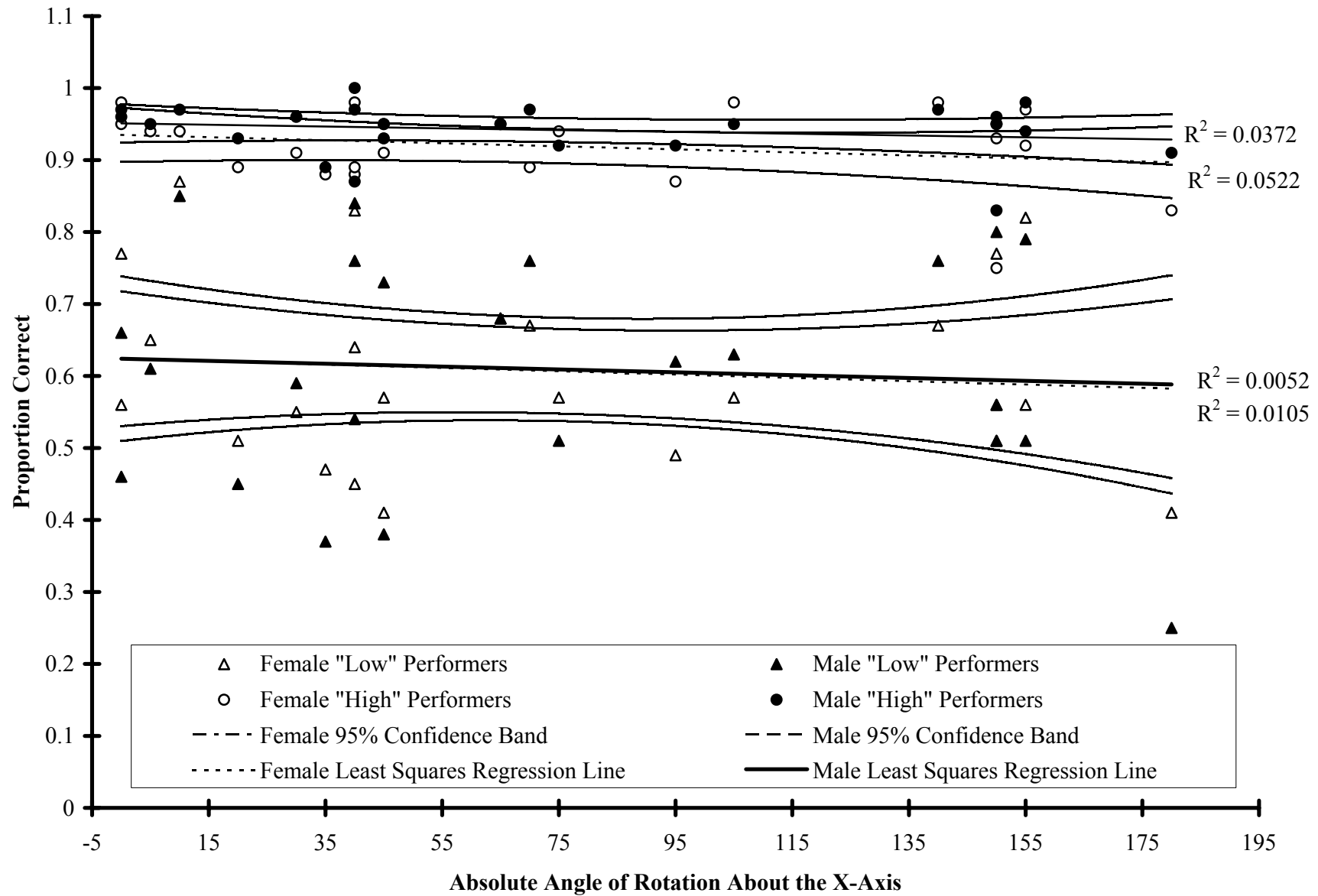


Figure 38. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Rotation on the X-Axis for Male and Female "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

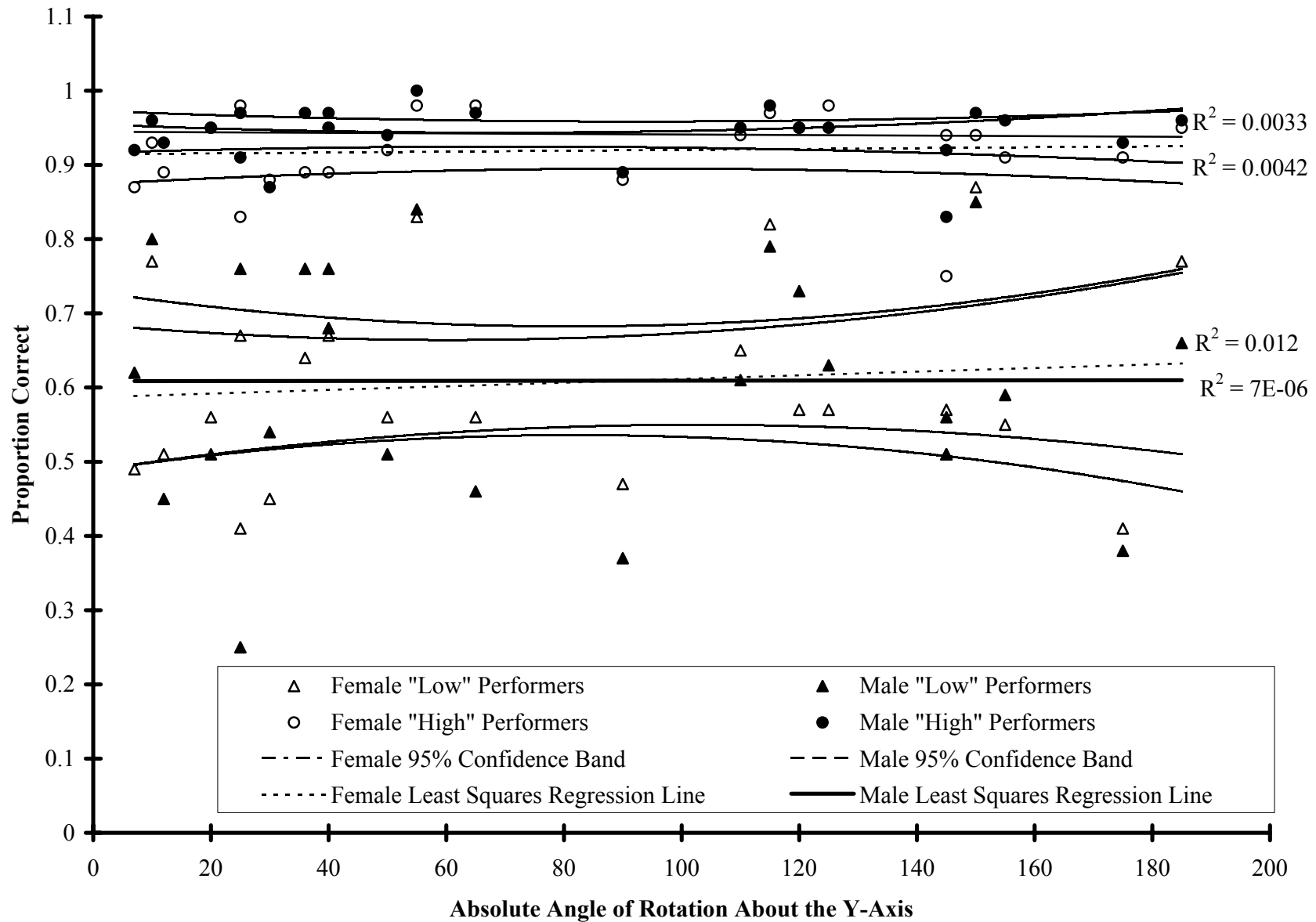


Figure 39. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Angle of Rotation About the Y-Axis for Male and Female "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

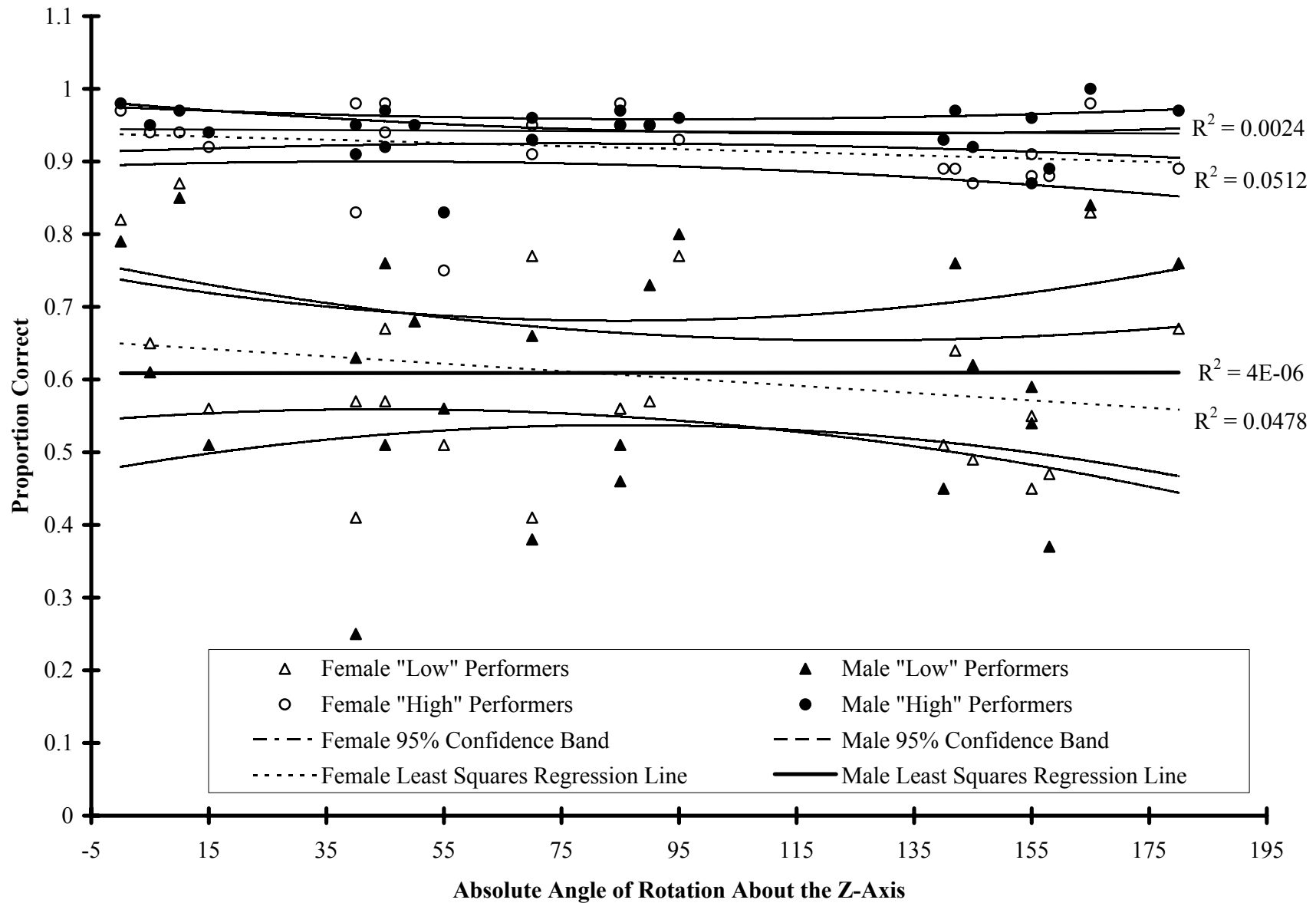


Figure 40. Scatterplot, Regression Lines, and 95% Confidence Bands for Accuracy Vs. Angle of Rotation About the Z-axis for Male and Female "High" and "Low" Performers. Encircled r^2 's are Significantly Different from Zero.

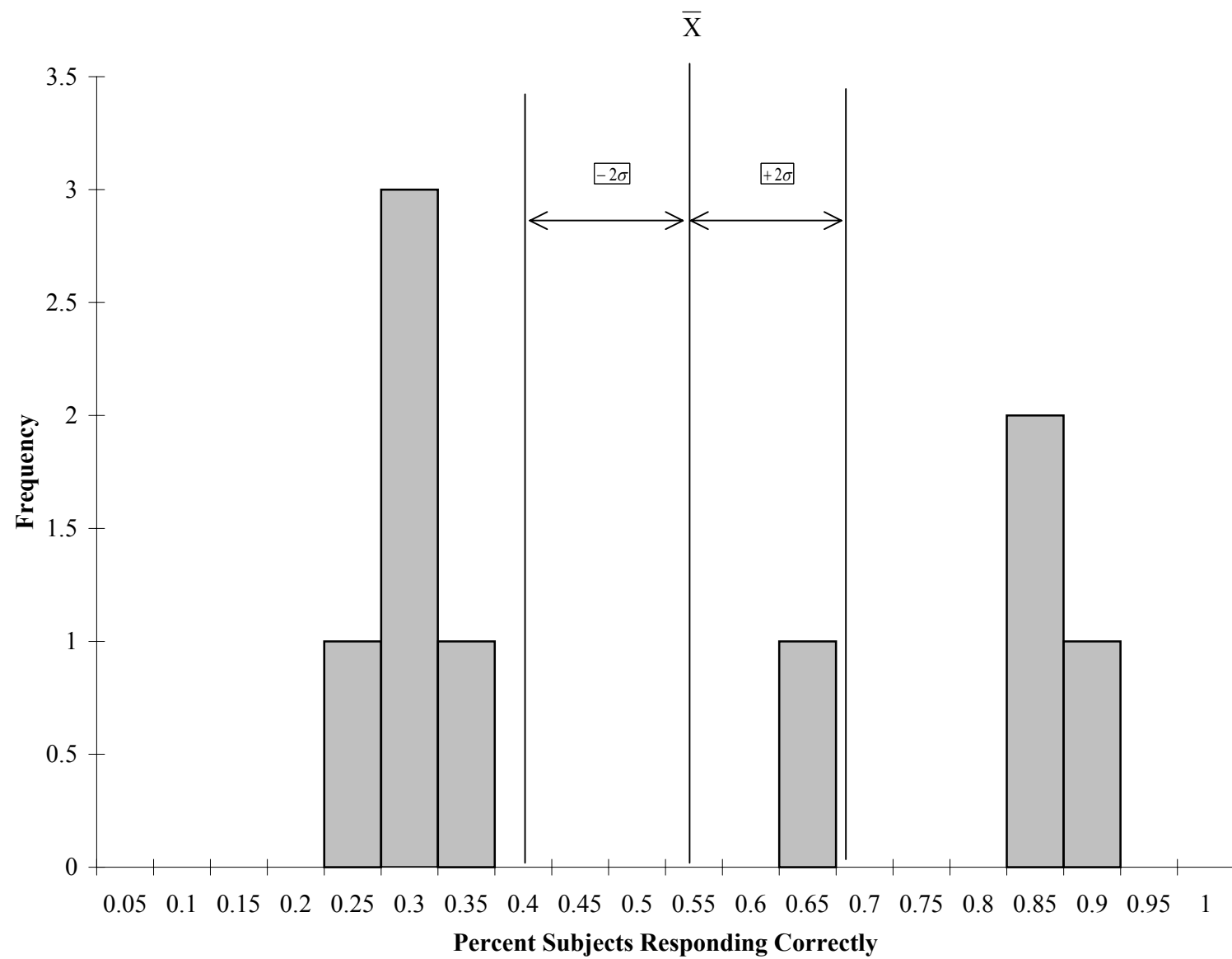


Figure 41. "Low" Component Success Rates on Old-same Items at Time 2.

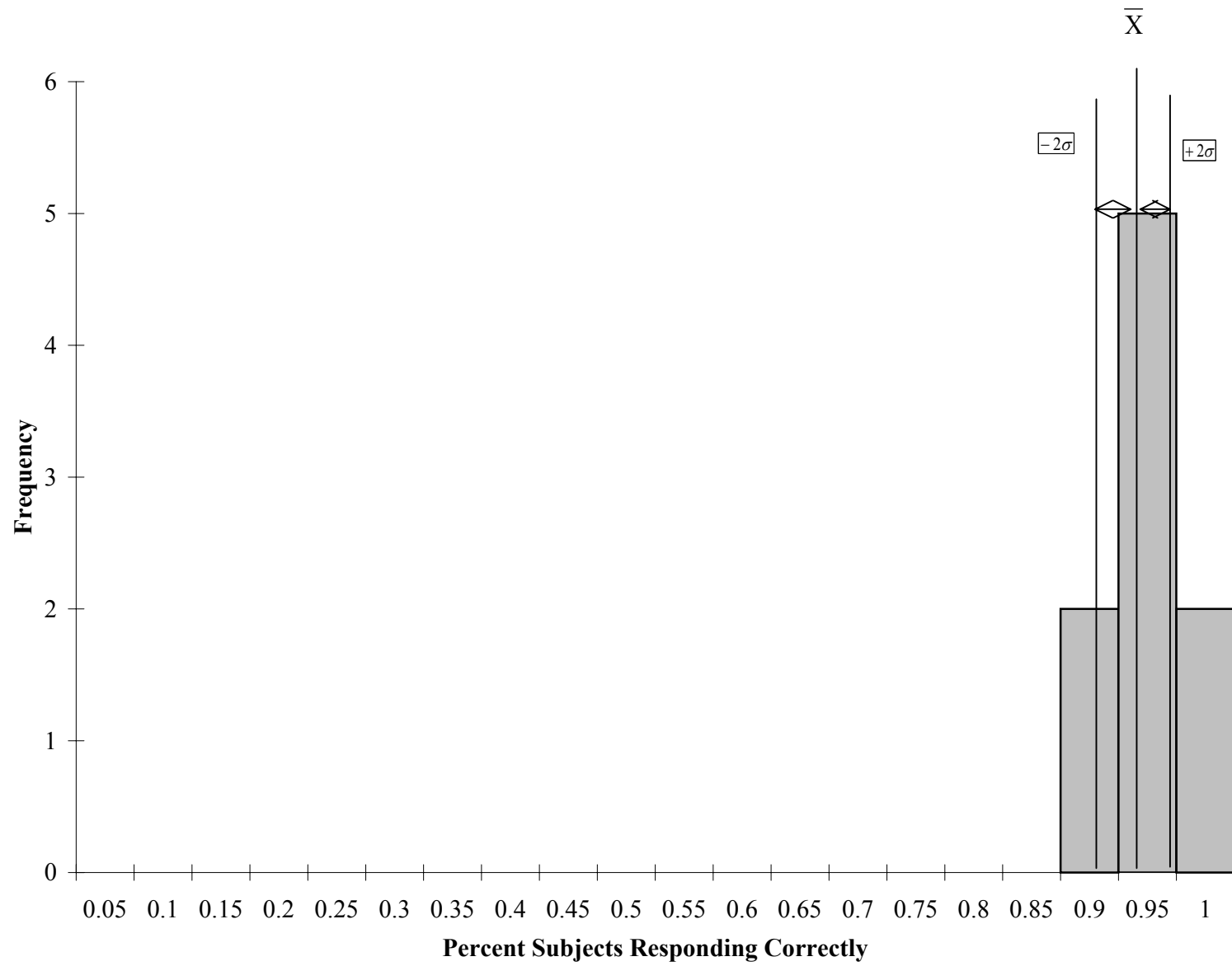


Figure 42. "High" Component Success Rates on Old-Same Items at Time 2.

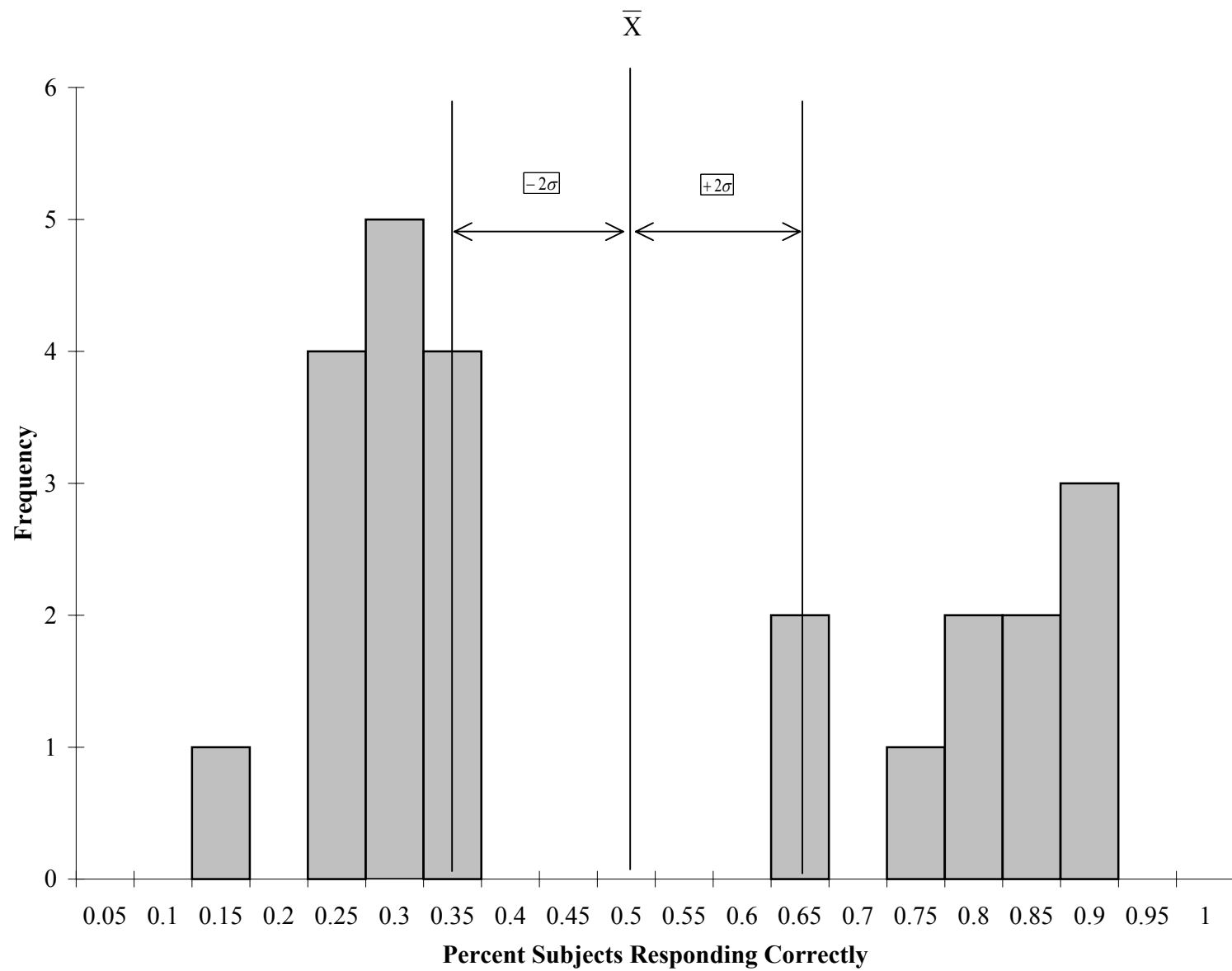


Figure 43. "Low" Component Success Rates on Old Items at Time 2.

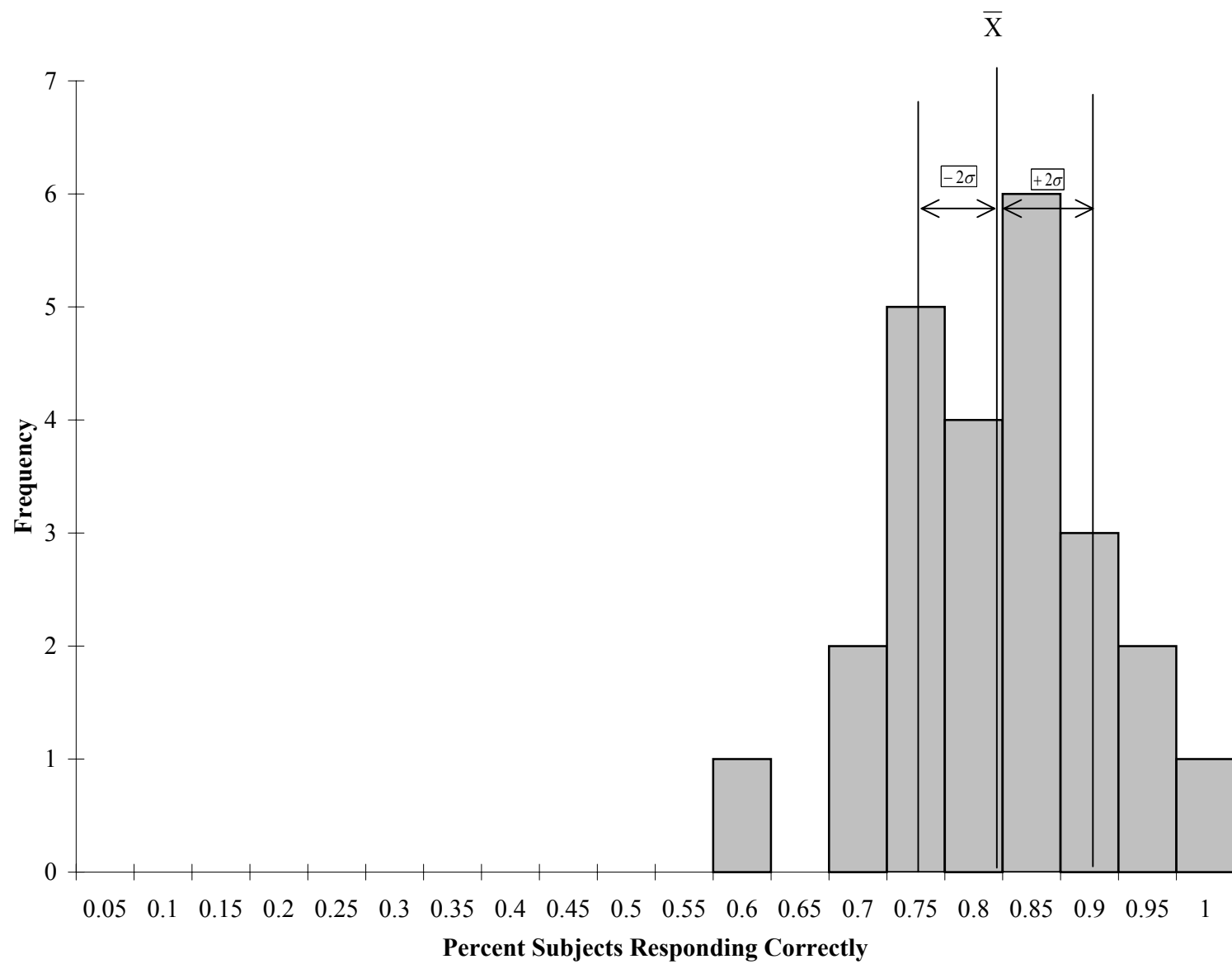


Figure 44. "Intermediate" Component Success Rates on Old Items at Time 2.

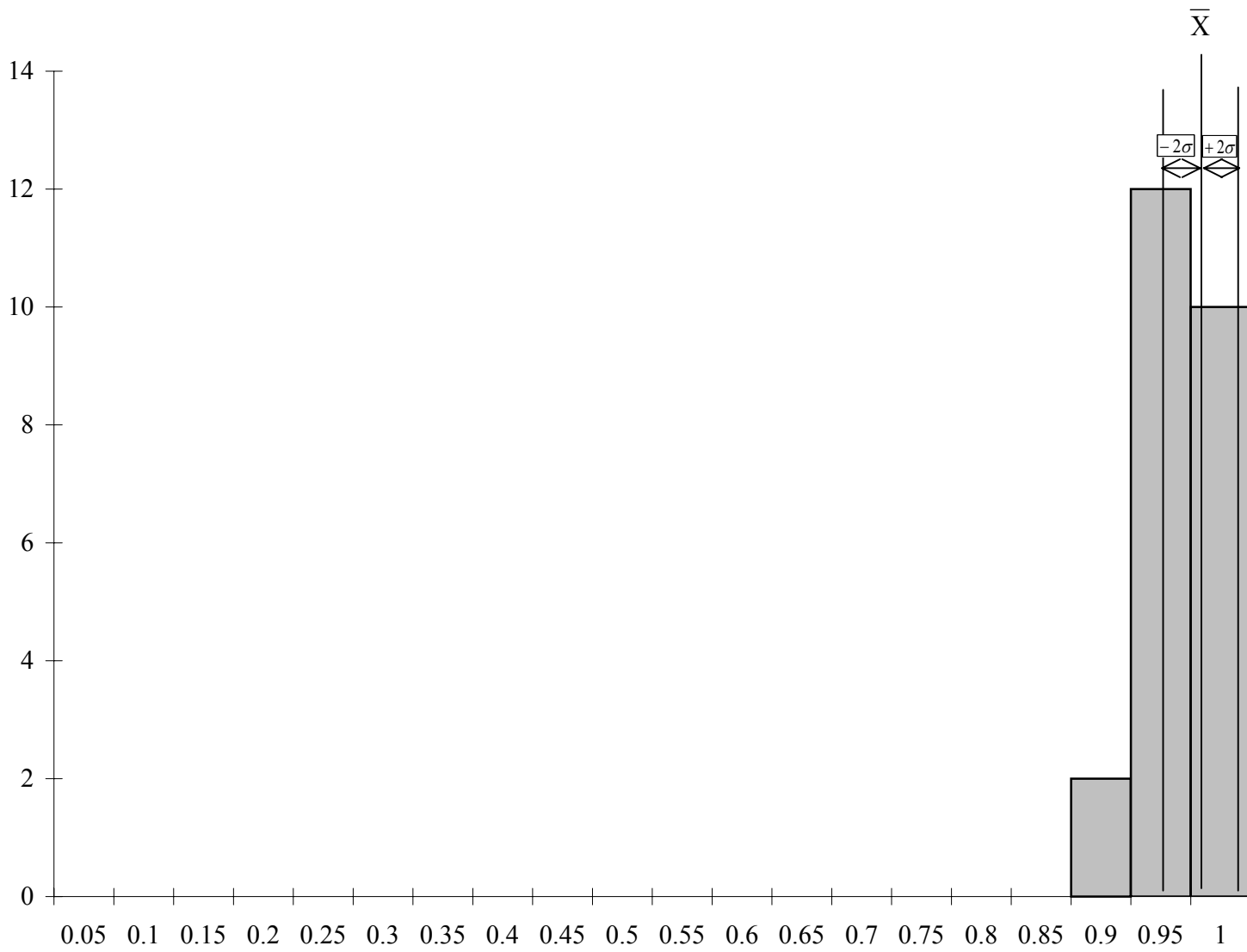


Figure 45. "High" Component Success Rates on Old Items at Time 2.

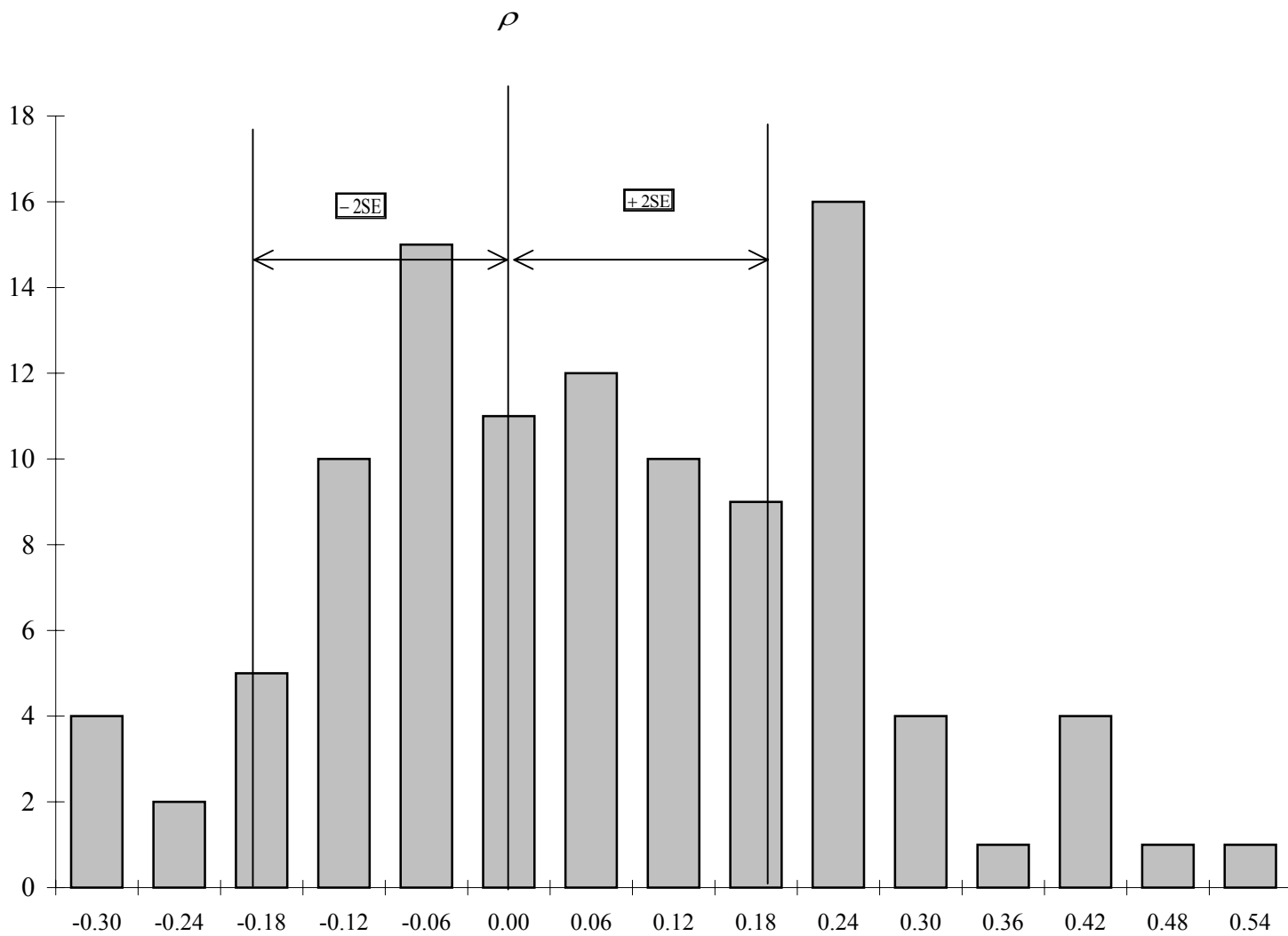


Figure 46. Inter-item Correlations on Old-different Items at Time 2 for the "Low" Component.

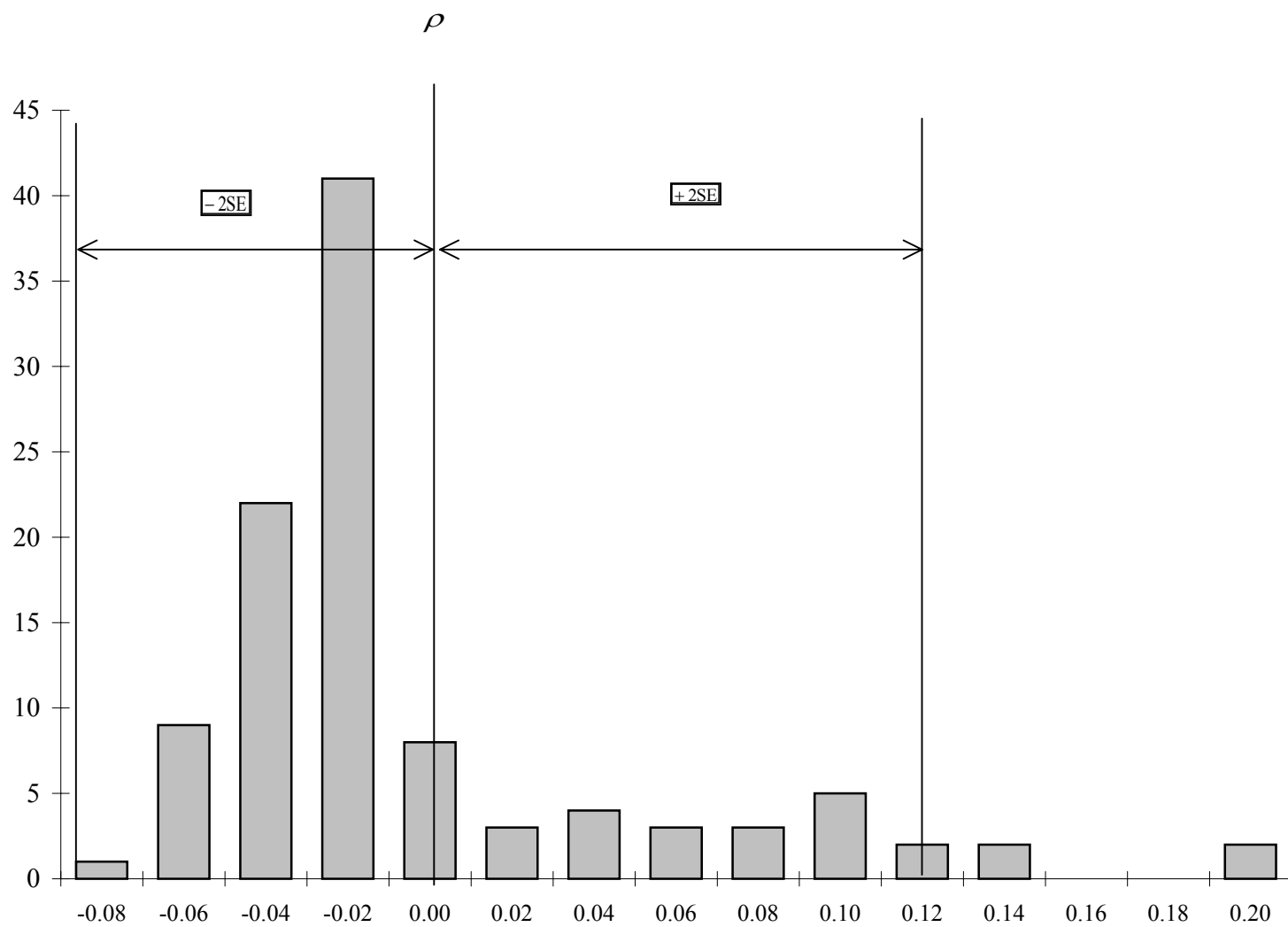


Figure 47. Inter-item Correlations on Old-different Items at Time 2 for the "High" Component.

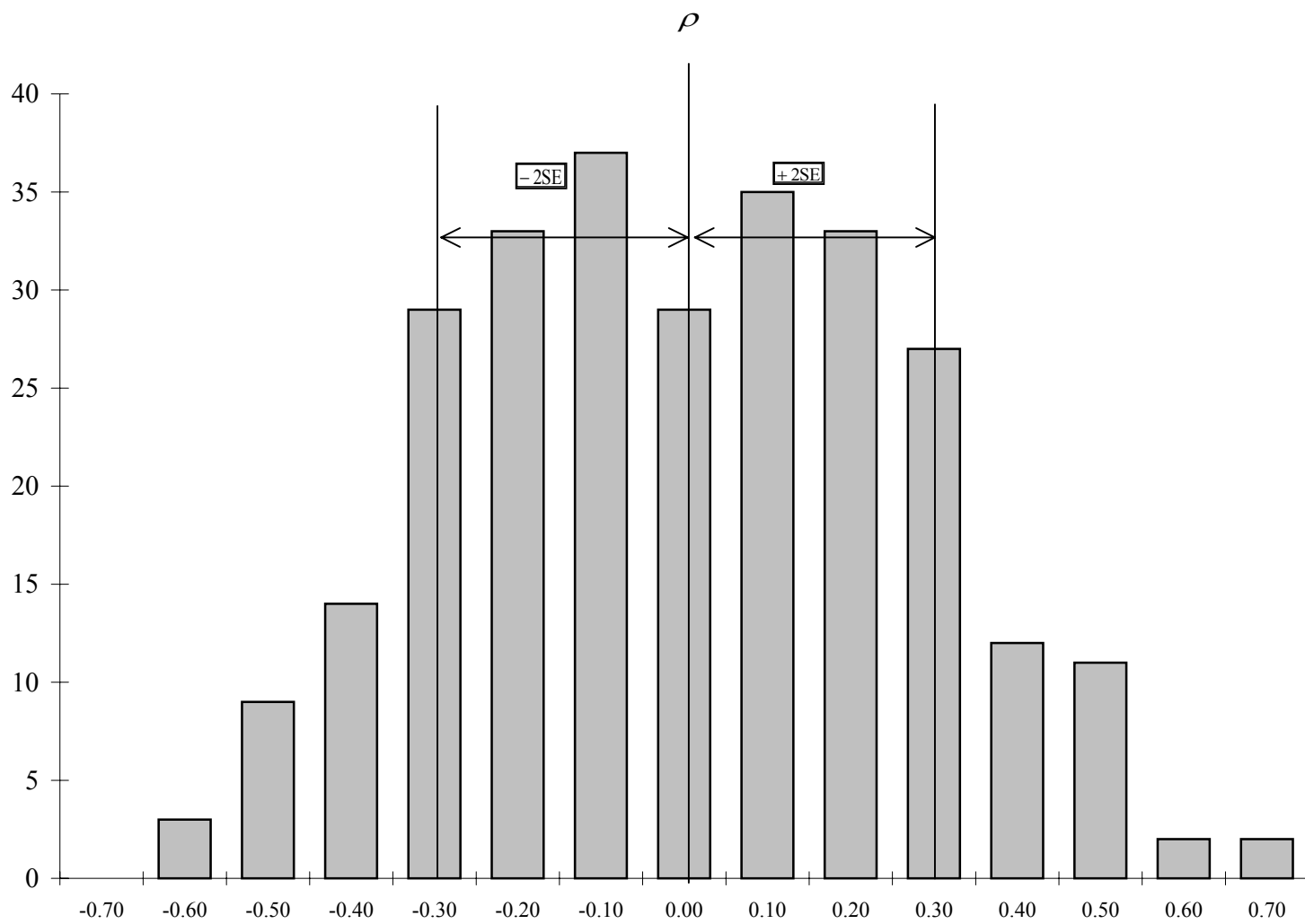


Figure 48. Inter-item Correlations on Old Items at Time 2 for the "Low" Component.

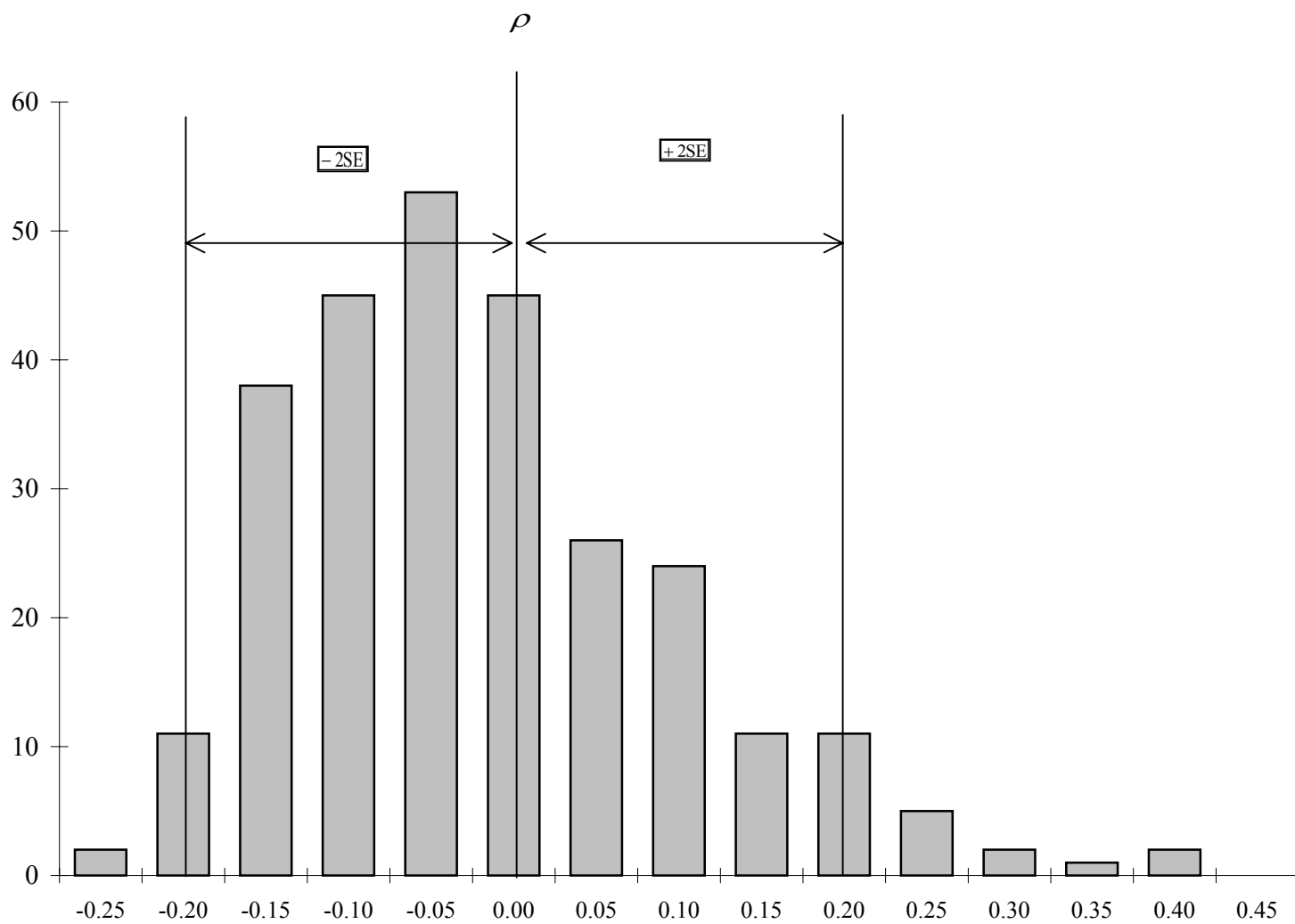


Figure 49. Inter-item Correlations on Old Items at Time 2 for the "Intermediate" Component.

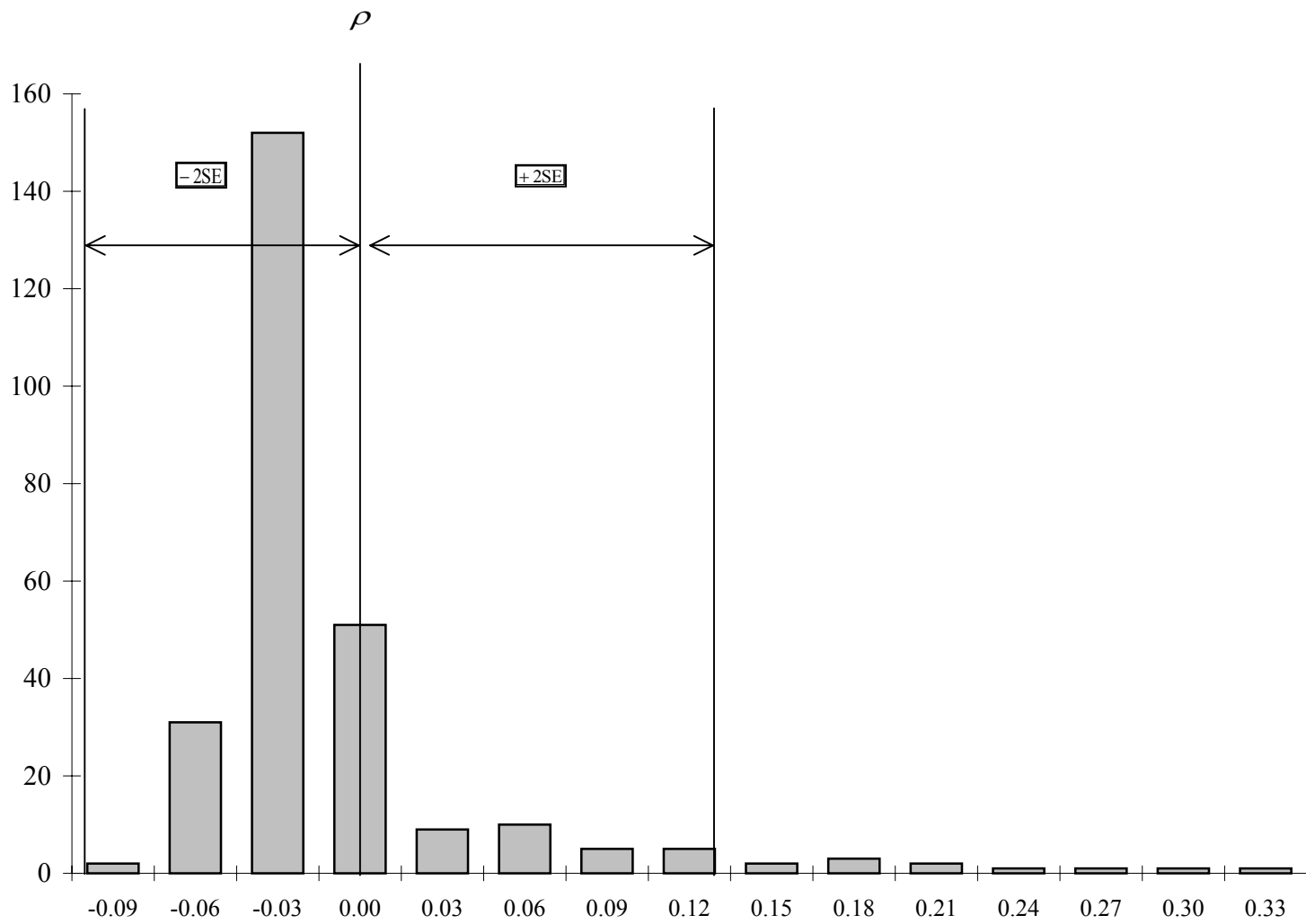


Figure 50. Inter-item Correlations on Old-items at Time 2 for the "High" Component.

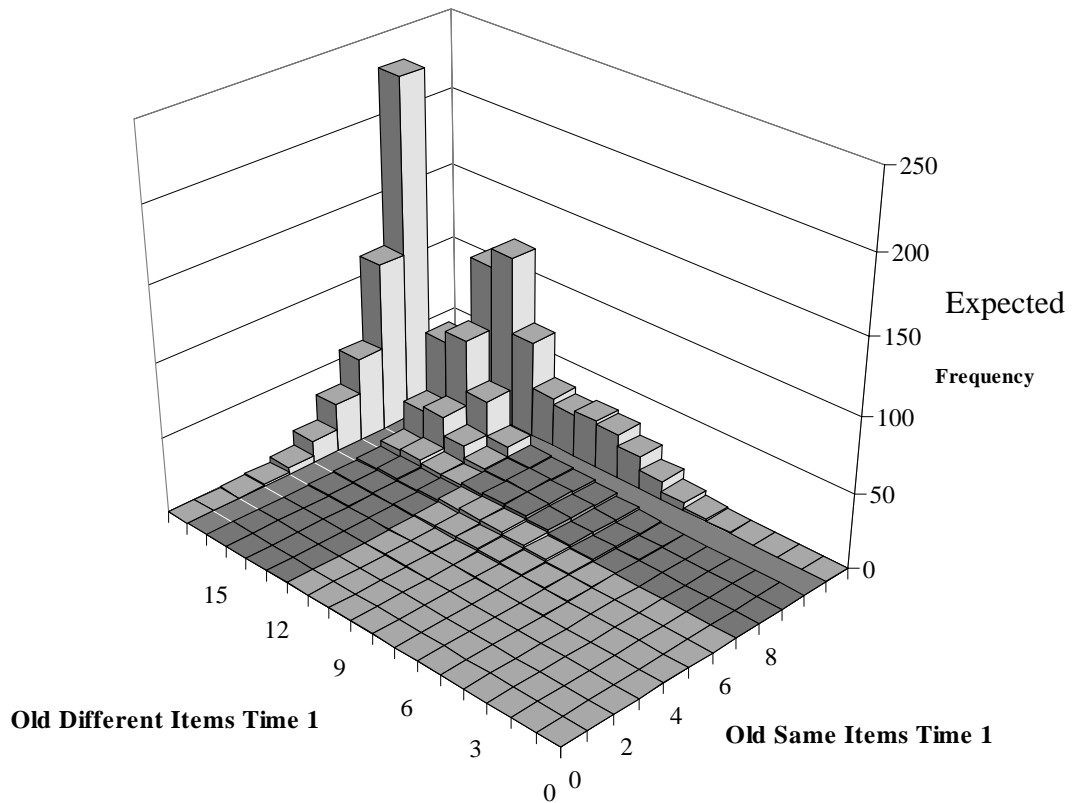
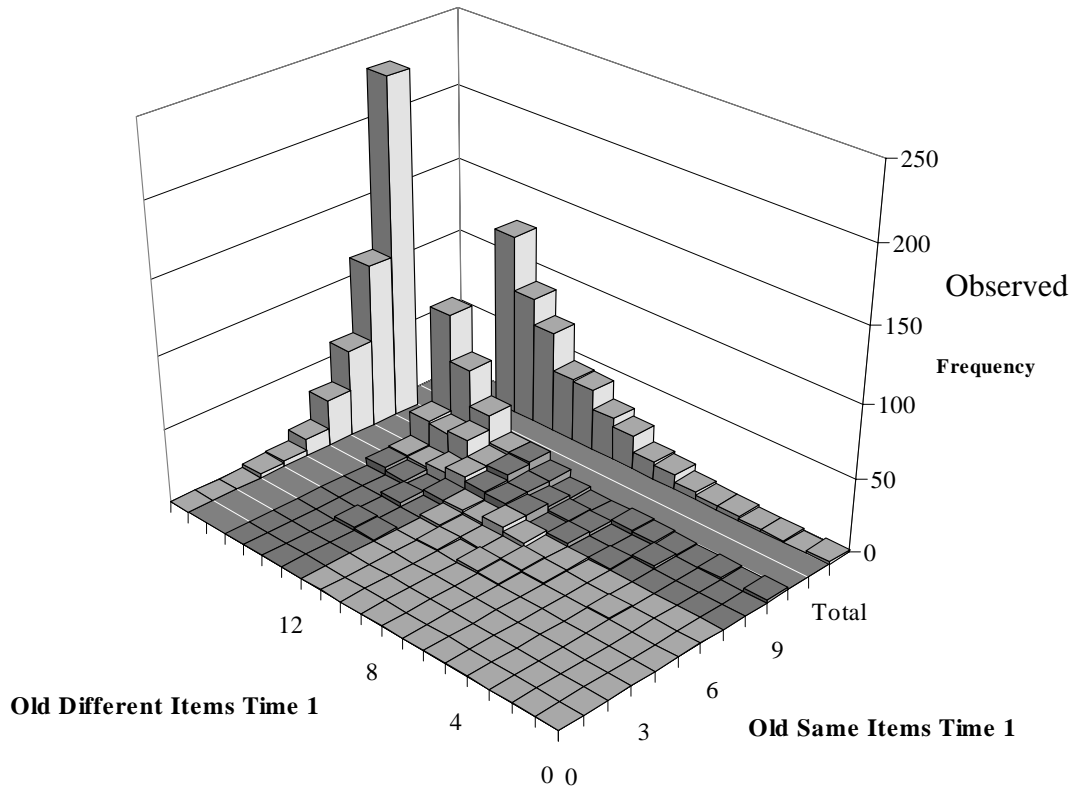


Figure 51. Observed (top) and Expected (bottom) Frequencies for Old-same and Old - different Items at Time 1, Showing Both Joint and Marginal Distributions.

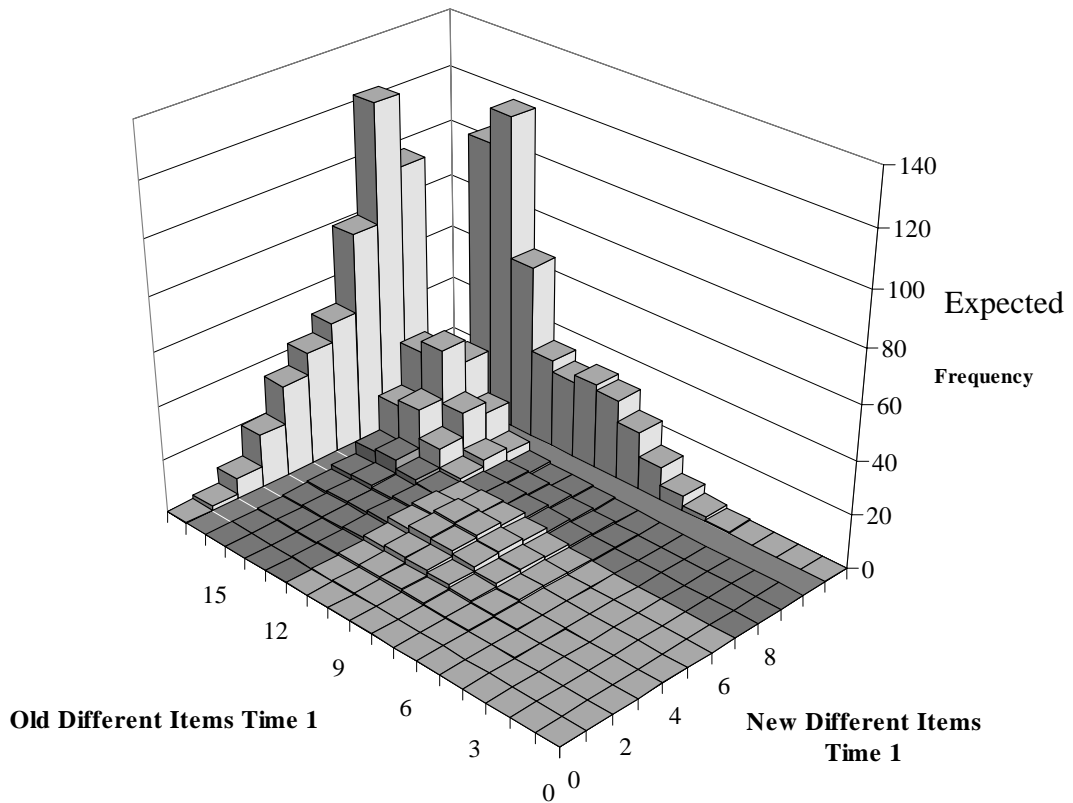
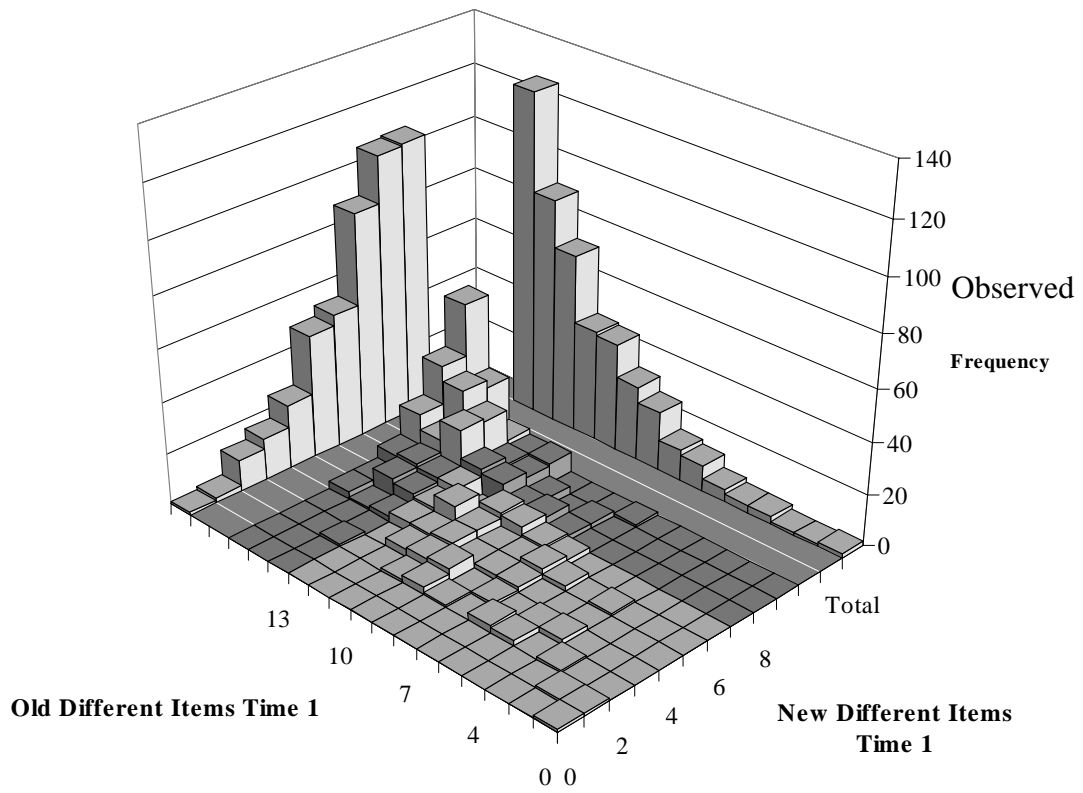


Figure 52. Observed (top) and Expected (bottom) Frequencies for Old-different and New-different Items at Time 1, Showing Both Joint and Marginal Distributions.

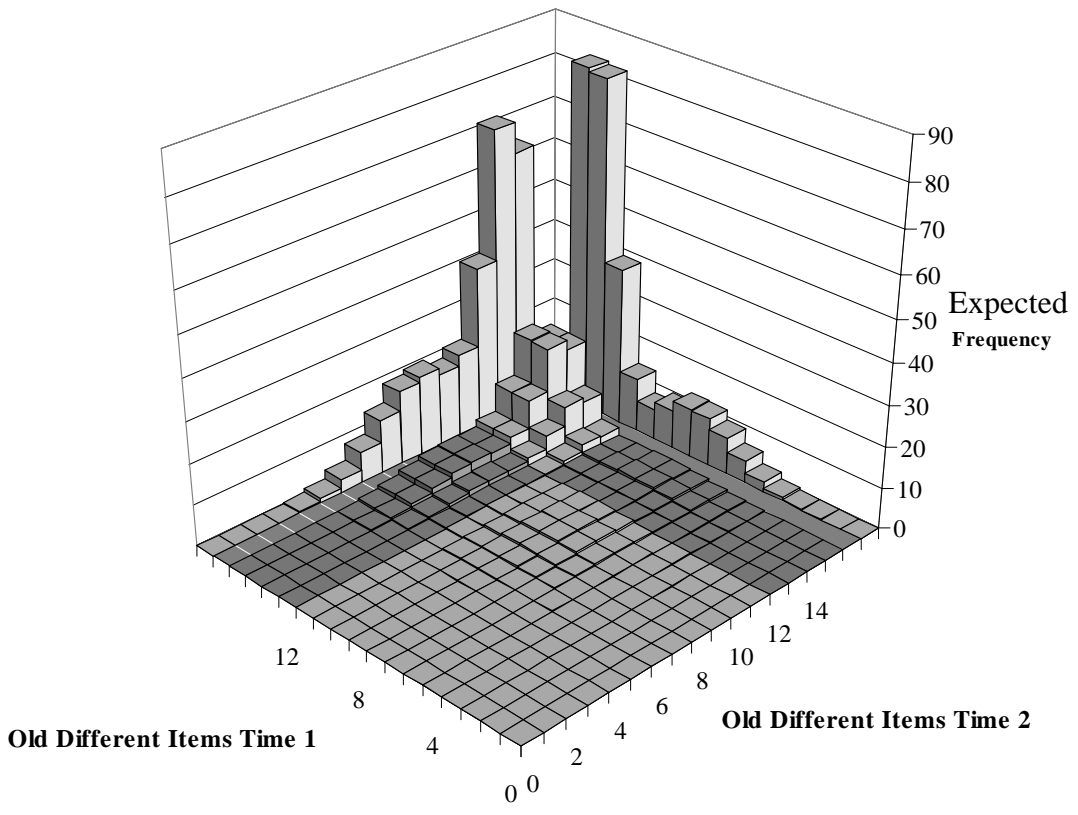
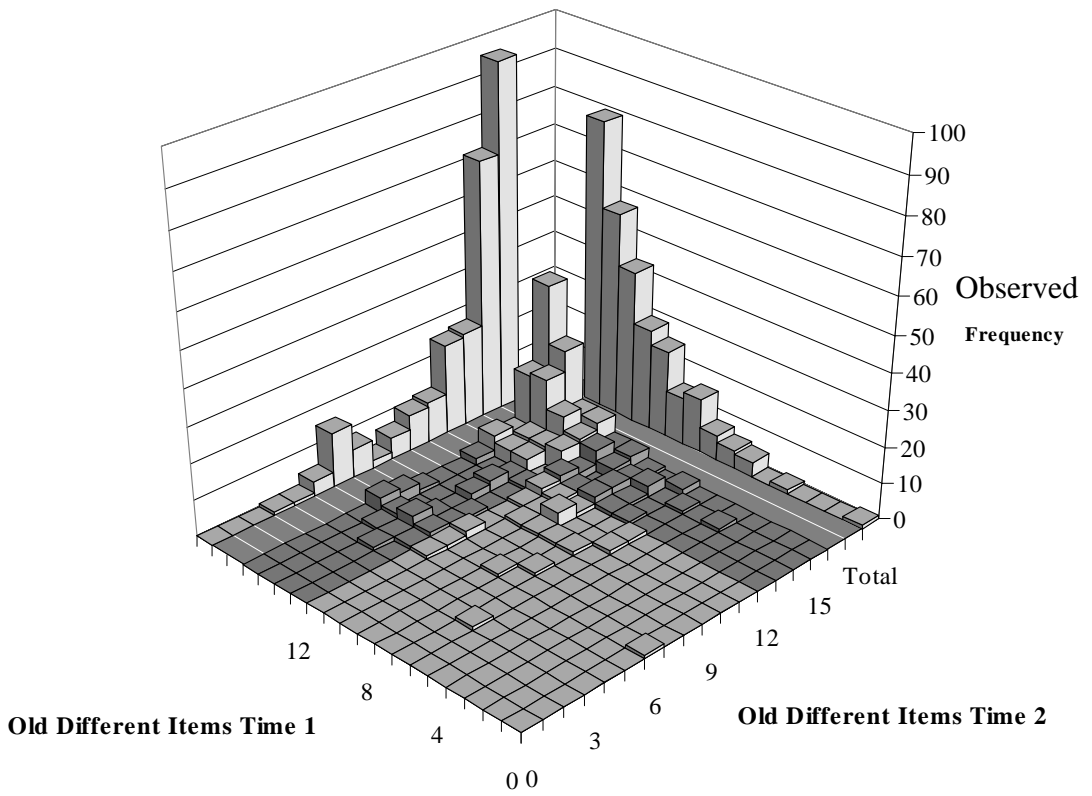


Figure 53. Observed (top) and Expected (bottom) Frequencies for Old-different Items at Times 1 and 2, Showing Both Joint and Marginal Distributions.

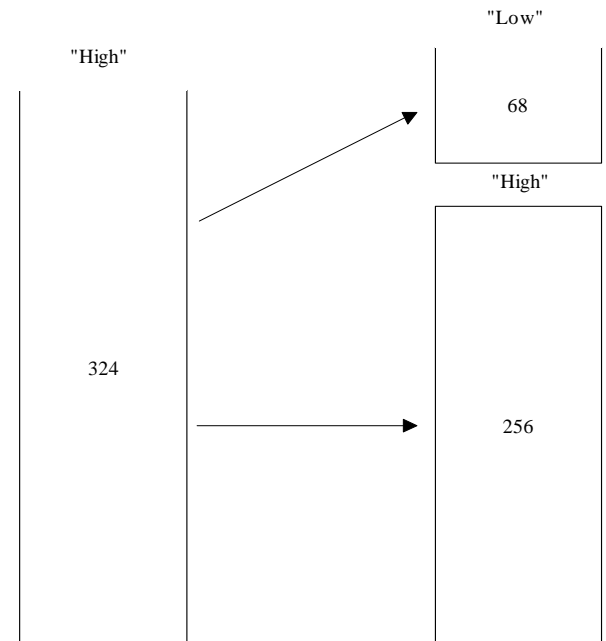
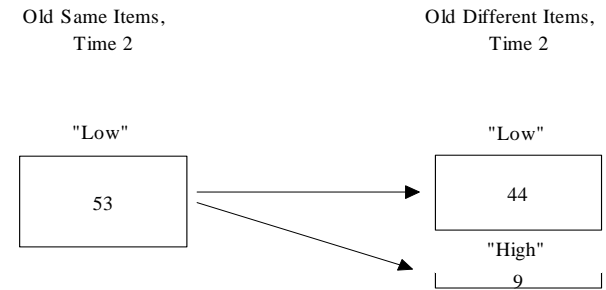
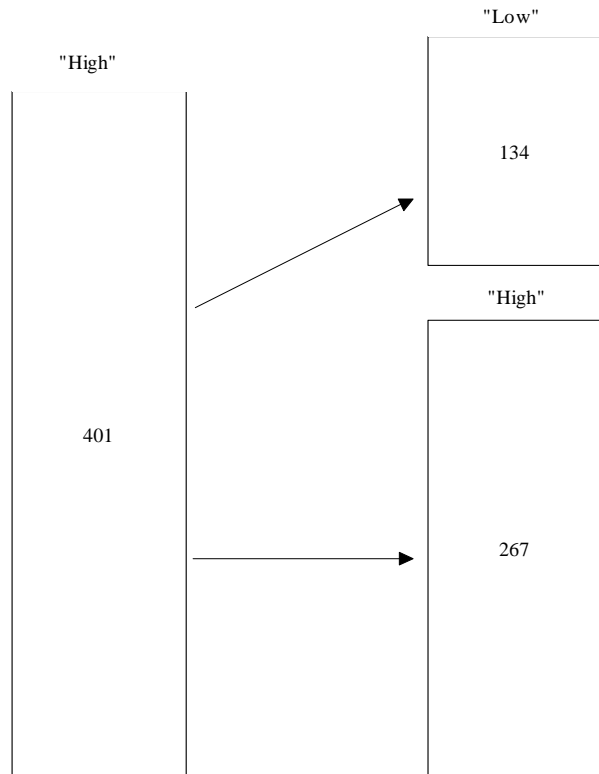
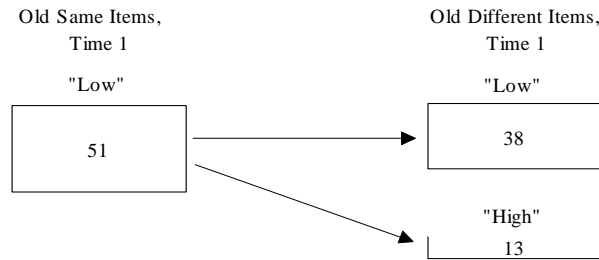


Figure 54. Within- and Across-Component Transition Frequencies within Item-type and Time.

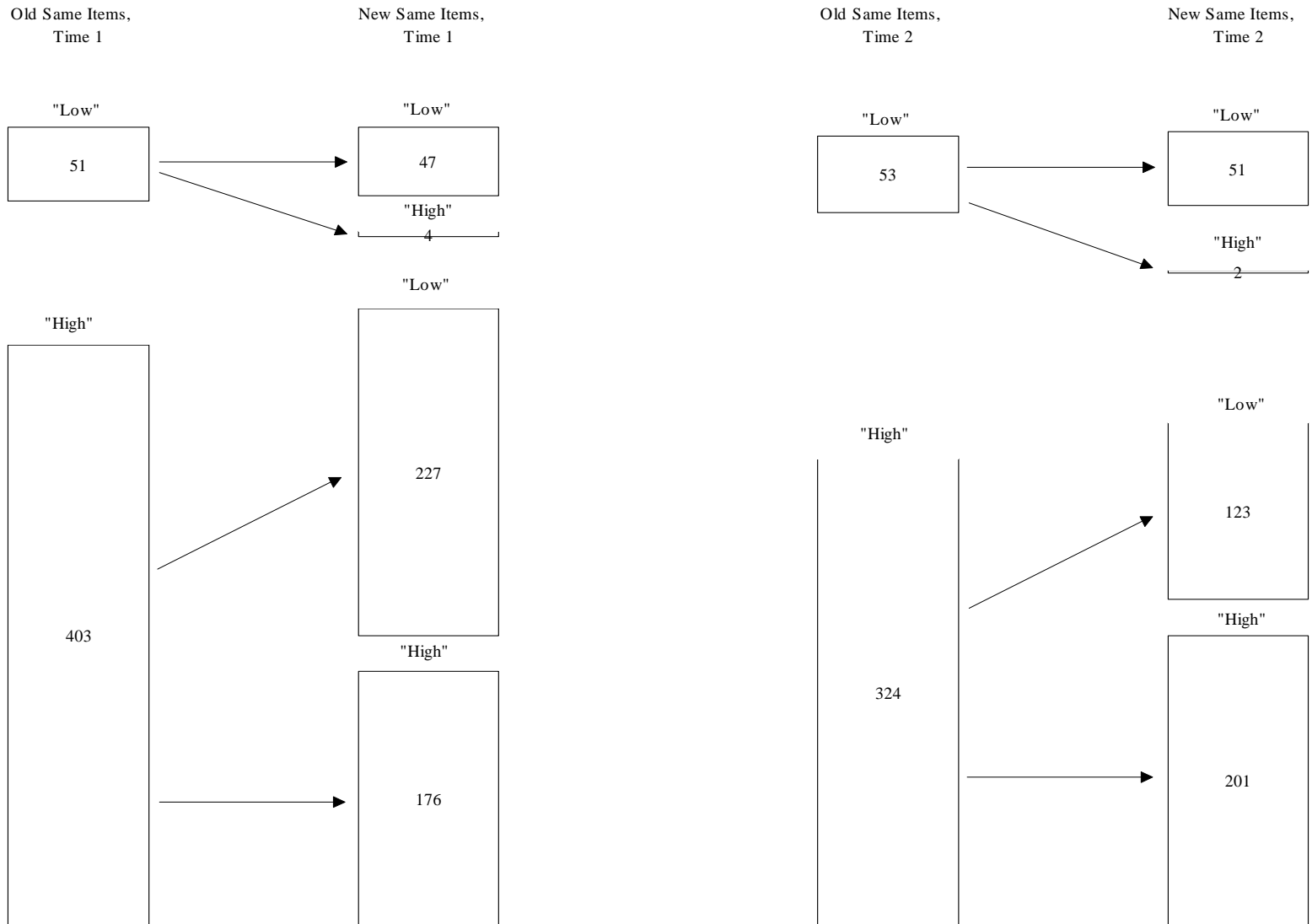


Figure 55. Within- and Across-Component Transition Frequencies within Item-status and Time.

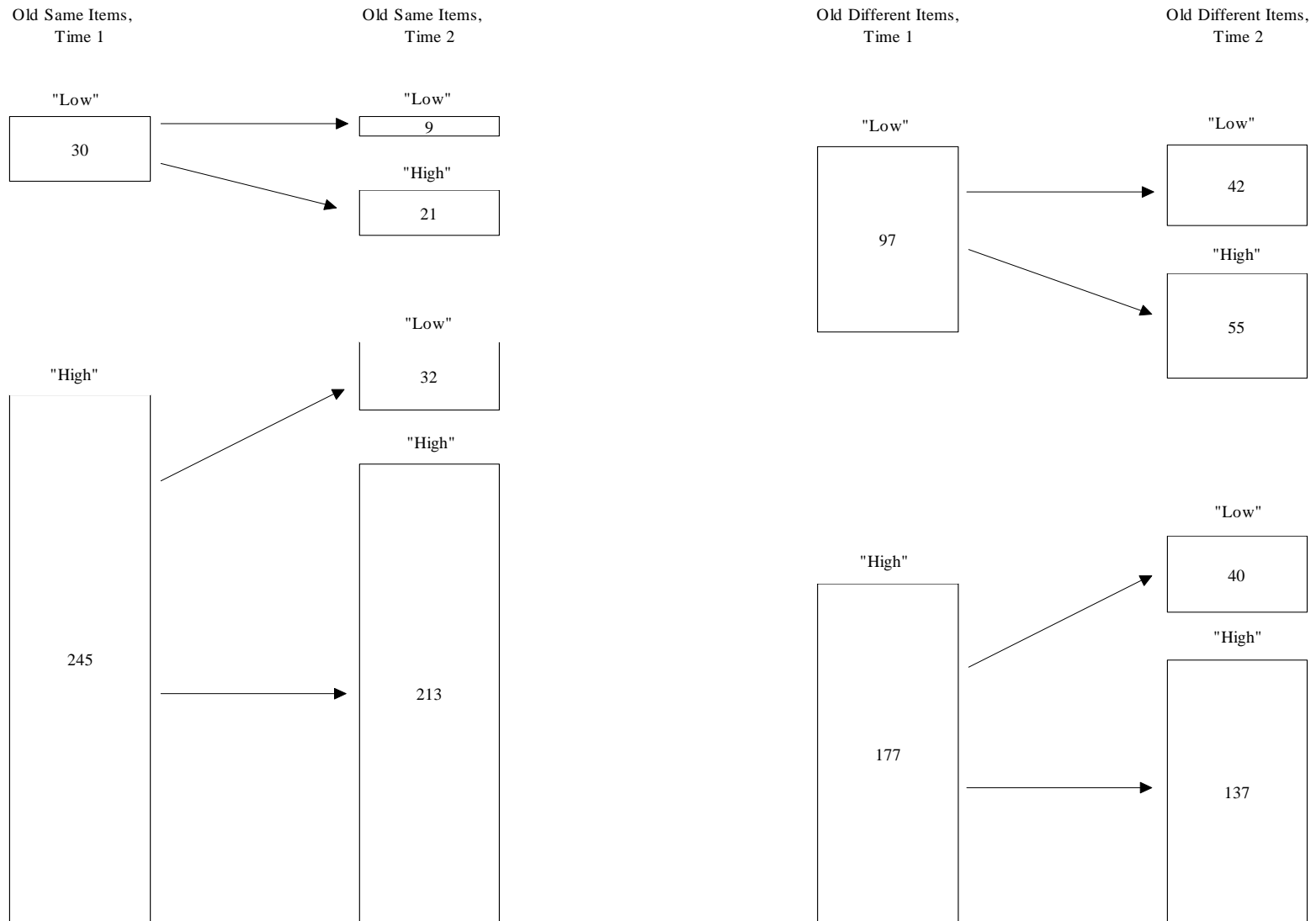
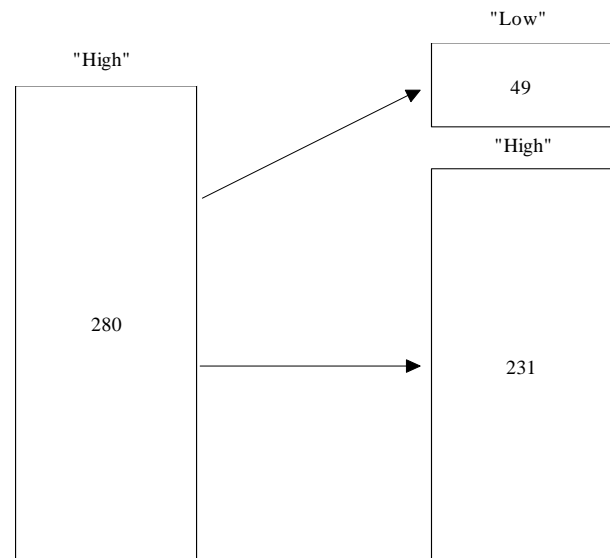
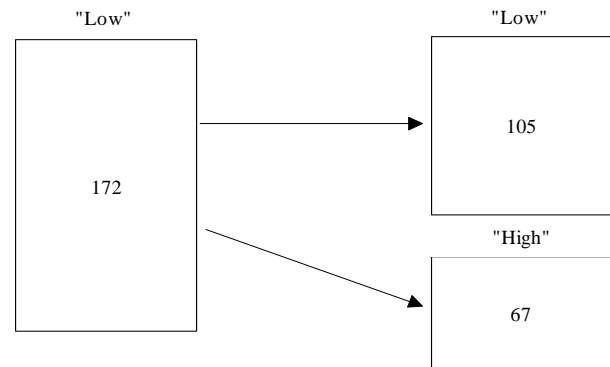


Figure 56. Within- and Across-Component Transition Frequencies within Item-type and Item-status.

Old Different Items,
Time 1

New Different Items,
Time 1



Old Different Items,
Time 2

New Different Items,
Time 2

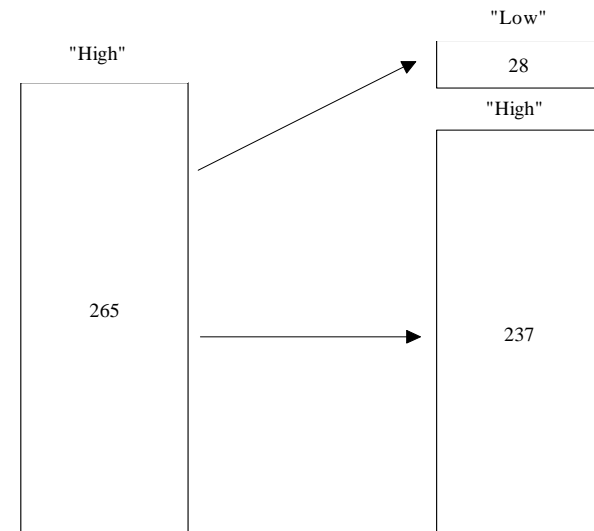
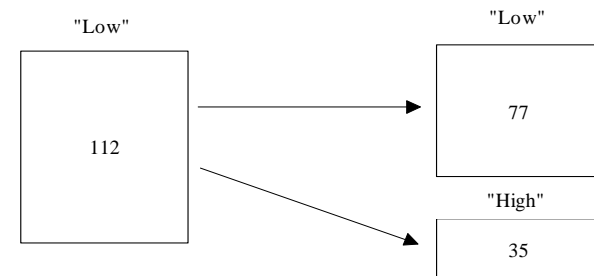


Figure 57. Within- and Across-Component Transition Frequencies within Item-status and Time.

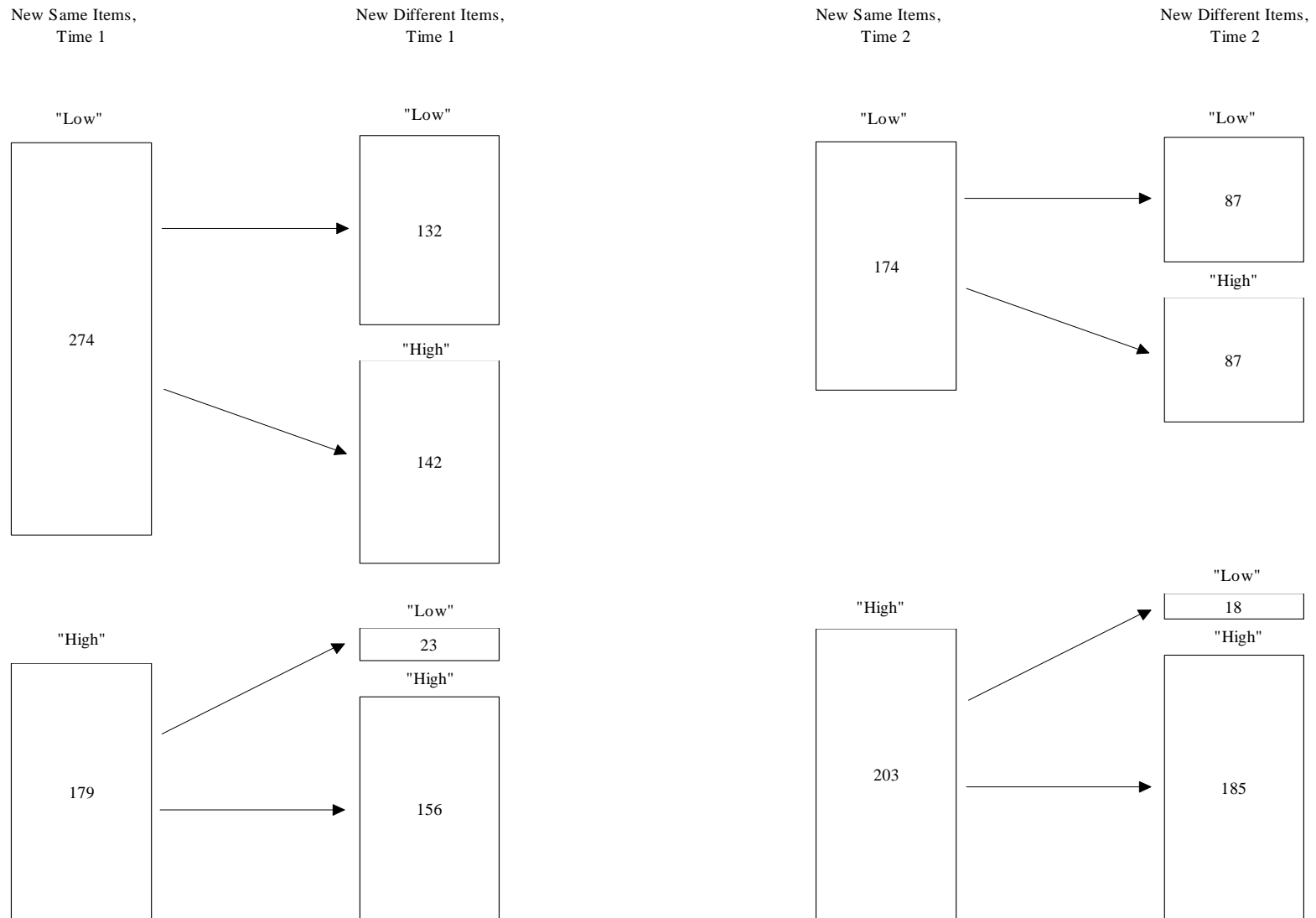


Figure 58. Within- and Across-Component Transition Frequencies within Item-type and Time.

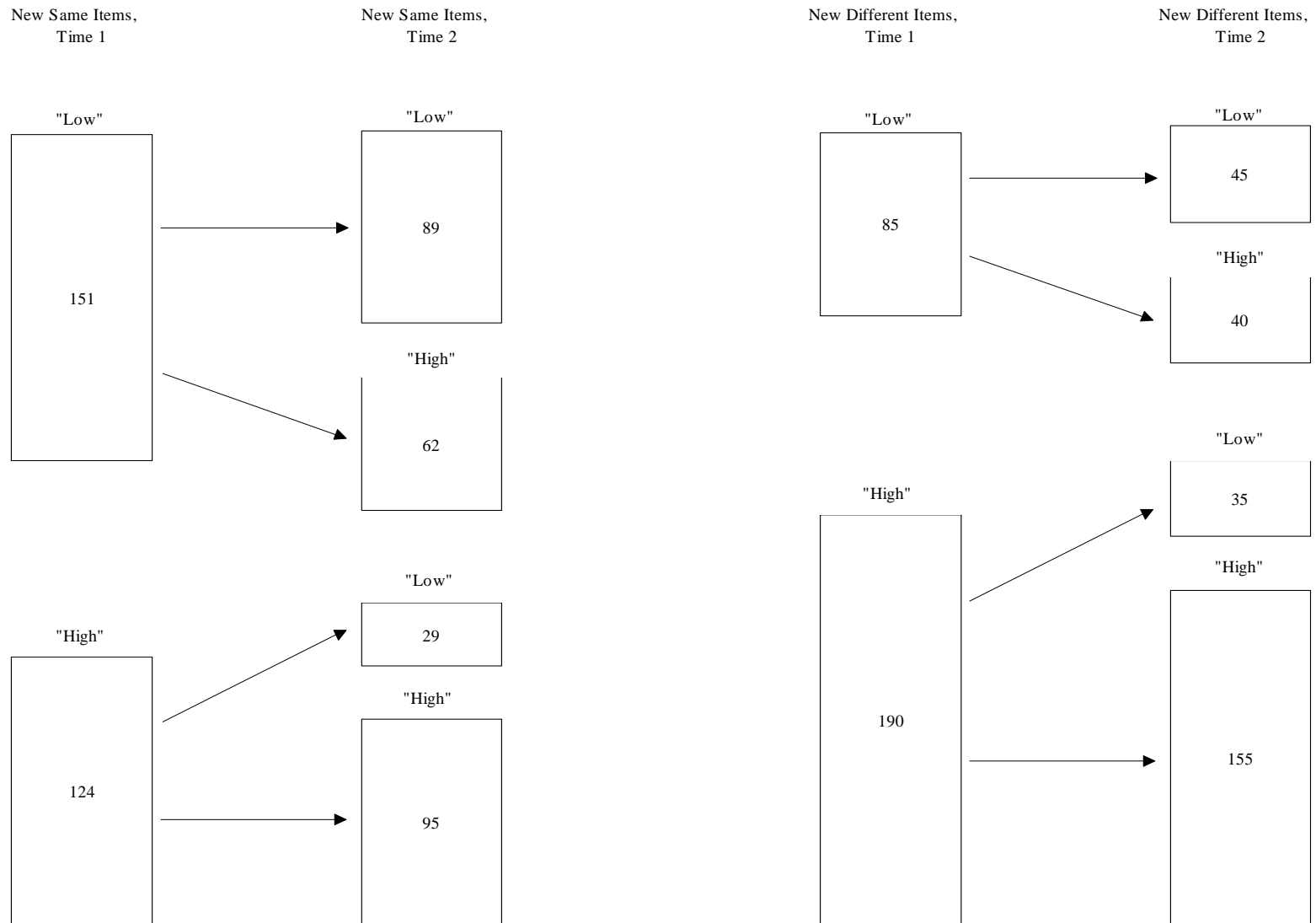


Figure 59. Within- and Across-Component Transition Frequencies within Item-type and Item-status.

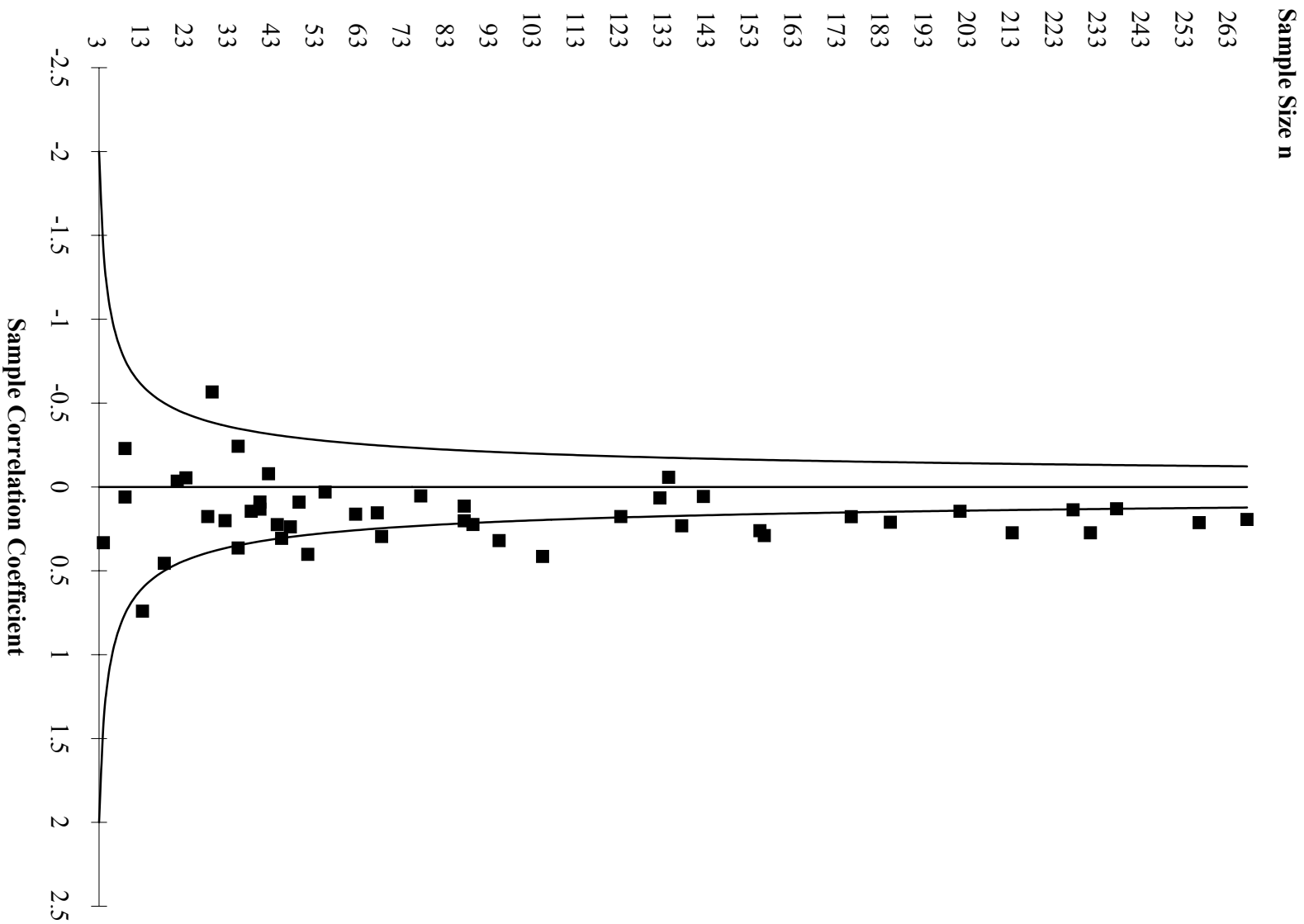


Figure 60. Funnel Graph of Within-cell Correlations from the Summary-level Item-sets.

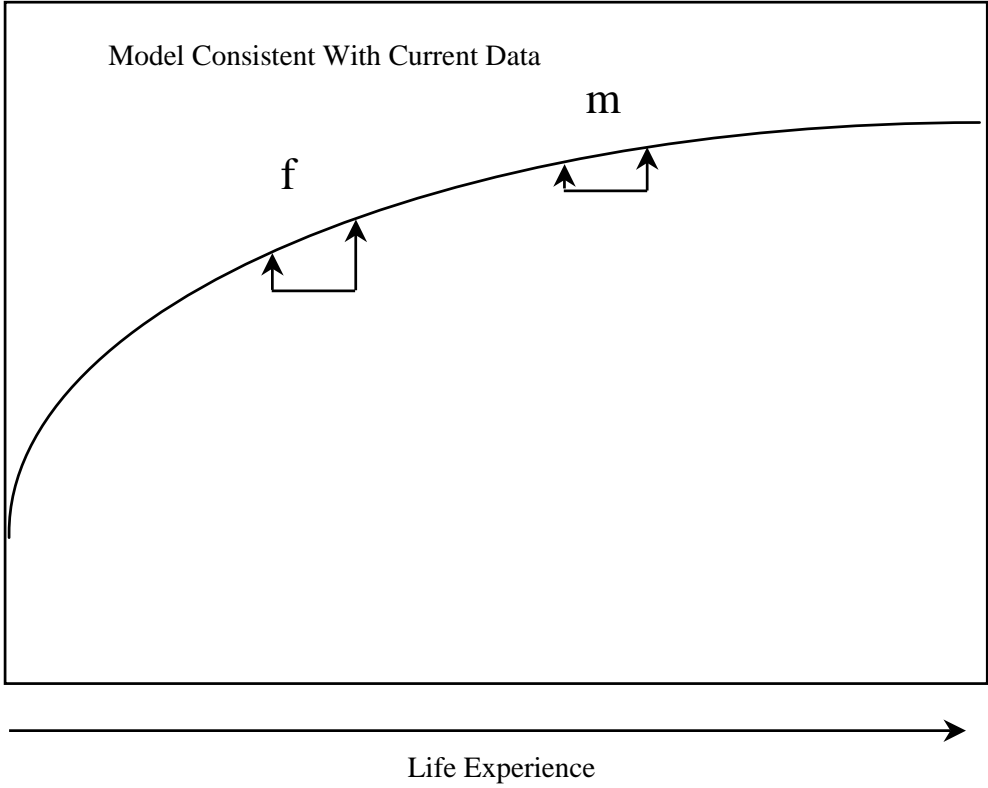
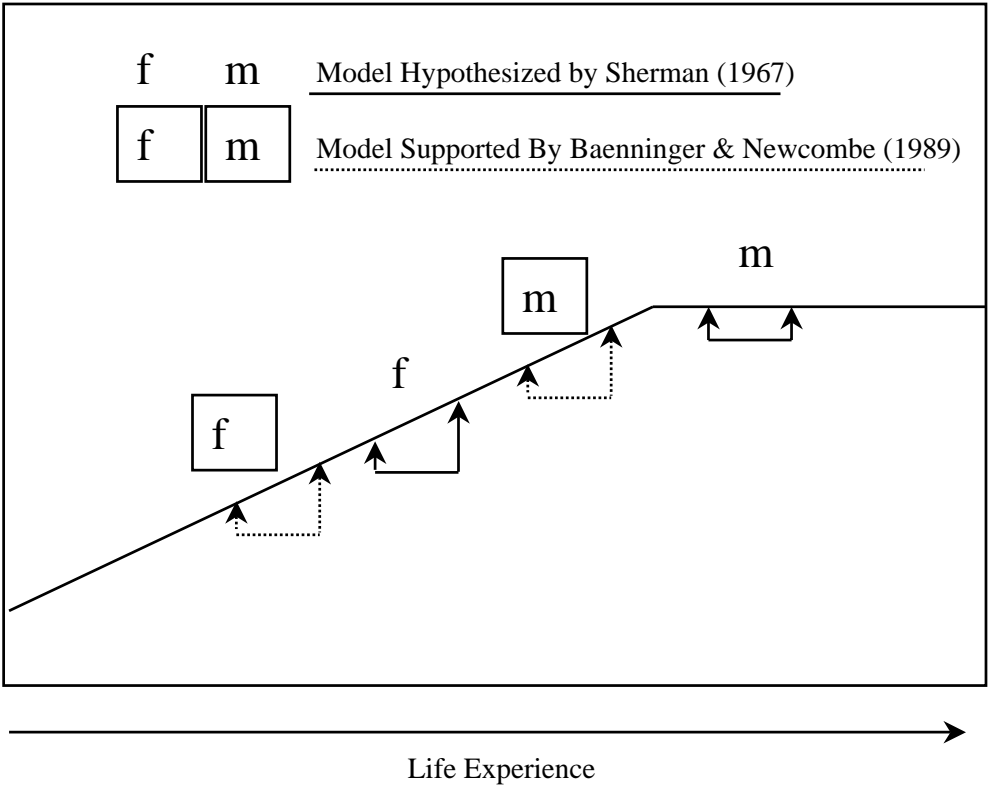


Figure 61. Models of a Sex by Training Interaction.

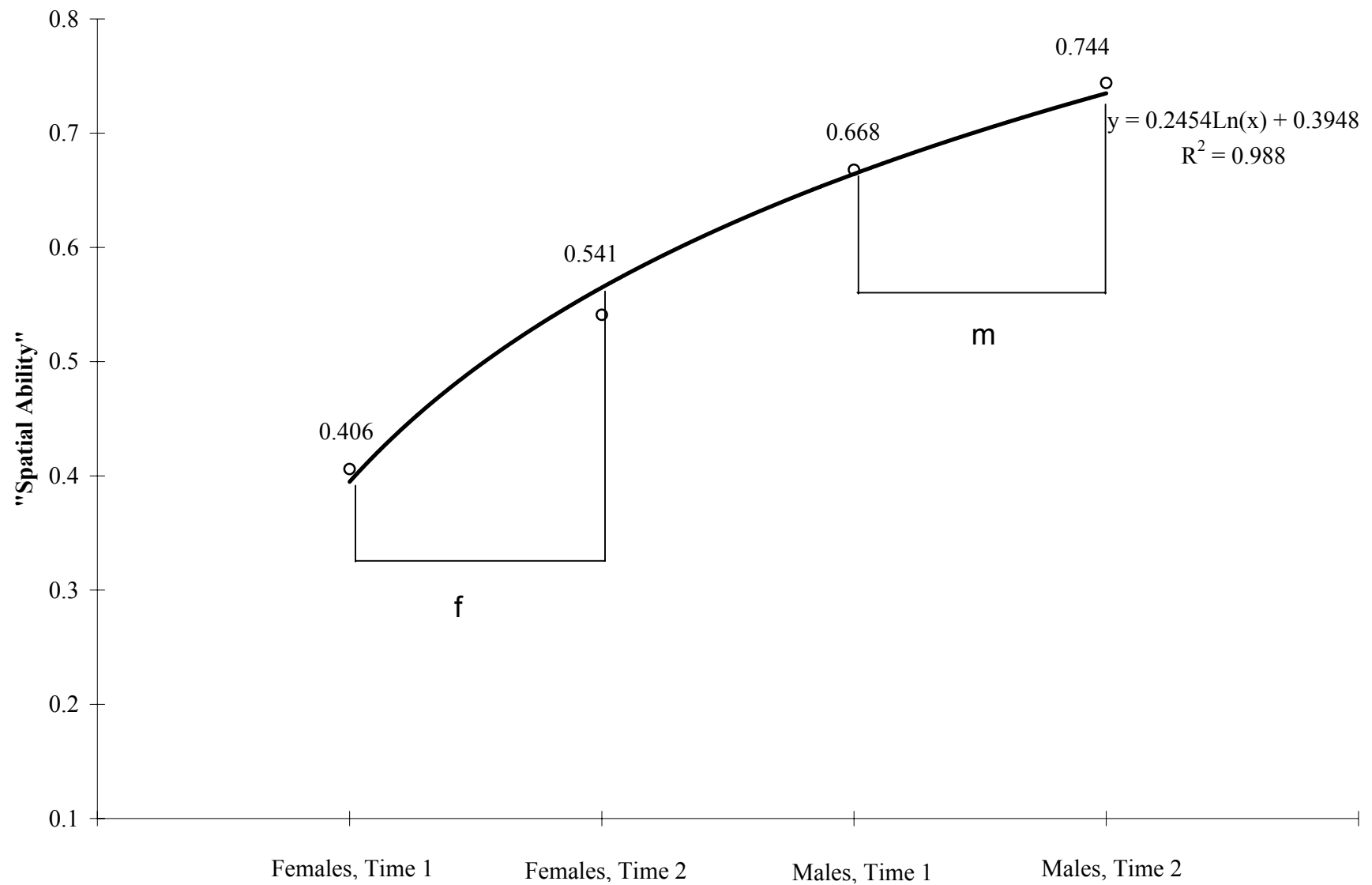


Figure 62. Male and Female "High" Component Proportion Estimates over Time with a Logarithmic Regression Function.

REFERENCES

- Alington, D.E., Leaf, R.C. & Monaghan, J.R. (1992). Effects of stimulus color, pattern, and practice on sex differences in mental rotations task performance. The Journal of Psychology, 126, 539-553.
- American Council on Education (1992). American Universities and Colleges (14th ed.). Hawthorne, NY: Walter de Gruyter.
- Anderson, J.R. (1978). Arguments concerning representations for mental imagery. Psychological Review, 85, 249-277.
- Baenninger, M. & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. Sex Roles, 20, 327-344.
- Bejar, I.I. (1990). A generative analysis of a three-dimensional spatial task. Applied Psychological Measurement, 14, 237-245.
- Berg, C., Hertzog, C. & Hunt, E. (1982). Age differences in the speed of mental rotation. Developmental Psychology, 18, 95-107.
- Bethel-Fox, C.E. & Shepard, R.N. (1988). Mental rotation: Effects of stimulus complexity on familiarity. Journal of Experimental Psychology, 14, 12-23.
- Blade, M.F. & Watson, W.S. (1955). Increase in spatial visualization test scores during engineering study. Psychological Monographs, 69, 1-13.
- Blischke, W.R. (1964). Estimating the parameters of mixtures of binomial distributions. Journal of the American Statistical Association, 59, 510-528.
- Box, G.E.P. (1976). Science and Statistics. Journal of the American Statistical Association, 71, 791-799.

Brainerd, C.J. (1982). Children's concept learning as rule-sampling systems with Markovian properties. In C.J. Brainerd (Ed.), Children's logical and mathematical cognition: Progress in cognitive development research (pp. 177-212). New York: Springer-Verlag.

Brainerd, C.J. (1993). Cognitive development is abrupt (but not stage-like) [Commentary]. Monographs of the Society for Research in Child Development, 58(9, Serial No. 237).

Brinkman, E.H. (1966). Programmed instruction as a technique for improving spatial visualization. Journal of Applied Psychology, 50, 179-184.

Bryden, M.P., George, J. & Inch, R. (1990). Sex differences and the role of figural complexity in determining the rate of mental rotation. Perceptual and Motor Skills, 70, 467-477.

Carpenter, P.A & Just, M.A. (1978). Eye fixations during mental rotation. In J.W. Senders, D.F. Fisher, & R.A. Monty (Eds.), Eye Movements and the Higher Psychological Functions. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carpenter, P.A. & Just, M.A. (1986). Spatial ability: An information processing approach to psychometrics. In R.J. Sternberg (Ed.), Advances in the Psychology of Human Intelligence (Vol. 3, pp. 221-253). Hillsdale, NJ: Lawrence Erlbaum.

Carter, P., Pazak, B. & Kail, R. (1983). Algorithms for processing spatial information. Journal of Experimental Child Psychology, 36, 284-304.

Clogg, C.C. (1979). Some latent structure models for the analysis of likert-type data. Social Science Research, 8, 287-301.

Connor, J.M., Schackman, M., & Serbin, L.A. (1978). Sex-related differences in response to practice on a visual-spatial test and generalization to a related test. Child Development, 49, 24-29.

Connor, J.M., Serbin, L.A. & Schackman, M. (1977). Sex differences in children's response to training on a visual-spatial test. Developmental Psychology, 13, 293-294.

Coombs, C.H., Dawes, R.M. & Tversky, A. (1970). Mathematical Psychology: An Elementary Introduction. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Cooper, L.A. & Mumaw, R.J. (1985). Spatial aptitude. In R.F. Dillon (Ed.), Individual Differences in Cognition (Vol. 2). New York: Academic Press.

Cooper, L.A. & Shepard, R.N. (1973). The time required to prepare for a rotated stimulus. Memory and Cognition, 1, 246-250.

Cooper, L.A. (1982). Strategies for visual comparison and representation: Individual differences. In R.J. Sternberg (Ed.) Advances in the Psychology of Human Intelligence (Vol. 1, pp. 77-124). Hillsdale, NJ: Lawrence Erlbaum.

Damos, D.L. (1990). Using the fold point to analyze mental rotation data: A second look. Bulletin of the Psychonomic Society, 28, 23-26.

Diaconis, P. & Efron, B. (1985). Testing for independence in a two-way contingency tables: New interpretations of the chi-square statistic. Annals of Statistics, 13, 845-874.

Egan, D.E. (1979). Testing based on understanding: Implications from studies of spatial ability. Intelligence, 3, 1-15.

Eliot, J. (1987). Models of Psychological Space. New York: Springer-Verlag.

Ernest, C.H. (1977). Imagery ability and cognition: A critical review. Journal of Mental Imagery, 2, 181-216.

Everitt, B.S. & Hand, D.J. (1981). Finite Mixture Distributions. London: Chapman & Hall.

Folk, M.D. & Luce, R.D. (1987). Effects of stimulus complexity on mental rotation rate of polygons. Journal of Experimental Psychology: Human Perception and Performance, 13, 395-404.

Funt, B.V. (1983). A parallel-process model of mental rotation. Cognitive Science, 7, 67-93.

Ghiselli, E.E.(1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Goebel, R.P. (1990). The mathematics of mental rotations. Journal of Mathematical Psychology, 34, 435-444.

Goldstein, D., Haldane, D. & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. Memory and Cognition, 18, 546-550.

Gordon, H.W. & Leighty, R. (1988). Importance of specialized cognitive function in the selection of military pilots. Journal of Applied Psychology, 73, 38-45.

Gordon, R. (1949). An investigation into some of the factors that favour the formation of stereotyped images. British Journal of Psychology, 39, 156-167.

Hochberg, J. & Gellman, L. (1977). The effect of landmark features on mental rotation times. Memory and Cognition, 5, 23-26.

Hock, H.S. & Ross, K. (1975). The effect of familiarity on rotational transformation. Perception and Psychophysics, 18, 15-20.

- Huber, P.J. (1977). Robust Statistical Procedures. Philadelphia: Society for Industrial and Applied Mathematics.
- Johnson, E.S. & Meade, A.C. (1987). An early sex difference. Child Development, 58, 725-740.
- Juhel, J. (1991). Spatial abilities and individual differences in visual information processing. Intelligence, 15, 117-137.
- Just, M.A. & Carpenter, P.A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. Psychological Review, 92, 137-172.
- Just, M.A. & Carpenter, P.A. (1976). Eye fixations and cognitive processes. Cognitive Psychology, 8, 441-480.
- Kail, R. & Park, Y. (1990). Impact of practice on speed of mental rotation. Journal of Experimental Child Psychology, 49, 227-244.
- Kail, R. (1986). The impact of extended practice on rate of mental rotation. Journal of Experimental Child Psychology, 42, 378-391.
- Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. Psychological Bulletin, 109, 490-501.
- Kail, R., Carter, P. & Pellegrino, J. (1979). The locus of sex differences in spatial ability. Perception and Psychophysics, 26, 182-186.
- Kail, R., Pellegrino, J. & Carter, P. (1980). Developmental changes in mental rotation. Journal of Experimental Child Psychology, 29, 102-116.
- Kail, R., Stevenson, M.R. & Black, K.N. (1984). Absence of a sex difference in algorithms for spatial problem solving. Intelligence, 8, 37-46.

Kaplan, B.J. & Weisberg, F.B. (1987). Sex differences and practice effects on two visual-spatial tasks. Perceptual and Motor Skills, 64, 139-142.

Kosslyn, S.M. (1980). Image and Mind. Cambridge, MA: Harvard University Press.

Kyllonen, P.C., Lohman, D.F., & Snow, R.E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. Journal of Educational Psychology, 76, 130-145.

Kyllonen, P.C. Lohman, D.F., & Woltz, D.J. (1984). Componential modeling of alternative strategies for performing spatial tasks. Journal of Educational Psychology, 76, 1325-1345.

Lazarsfeld, P.F. & Henry, N.W. (1968). Latent Structure Analysis. New York: Houghton Mifflin Company.

Linn, M.C. & Petersen, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. Child Development, 56, 1479-1498.

Lohman, D.F. (1986). The effect of speed-accuracy tradeoff on sex differences in mental rotation. Perception and Psychophysics, 39, 427-436.

Lohman, D.F. (1988). Spatial abilities as traits, process, and knowledge. In R.J. Sternberg (Ed.). Advances in the Psychology of Human Intelligence (Vol. 4, pp. 181-248). New York: Lawrence Erlbaum.

Lohman, D.F. & Kyllonen, P.C.(1983). Individual differences in solution strategy on spatial tasks. In R.F. Dillon and R.R. Schmeck (Eds.) Individual Differences in Cognition (Vol. 1, pp. 105-135). New York: Academic Press.

Lohman, D.F. and Nichols, P.D. (1990). Training spatial abilities: Effects of practice on rotation and synthesis tasks. Learning and Individual Differences, 2, 67-93.

Lord, F.M. (1965). A strong true-score theory, with applications. Psychometrika, 30, 239-270.

McCutcheon, A.L. (1987). Latent Class Analysis. Beverly Hills, CA: Sage Publications.

McGee, M.G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. Psychological Bulletin, 86, 889-918.

McLachlan, G.J. & Basford, K.E. (1988). Mixture Models: Inference and Applications to Clustering. New York: Dekker.

McLurg, P.A. & Chaille, C. (1987). Computer games: Environments for developing spatial cognition? Journal of Educational Computing Research, 3, 95-110.

Merriman, W.E., Keating, D.P. & List, J.A. (1985). Mental rotation of facial profiles: Age-, sex-, and ability-related differences. Developmental Psychology, 21, 888-900.

Miller, R.B., Kelly, G.N., & Kelly, J.T. (1988). Effects of logo computer programming experience on problem solving and spatial relations ability. Contemporary Educational Psychology, 13, 348-357.

Mislevy, R.J., Wingersky, M.S., Irvine, S.H. & Dann, P.L. (1991). Resolving mixtures of strategies in spatial visualization tasks. British Journal of Mathematical and Statistical Psychology, 44, 265-288.

Mumaw, R.J., Pellegrino, J.W., Kail, R.V., & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. Memory and Cognition, 12, 515-521.

Olson, D.M., Eliot, J. & Hardy, R.C. (1988). Relationships between activities and sex-related differences in performance on spatial tests. Perceptual and Motor Skills, 67, 223-232.

Olson, D.R., & Bialystok, E. (1983). Spatial Cognition: The Structure and Development of Mental Representations of Spatial Relations. Hillsdale, NJ: Lawrence Earlbaum Associates.

Ozer, D.J. (1987). Personality, intelligence, and spatial visualization: Correlates of mental rotations test performance. Journal of Personality and Social Psychology, 53, 129-134.

Paivio, A. (1971). Imagery and Verbal Processes. New York: Holt, Rinehart & Winston.

Pellegrino, J.W. & Kail, R. (1982). Process analyses of spatial aptitude. In R.J. Sternberg (Ed.), Advances in the Psychology of Human Intelligence (Vol. 1, pp. 311-365). Hillsdale, NJ: Lawrence Erlbaum.

Phillips, G.M. & Taylor, P.J. (1973). Theory and Applications of Numerical Analysis. New York: Academic Press.

Piaget, J. & Inhelder, B. (1956). The Child's Conception of Space. New York: Norton.

Piaget, J. & Inhelder, B. (1971). Mental Imagery in the Child. New York: Basic Books.

Poltrock, S.E. & Agnoli, F. (1986). Are spatial visualization ability and visual imagery ability equivalent. In R.J. Sternberg (Ed.) Advances in the Psychology of Human Intelligence (Vol. 3, pp.255-296). Hillsdale, NJ: Lawrence Erlbaum.

Poltrock, S.E. & Brown, P. (1984). Individual differences in visual imagery and spatial ability. Intelligence, 8, 93-138.

Pylyshyn, Z.W. (1981). The imagery debate: Analogue media versus tacit knowledge. Psychological Review, 88, 16-45.

Richardson, A. (1969). Mental Imagery. New York: Springer.

Robertson, L.C. & Palmer, S.E. (1983). Holistic processes in the perception and transformation of disoriented figures. Journal of Experimental Psychology: Human Perception and Performance, 9, 203-214.

Roscoe, J.T. (1975). Fundamental Research Statistics for the Behavioral Sciences. NY: Holt, Rinehart, and Winston, Inc.

Seddon, G.M., Eniaiyaju, P.A., & Jusoh, I. (1984). The visualization of rotation in diagrams of three-dimensional structures. American Educational Research Journal, 21, 25-38.

Shepard, R.N. & Cooper, L.A. (1982). Mental Images and Their Transformations. Cambridge, MA: The MIT Press.

Shepard, R.N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

Sherman, J.A. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. Psychological Review, 74, 290-299.

Sholl, M.J. & Liben, L.S. (1995). Illusory tilt and Euclidean schemes as factors in performance on the water-level task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1624-1638.

Shubbar, K.E. (1990). Learning the visualisation of rotations in diagrams of three-dimensional structures. Research in Science and Technological Education, 8, 145-154.

Snedecor, G.W. & Cochran, W.G. (1980). Statistical Methods. Ames, IA: The Iowa State University Press.

Straughn, C.T. & Lovejoy-Straughn, B. (1995). Lovejoy's College Guide. NY: Macmillan.

Stumpf, H. (1993). Performance factors and gender-related differences in spatial ability: Another assessment. Memory and Cognition, 21, 828-836.

Tapley, S.M. & Bryden, M.P. (1977). An investigation of sex differences in spatial ability: Mental rotation of three-dimensional objects. Canadian Journal of Psychology, 31, 122-130.

Thomas, H. & Kail, R. (1991). Sex differences in speed of mental rotation and the X-linked genetic hypothesis. Intelligence, 15, 17-32.

Thomas, H. (1977). Fitting cross-classification table data to models when observations are subject to classification error. Psychometrika, 42, 199-206.

Thomas, H. (1993). Individual differences in children, studies, and statistics: Applications of empirical Bayes methodology. In M.L. Howe & R. Pasnak (Eds.), Emerging Themes in Cognitive Development. Vol. 1: Foundations. New York: Springer-Verlag.

Thomas, H. & Lohaus, A. (1993). Modeling growth and individual differences in spatial tasks. Monographs of the Society for Research in Child Development, 58(9, Serial No. 237).

Thomas, H. & Turner, G.F.W. (1991). Individual differences and development in water-level task performance. Journal of Experimental Child Psychology, 51, 171-194.

Thurstone, L.L. (1938). Primary Mental Abilities. Chicago: University of Chicago Press.

Turner, G.F.W. (1991). Modeling individual differences in spatial task performance with binomial mixtures. Unpublished masters thesis, The Pennsylvania State University, University Park.

Van Voorhis, W.R. (1941). The improvement of space perception ability by training. Unpublished doctoral dissertation, The Pennsylvania State University, University Park.

Vandenberg, S.G. & Kuse, A.R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. Perceptual and Motor Skills, 47, 599-604.

Vasta, R. & Liben, L.S. (1996). The water-level task: An intriguing puzzle. Current Directions in Psychological Science, 5, 171-177.

Voyer, D. & Bryden, M.P. (1990). Gender, level of spatial ability, and lateralization of mental rotation. Brain and Cognition, 13, 18-29.

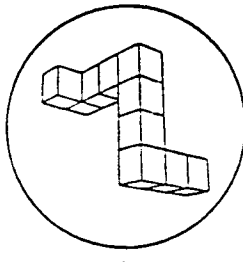
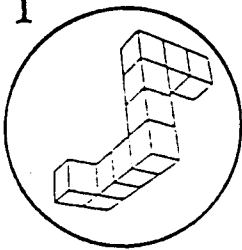
Wallace, B. & Hofelich, B.G. (1992). Process generalization and the prediction of performance on mental imagery tasks. Memory and Cognition, 20, 695-704.

Willis, S.L. & Shaie, K.W. (1988). Gender differences in spatial ability in old age: Longitudinal and intervention findings. Sex Roles, 18, 189-203.

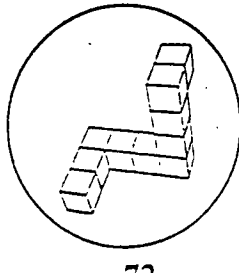
Yuille, J.C. & Steiger, J.H. (1982). Nonholistic processing in mental rotation:
Some suggestive evidence. Perception and Psychophysics, 31, 201-209.

APPENDIX A
PSU MRT TEST

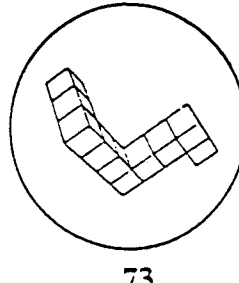
I



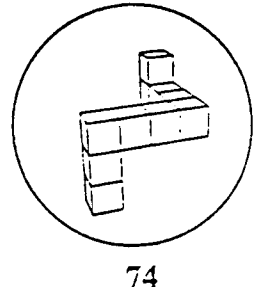
71



72

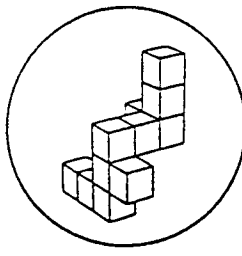
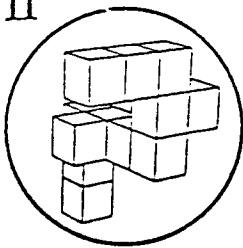


73

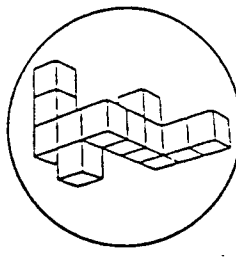


74

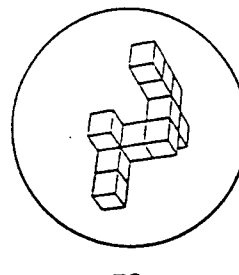
II



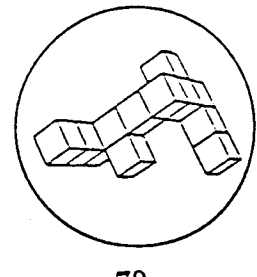
76



77

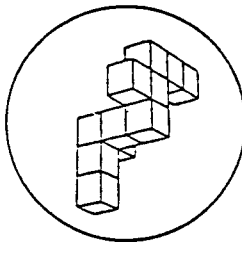
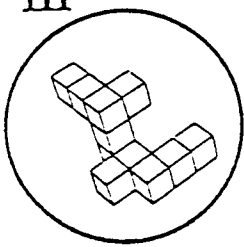


78

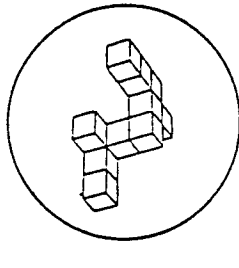


79

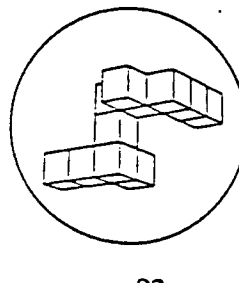
III



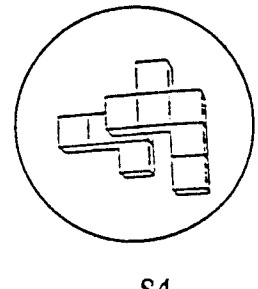
81



82

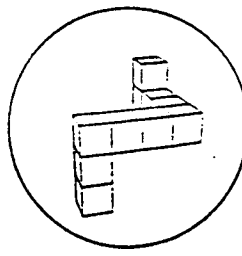
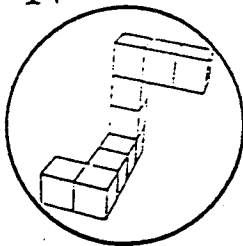


83

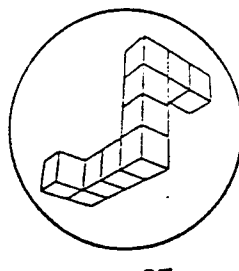


84

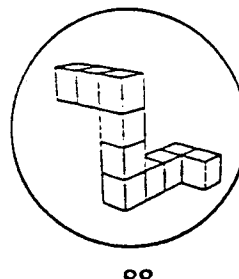
IV



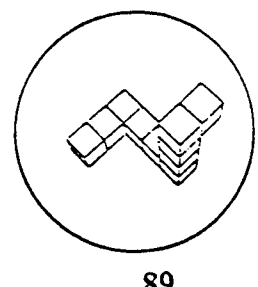
86



87



88

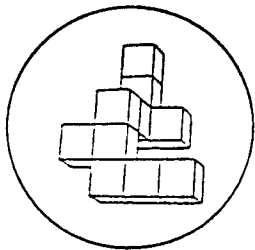
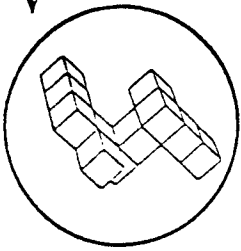


89

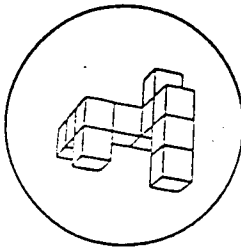
TURN PAGE AND CONTINUE

PSU MRT TEST

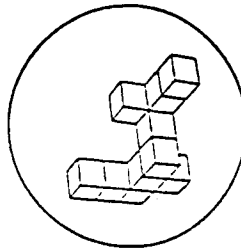
V



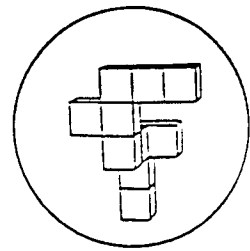
91



92

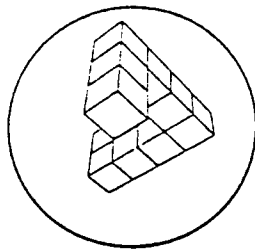
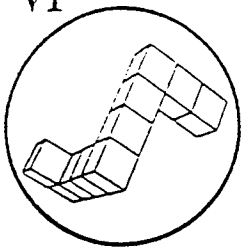


93

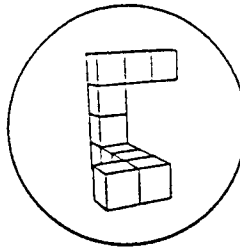


94

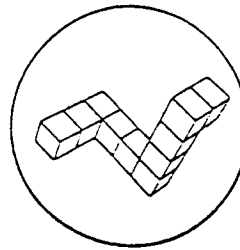
VI



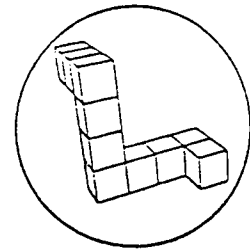
96



97

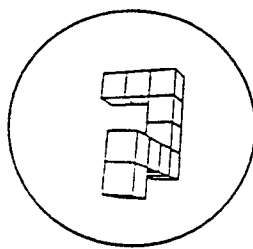
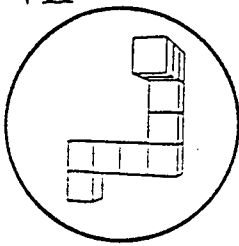


98

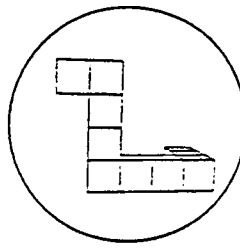


99

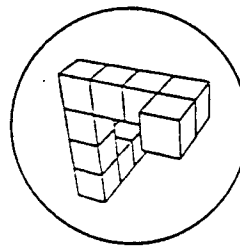
VII



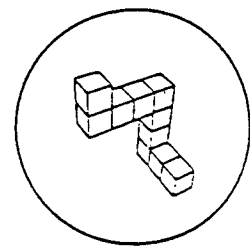
101



102

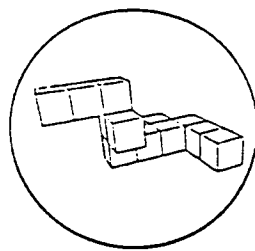
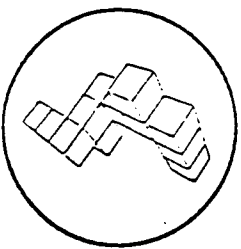


103

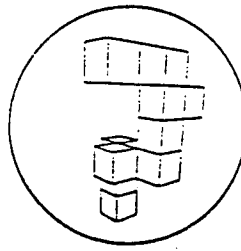


104

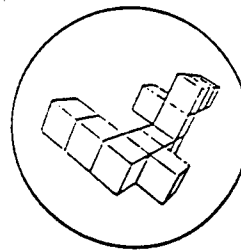
VIII



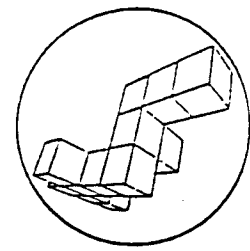
106



107



108

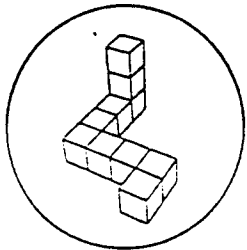
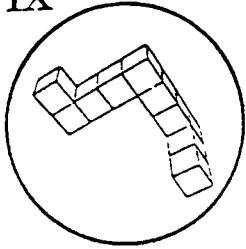


109

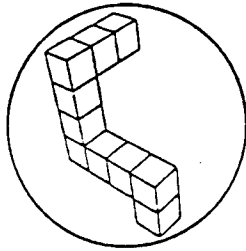
TURN PAGE AND CONTINUE

PSU MRT TEST

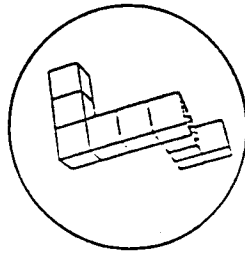
IX



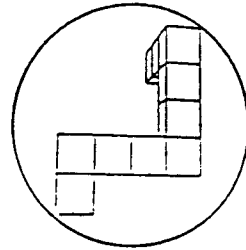
111



112

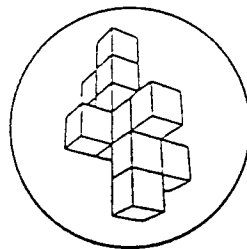
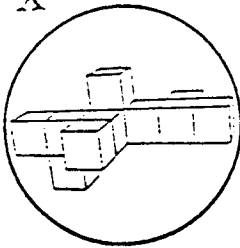


113

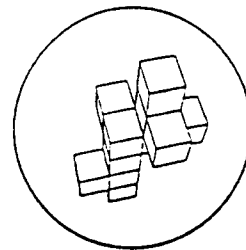


114

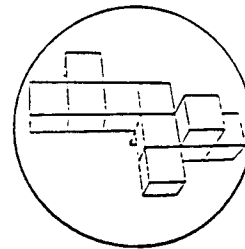
X



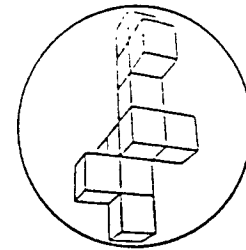
116



117

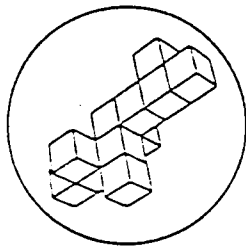
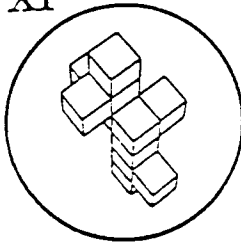


118

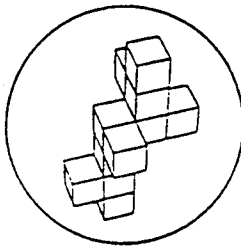


119

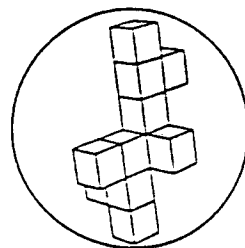
XI



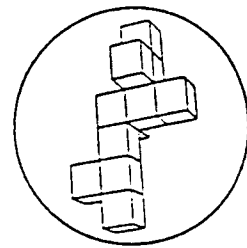
121



122

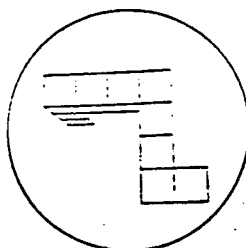
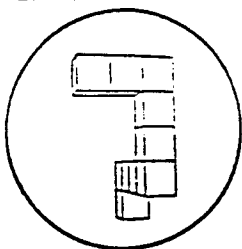


123

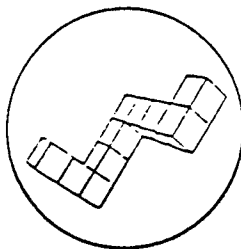


124

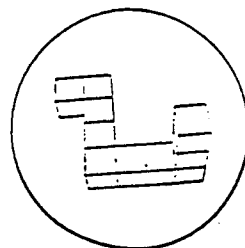
XII



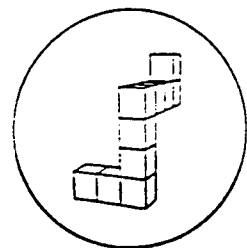
126



127



128



129

APPENDIX B

Model Development and Estimation

in Bivariate Binomial Mixtures

The joint three component binomial model is described below, while the joint two component model is described in Thomas and Lohaus (1993). Both joint models are constrained so that the marginal distributions are mixed binomial. The three component bivariate model is based on three component mixed binomial marginal distributions given by

$$f(x) = \pi_{x1} b(x; \theta_{x1}) + \pi_{x2} b(x; \theta_{x2}) + (1 - \pi_{x1} - \pi_{x2}) b(x; \theta_{x3}) \quad (25)$$

$$f(y) = \pi_{y1} b(y; \theta_{y1}) + \pi_{y2} b(y; \theta_{y2}) + (1 - \pi_{y1} - \pi_{y2}) b(y; \theta_{y3}) \quad (26)$$

Equation 25 represents the same distribution as Equation 4 (p. 25), but π_{x1} replaces π_1 to

distinguish it from the distribution of Y. Here $b(x; \theta_{x1}) = b(x; \theta_{x1}, n)$ where n is suppressed

defines the familiar binomial distribution with success parameter θ . Now define the conditional distribution of Y given $X=x$ by

$$f(y|x) = \pi_{y1|x} b(y; \theta_{y1}) + \pi_{y2|x} b(y; \theta_{y2}) + (1 - \pi_{y1|x} - \pi_{y2|x}) b(y; \theta_{y3}) \quad (27)$$

The conditional weight $\pi_{ya|x} = \tau_{1b} P(1|X) + \tau_{2b} P(2|X) + \tau_{3b} P(3|X)$, where the τ 's represent

transition parameters. As in the two component joint model from Thomas and Lohaus (1993), the τ_{ab} 's are transition parameters which represent state change probabilities, and a and b index the

components of X and Y respectively, $a = 1, 2, \dots, A$ and $b = 1, 2, \dots, B$. With X_1, X_2, X_3, Y_1, Y_2

and Y_3 denoting the first, second, and third components of X and Y respectively, then τ_{11} is the

probability of changing from X_1 to Y_1 , τ_{12} is the probability of changing from X_1 to Y_2 , and (1

$1 - \tau_{11} - \tau_{12} = \tau_{13}$ is the probability of changing from X1 to Y3. Similarly, τ_{2b} and τ_{3b} represent the probability of changing from X2 and X3, and $(1 - \tau_{a1} - \tau_{a2}) = \tau_{a3}$, so

$$\sum_{b=1}^B \tau_{ab} = 1 \quad (29)$$

for all a, leaving AB - A τ parameters free. To show the constraints on the τ , there are AB τ values; from Equation 29, this number is reduced to AB - A. From Equation 32 below,

$$\begin{aligned} f(y) &= \sum_x f(y|x) = \sum_x \sum_{a=1}^A \sum_{b=1}^B t_{ab} \pi_{xa} b(x; \theta_{xa}) b(y; \theta_{yb}) \\ &= \sum_{a=1}^A \sum_{b=1}^B t_{ab} \pi_{xa} b(y; \theta_{yb}) \sum_x b(x; \theta_{xa}) \\ &= \sum_{a=1}^A \sum_{b=1}^B \tau_{ab} \pi_{xa} b(y; \theta_{yb}), \end{aligned} \quad (30)$$

so

$$\sum_{a=1}^A \sum_{b=1}^B \tau_{ab} \pi_{xa} = \pi_{yb}. \quad (31)$$

This constraint forces B - 1 additional constraints, so the number of free parameters in the model is AB - A - (B-1) = (A - 1)(B - 1).

To find the joint distribution of X and Y,

$$\begin{aligned} f(x,y) &= f(y|x) f(x) = \\ &= \tau_{11} \pi_{x1} b(x; \theta_{x1}) b(y; \theta_{y1}) + \tau_{21} \pi_{x2} b(x; \theta_{x2}) b(y; \theta_{y1}) + \\ &\quad \tau_{31} (1 - \pi_{x1} - \pi_{x2}) b(x; \theta_{x3}) b(y; \theta_{y1}) + \tau_{12} \pi_{x1} b(x; \theta_{x1}) b(y; \theta_{y2}) + \end{aligned}$$

$$\begin{aligned}
& \tau_{22} \pi_{\underline{x}2} b(\underline{x}; \theta_{\underline{x}2}) b(\underline{y}; \theta_{\underline{y}2}) + \tau_{32} (1 - \pi_{\underline{x}1} - \pi_{\underline{x}2}) b(\underline{x}; \theta_{\underline{x}3}) b(\underline{y}; \theta_{\underline{y}2}) + \\
& \tau_{13} \pi_{\underline{x}1} b(\underline{x}; \theta_{\underline{x}1}) b(\underline{y}; \theta_{\underline{y}3}) + \tau_{23} \pi_{\underline{x}2} b(\underline{x}; \theta_{\underline{x}2}) b(\underline{y}; \theta_{\underline{y}3}) + \\
& \tau_{33} (1 - \pi_{\underline{x}1} - \pi_{\underline{x}2}) b(\underline{x}; \theta_{\underline{x}3}) b(\underline{y}; \theta_{\underline{y}3}) \\
& = \sum_{a=1}^A \sum_{b=1}^B \tau_{ab} \pi_{xa} b(\underline{x}; \theta_{xa}) b(\underline{y}; \theta_{yb})
\end{aligned} \tag{32}$$

The mean and variance of X and Y are given above by Equations 6 and 7, (p. 28). In the bivariate three component case

$$E(Y|x) = m[\pi_{y1|x} \theta_{y1} + \pi_{y2|x} \theta_{y2} + (1 - \pi_{y1|x} - \pi_{y2|x}) \theta_{y3}], \tag{33}$$

where m is the number of Bernoulli trials. $E(X|y)$ is calculated similarly with the role of X and Y interchanged. The $E(XY)$ is given by

$$E(XY) = m^2 \left[\sum_{a=1}^A \sum_{b=1}^B \tau_{ab} \pi_{xa} \theta_{xa} \theta_{yb} \right], \tag{34}$$

where m is the number of Bernoulli trials. From this, the correlation between X and Y can be calculated.

$$\rho = \text{corr}(X, Y) = \frac{E(XY) - E(X)E(Y)}{[V(X)V(Y)]^{1/2}}. \tag{35}$$

By substituting estimates for the parameters in Equation (32), the model correlation coefficient

$\hat{\rho}$ can be computed.

Model Chi-Square Goodness-of-Fit Statistics

Basic- Level Group	<u>Binomial Mixture</u>							
	<u>Normal Distribution</u>		<u>Two Component</u>			<u>Three Component</u>		
	χ^2	df	χ^2	L^2	df	χ^2	L^2	df
PFOS1	66.95*	12	6.49	7.27	6	6.49	7.27	4
PFOD1	111.79*	21	10100*	25.90	12	8.02	8.36	10
PFNS1	16.75	12	3.13	3.18	6	3.13	3.18	4
PFND1	21.33	21	1.53	1.65	12	1.53	1.65	10
PFOS2	63.637*	12	6.83	8.97	6	6.83	8.97	4
PFOD2	78.91*	21	21.10	16.80	12	20.00	16.60	10
PFNS2	30.68*	12	8.37	8.74	6	7.33	7.72	4
PFND2	28.63	21	4.75	4.87	12	4.74	4.87	10
PMOS1	275.26*	12	0.69	0.92	6	0.74	1.00	4
PMOD1	3724.0*	21	118000*	53.8*	12	22.90	17.80	10
PMNS1	101.25*	12	17.00	16.50	6	9.04	11.30	4
PMND1	140.64*	21	20.10	10.60	12	2.00	1.98	10
PMOS2	328.40*	12	18.40	17.70	6	18.20*	17.50*	4
PMOD2	317.44*	21	28.30*	27.30	12	9.51	10.20	10
PMNS2	144.11*	12	20.50*	23.70*	6	13.90	17.10*	4
PMND2	119.45*	21	2.88	3.46	12	2.89	3.46	10
CFOS1	15.00	12	3.57	3.82	6	3.41	3.69	4
CFOD1	33.70	21	19.60	12.30	12	14.30	11.70	10
CFNS1	8.99	12	15.40	11.20	6	6.49	7.13	4
CFND1	11.81	21	1.62	1.79	12	1.62	1.79	10
CFOS2	13.71	12	1.59	1.78	6	1.59	1.78	4
CFOD2	20.06	21	4.19	4.09	12	1.09	1.41	10
CFNS2	13.67	12	6.42	7.25	6	6.42	7.25	4
CFND2	11.73	21	0.73	0.99	12	0.60	0.88	10
CMOS1	115.58*	12	4.68	5.78	6	4.08	5.34	4
CMOD1	209.79*	21	557.00*	19.10	12	4.19	5.28	10
CMNS1	63.34*	12	190.00*	14.90	6	66.00*	9.71	4
CMND1	72.48*	21	11.50	11.10	12	8.13	8.10	10
CMOS2	113.63*	12	6.10	6.91	6	4.44	5.19	4
CMOD2	496.11*	21	78.90*	12.80	12	6.51	7.60	10
CMNS2	67.62*	12	75.90*	12.50	6	60.50*	12.00	4
CMND2	58.45*	21	0.43	0.66	12	0.44	0.67	10

Note: $\chi^2_{.05}(4) = 14.86$, $\chi^2_{.05}(6) = 18.54$, $\chi^2_{.05}(10) = 25.18$, $\chi^2_{.05}(12) = 28.30$, $\chi^2_{.05}(21) = 41.40$.

Test statistics with *'s represent χ^2 's significant at the $\alpha = .05$ level.

APPENDIX D
One, Two, Three and Four Component Model Solutions for All Basic-Level Groups

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
PFOS1	Penn State	Female	1	Old	Same	1	0.811 (0.016)								0.309	2.5E+04	104.0	14
						2	0.654 (0.020)	0.947 (0.009)			0.464 (0.056)	0.536 (0.056)			0.908	76.0	12.5	12
						3	0.338 (0.070)	0.698 (0.019)	0.953 (0.009)		0.038 (0.021)	0.467 (0.056)	0.495 (0.056)		0.999	5.1	5.6	10
						4	0.333 (0.071)	0.693 (0.020)	0.944 (0.010)	1.000 (0.001)	0.036 (0.021)	0.456 (0.056)	0.448 (0.056)	0.059 (0.026)	1.003	5.0	5.5	8
PFOD1					Different	1	0.761 (0.016)								0.316	2.3E+07	115.5	14
						2	0.576 (0.021)	0.887 (0.011)			0.404 (0.051)	0.596 (0.051)			0.885	1.0E+04	25.9	12
						3	0.000 (0.000)	0.606 (0.020)	0.895 (0.011)		0.011 (0.011)	0.428 (0.051)	0.561 (0.051)		0.962	8.0	8.4	10
						4	0.000 (0.000)	0.522 (0.032)	0.699 (0.021)	0.908 (0.011)	0.011 (0.011)	0.174 (0.039)	0.335 (0.049)	0.481 (0.052)	0.992	5.8	7.3	8
PFNS1					New Same	1	0.675 (0.018)								0.504	82.3	42.7	14
						2	0.600 (0.015)	0.878 (0.017)			0.732 (0.046)	0.268 (0.046)			0.992	3.1	3.2	12
						3	0.597 (0.031)	0.601 (0.018)	0.878 (0.017)		0.180 (0.040)	0.552 (0.052)	0.268 (0.046)		0.992	3.1	3.2	10
						4	0.599 (0.017)	0.604 (0.036)	0.878 (0.019)	0.878 (0.039)	0.602 (0.051)	0.130 (0.035)	0.217 (0.043)	0.051 (0.023)	0.992	3.1	3.2	8
PFND1					Different	1	0.677 (0.023)								0.542	43.9	27.6	8
						2	0.491 (0.027)	0.802 (0.018)			0.401 (0.051)	0.599 (0.051)			1.001	1.5	1.6	6
						3	0.491 (0.045)	0.492 (0.034)	0.802 (0.018)		0.149 (0.037)	0.252 (0.045)	0.599 (0.051)		1.001	1.5	1.6	4
						4	0.491 (0.027)	0.802 (0.024)	0.802 (0.029)	0.802 (0.073)	0.401 (0.051)	0.338 (0.049)	0.226 (0.043)	0.036 (0.019)	1.001	1.5	1.6	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
PFOS2			2	Old	Same	1	0.845 (0.020)								0.456	197.7	40.3	8
						2	0.630 (0.035)	0.933 (0.012)			0.290 (0.054)	0.710 (0.054)			0.983	6.8	9.0	6
						3	0.536 (0.074)	0.678 (0.038)	0.936 (0.012)		0.072 (0.031)	0.242 (0.051)	0.687 (0.055)		0.993	6.7	8.9	4
						4	0.630 (0.035)	0.933 (0.026)	0.933 (0.020)	0.933 (0.017)	0.290 (0.054)	0.140 (0.041)	0.244 (0.051)	0.325 (0.056)	0.983	6.8	9.0	2
PFOD2					Different	1	0.801 (0.017)								0.242	1.7E+04	151.3	14
						2	0.568 (0.025)	0.942 (0.009)			0.376 (0.058)	0.624 (0.058)			0.936	21.1	16.8	12
						3	0.472 (0.036)	0.682 (0.030)	0.947 (0.009)		0.180 (0.046)	0.229 (0.050)	0.591 (0.059)		0.992	12.3	14.4	10
						4	0.472 (0.036)	0.682 (0.030)	0.947 (0.013)	0.947 (0.013)	0.181 (0.046)	0.228 (0.050)	0.298 (0.055)	0.293 (0.054)	0.992	12.3	14.4	8
PFNS2					New	1	0.749 (0.019)								0.359	255.3	75.0	14
						2	0.593 (0.022)	0.901 (0.013)			0.495 (0.060)	0.505 (0.060)			0.990	8.4	8.7	12
						3	0.593 (0.045)	0.594 (0.025)	0.901 (0.013)		0.112 (0.038)	0.383 (0.058)	0.505 (0.060)		0.990	8.4	8.7	10
						4	0.587 (0.022)	0.880 (0.022)	0.880 (0.020)	1.000 (0.000)	0.473 (0.060)	0.207 (0.048)	0.261 (0.053)	0.059 (0.028)	1.007	7.3	7.7	8
PFND2					Different	1	0.775 (0.024)								0.548	43.1	22.7	8
						2	0.597 (0.033)	0.875 (0.016)			0.361 (0.057)	0.639 (0.057)			0.997	4.7	4.9	6
						3	0.596 (0.064)	0.597 (0.038)	0.875 (0.016)		0.092 (0.035)	0.269 (0.053)	0.639 (0.057)		0.997	4.7	4.9	4
						4	0.597 (0.033)	0.875 (0.025)	0.875 (0.026)	0.875 (0.039)	0.361 (0.057)	0.270 (0.053)	0.254 (0.052)	0.115 (0.038)	0.997	4.7	4.9	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
PMOS1		Male	1	Old	Same	1	0.917 (0.008)								0.579	82.1	55.5	8
						2	0.807 (0.014)	0.981 (0.004)			0.365 (0.030)	0.635 (0.030)			1.004	0.7	0.9	6
						3	0.807 (0.038)	0.807 (0.014)	0.981 (0.004)		0.045 (0.013)	0.320 (0.029)	0.635 (0.030)		1.004	0.7	0.9	4
						4	0.807 (0.023)	0.807 (0.017)	0.975 (0.009)	0.982 (0.004)	0.129 (0.021)	0.236 (0.026)	0.131 (0.021)	0.505 (0.031)	1.004	0.7	0.9	2
PMOD1					Different	1	0.855 (0.008)								0.289	1.5E+10	317.4	14
						2	0.657 (0.014)	0.942 (0.004)			0.305 (0.029)	0.695 (0.029)			0.853	1.2E+05	53.8	12
						3	0.242 (0.047)	0.723 (0.012)	0.954 (0.004)		0.021 (0.009)	0.362 (0.030)	0.617 (0.030)		0.957	22.8	17.8	10
						4	0.076 (0.048)	0.577 (0.022)	0.828 (0.009)	0.975 (0.004)	8.035E-05 (0.006)	0.128 (0.021)	0.418 (0.031)	0.447 (0.031)	0.989	7.7	8.6	8
PMNS1				New	Same	1	0.779 (0.009)								0.382	640.1	218.4	14
						2	0.655 (0.010)	0.927 (0.006)			0.545 (0.031)	0.455 (0.031)			0.952	17.0	16.5	12
						3	0.517 (0.024)	0.711 (0.010)	0.938 (0.006)		0.114 (0.020)	0.489 (0.031)	0.398 (0.030)		1.008	9.0	11.3	10
						4	0.517 (0.024)	0.711 (0.010)	0.938 (0.008)	0.938 (0.010)	0.114 (0.020)	0.488 (0.031)	0.232 (0.026)	0.165 (0.023)	1.008	9.0	11.3	8
PMND1					Different	1	0.788 (0.012)								0.437	4.9E+03	133.2	8
						2	0.542 (0.019)	0.885 (0.008)			0.285 (0.028)	0.715 (0.028)			0.939	20.2	10.6	6
						3	0.336 (0.037)	0.704 (0.014)	0.924 (0.008)		0.071 (0.016)	0.430 (0.031)	0.499 (0.031)		0.995	2.0	2.0	4
						4	0.328 (0.038)	0.685 (0.016)	0.878 (0.012)	0.947 (0.009)	0.066 (0.015)	0.372 (0.030)	0.307 (0.029)	0.256 (0.027)	0.996	1.9	2.0	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
PMOS2			2	Old	Same	1	0.889 (0.010)								0.343	1.3E+03	198.7	8
						2	0.540 (0.029)	0.960 (0.005)			0.169 (0.026)	0.831 (0.026)			1.029	18.4	17.7	6
						3	0.540 (0.049)	0.540 (0.035)	0.960 (0.005)		0.058 (0.017)	0.111 (0.022)	0.831 (0.026)		1.029	18.4	17.7	4
						4	0.532 (0.029)	0.948 (0.010)	0.948 (0.009)	0.995 (0.004)	0.163 (0.026)	0.273 (0.032)	0.377 (0.034)	0.188 (0.028)	1.040	18.2	17.5	2
PMOD2					Different	1	0.865 (0.009)								0.225	6.7E+04	387.5	14
						2	0.492 (0.023)	0.937 (0.005)			0.163 (0.026)	0.837 (0.026)			0.952	28.3	27.3	12
						3	0.457 (0.024)	0.869 (0.010)	0.976 (0.004)		0.139 (0.024)	0.365 (0.034)	0.496 (0.035)		1.012	9.5	10.2	10
						4	0.457 (0.024)	0.869 (0.015)	0.870 (0.014)	0.976 (0.004)	0.139 (0.024)	0.170 (0.027)	0.196 (0.028)	0.495 (0.035)	1.012	9.5	10.2	8
PMNS2					New	1	0.817 (0.010)								0.307	2.3E+03	255.7	14
						2	0.560 (0.018)	0.907 (0.006)			0.260 (0.031)	0.741 (0.031)			0.970	20.5	23.7	12
						3	0.543 (0.032)	0.569 (0.021)	0.907 (0.006)		0.083 (0.019)	0.178 (0.027)	0.740 (0.031)		0.970	20.5	23.7	10
						4	0.530 (0.020)	0.860 (0.013)	0.860 (0.012)	0.964 (0.007)	0.217 (0.029)	0.247 (0.030)	0.262 (0.031)	0.273 (0.032)	1.018	13.9	17.1	8
PMND2					Different	1	0.836 (0.012)								0.562	72.5	55.2	8
						2	0.687 (0.017)	0.931 (0.008)			0.392 (0.035)	0.608 (0.035)			1.026	2.9	3.5	6
						3	0.687 (0.039)	0.687 (0.020)	0.931 (0.008)		0.080 (0.019)	0.312 (0.033)	0.608 (0.035)		1.026	2.9	3.5	4
						4	0.687 (0.021)	0.687 (0.033)	0.931 (0.011)	0.931 (0.010)	0.282 (0.032)	0.110 (0.022)	0.284 (0.032)	0.324 (0.033)	1.026	2.9	3.5	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
CFOS1	Cooper-Union	Female	1	Old	Same	1	0.849 (0.032)								0.733	7.8	5.4	8
						2	0.596 (0.108)	0.872 (0.022)			0.083 (0.052)	0.917 (0.052)			0.999	3.6	3.8	6
						3	0.596 (0.227)	0.596 (0.122)	0.872 (0.022)		0.019 (0.026)	0.064 (0.046)	0.917 (0.052)		0.999	3.6	3.8	4
						4	0.582 (0.119)	0.859 (0.032)	0.859 (0.034)	1.000 (0.001)	0.068 (0.048)	0.461 (0.094)	0.406 (0.093)	0.065 (0.046)	1.026	3.4	3.7	2
CFOD1					Different	1	0.817 (0.027)								0.320	7.5E+03	36.0	14
						2	0.592 (0.046)	0.900 (0.017)			0.270 (0.084)	0.730 (0.084)			0.878	19.6	12.3	12
						3	0.318 (0.107)	0.689 (0.041)	0.912 (0.017)		0.046 (0.039)	0.306 (0.087)	0.648 (0.090)		0.986	8.4	9.6	10
						4	0.318 (0.107)	0.689 (0.041)	0.912 (0.018)	0.912 (0.057)	0.046 (0.039)	0.306 (0.087)	0.589 (0.093)	0.059 (0.044)	0.986	8.4	9.6	8
CFNS1					New	1	0.693 (0.032)								0.703	15.6	11.2	14
						2	0.650 (0.026)	0.868 (0.037)			0.804 (0.075)	0.196 (0.075)			1.049	6.8	7.4	12
						3	0.650 (0.062)	0.650 (0.029)	0.868 (0.037)		0.143 (0.066)	0.661 (0.089)	0.196 (0.075)		1.049	6.8	7.4	10
						4	0.647 (0.031)	0.647 (0.051)	0.834 (0.040)	1.000 (0.000)	0.560 (0.094)	0.212 (0.077)	0.207 (0.077)	0.021 (0.027)	1.060	6.5	7.1	8
CFND1					Different	1	0.698 (0.041)								0.450	28.7	17.7	8
						2	0.562 (0.040)	0.921 (0.028)			0.621 (0.092)	0.380 (0.092)			0.969	1.6	1.8	6
						3	0.561 (0.069)	0.563 (0.048)	0.921 (0.028)		0.204 (0.076)	0.417 (0.093)	0.380 (0.092)		0.969	1.6	1.8	4
						4	0.562 (0.045)	0.563 (0.082)	0.921 (0.039)	0.921 (0.039)	0.477 (0.094)	0.144 (0.066)	0.193 (0.075)	0.187 (0.074)	0.969	1.6	1.8	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
CFOS2			2	Old	Same	1	0.860 (0.031)								0.757	3.9	2.9	8
						2	0.755 (0.050)	0.905 (0.022)			0.300 (0.088)	0.701 (0.088)			0.997	1.6	1.8	6
						3	0.755 (0.140)	0.755 (0.054)	0.905 (0.022)		0.039 (0.037)	0.261 (0.085)	0.700 (0.088)		0.997	1.6	1.8	4
						4	0.755 (0.050)	0.905 (0.039)	0.905 (0.036)	0.905 (0.043)	0.300 (0.088)	0.229 (0.081)	0.279 (0.086)	0.192 (0.076)	0.997	1.6	1.8	2
CFOD2					Different	1	0.832 (0.026)								0.315	264.3	29.6	14
						2	0.602 (0.048)	0.913 (0.016)			0.260 (0.084)	0.740 (0.084)			0.902	4.2	4.1	12
						3	0.544 (0.058)	0.853 (0.024)	0.985 (0.012)		0.179 (0.074)	0.557 (0.096)	0.264 (0.085)		0.987	1.1	1.4	10
						4	0.544 (0.058)	0.853 (0.033)	0.853 (0.033)	0.985 (0.012)	0.180 (0.074)	0.278 (0.086)	0.279 (0.086)	0.263 (0.085)	0.987	1.1	1.4	8
CFNS2					New	1	0.701 (0.032)								0.407	56.6	29.3	14
						2	0.606 (0.029)	0.924 (0.024)			0.701 (0.088)	0.300 (0.088)			0.981	6.4	7.2	12
						3	0.606 (0.058)	0.606 (0.034)	0.924 (0.024)		0.177 (0.073)	0.524 (0.096)	0.300 (0.088)		0.981	6.4	7.2	10
						4	0.606 (0.032)	0.606 (0.067)	0.924 (0.030)	0.924 (0.041)	0.568 (0.095)	0.133 (0.065)	0.194 (0.076)	0.105 (0.059)	0.981	6.4	7.2	8
CFND2					Different	1	0.696 (0.042)								0.396	36.8	21.4	8
						2	0.495 (0.046)	0.887 (0.028)			0.489 (0.096)	0.511 (0.096)			0.970	0.7	1.0	6
						3	0.495 (0.071)	0.495 (0.060)	0.887 (0.028)		0.202 (0.077)	0.287 (0.087)	0.511 (0.096)		0.970	0.7	1.0	4
						4	0.485 (0.047)	0.862 (0.044)	0.862 (0.047)	1.000 (0.006)	0.463 (0.096)	0.255 (0.084)	0.222 (0.080)	0.059 (0.046)	0.983	0.6	0.9	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
CMOS1		Male	1	Old	Same	1	0.893 (0.017)								0.471	616.2	31.2	8
						2	0.594 (0.054)	0.935 (0.010)			0.122 (0.038)	0.878 (0.038)			0.963	4.7	5.8	6
						3	0.534 (0.068)	0.866 (0.022)	0.960 (0.010)		0.080 (0.031)	0.348 (0.055)	0.573 (0.057)		0.990	4.2	5.4	4
						4	0.550 (0.063)	0.901 (0.021)	0.905 (0.020)	0.985 (0.009)	0.092 (0.033)	0.296 (0.053)	0.336 (0.055)	0.277 (0.052)	0.989	4.1	5.3	2
CMOD1					Different	1	0.856 (0.015)								0.318	5.9E+05	83.2	14
						2	0.709 (0.021)	0.956 (0.008)			0.404 (0.057)	0.596 (0.057)			0.850	559.7	19.1	12
						3	0.274 (0.081)	0.762 (0.019)	0.965 (0.008)		0.027 (0.019)	0.444 (0.057)	0.530 (0.058)		1.015	4.2	5.3	10
						4	0.274 (0.081)	0.762 (0.019)	0.964 (0.014)	0.965 (0.009)	0.027 (0.019)	0.444 (0.057)	0.166 (0.043)	0.363 (0.056)	1.015	4.2	5.3	8
CMNS1					New	1	0.795 (0.017)								0.306	1.7E+05	95.7	14
						2	0.612 (0.023)	0.917 (0.011)			0.401 (0.057)	0.599 (0.057)			0.894	190.4	14.9	12
						3	0.139 (0.089)	0.646 (0.022)	0.926 (0.010)		0.014 (0.013)	0.430 (0.057)	0.556 (0.057)		0.959	7.4	7.9	10
						4	0.138 (0.089)	0.615 (0.025)	0.877 (0.014)	1.000 (0.001)	0.013 (0.013)	0.344 (0.055)	0.499 (0.058)	0.144 (0.040)	0.998	3.8	4.4	8
CMND1					Different	1	0.825 (0.021)								0.407	1.0E+03	52.1	8
						2	0.487 (0.047)	0.895 (0.013)			0.170 (0.043)	0.830 (0.043)			0.936	11.5	11.1	6
						3	0.392 (0.059)	0.834 (0.017)	0.999 (0.003)		0.100 (0.035)	0.684 (0.054)	0.216 (0.048)		0.978	8.1	8.1	4
						4	0.332 (0.068)	0.691 (0.046)	0.865 (0.017)	1.000 (0.000)	0.070 (0.030)	0.151 (0.041)	0.599 (0.057)	0.179 (0.044)	0.998	7.5	7.8	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df
CMOS2			2	Old	Same	1	0.914 (0.015)								0.596	121.0	17.8	8
						2	0.702 (0.051)	0.941 (0.009)			0.114 (0.035)	0.887 (0.035)			0.942	6.1	6.9	6
						3	0.543 (0.098)	0.900 (0.013)	1.000 (0.001)		0.036 (0.021)	0.695 (0.052)	0.270 (0.050)		1.022	4.4	5.2	4
						4	0.544 (0.098)	0.900 (0.020)	0.900 (0.019)	1.000 (0.001)	0.036 (0.021)	0.329 (0.053)	0.366 (0.054)	0.269 (0.050)	1.022	4.4	5.2	2
CMOD2					Different	1	0.899 (0.012)								0.378	6.5E+04	59.9	14
						2	0.759 (0.022)	0.965 (0.006)			0.318 (0.052)	0.682 (0.052)			0.916	79.2	12.8	12
						3	0.378 (0.113)	0.797 (0.019)	0.970 (0.006)		0.015 (0.014)	0.356 (0.054)	0.629 (0.054)		1.022	6.5	7.6	10
						4	0.378 (0.113)	0.797 (0.019)	0.970 (0.012)	0.970 (0.007)	0.015 (0.014)	0.356 (0.054)	0.182 (0.043)	0.447 (0.056)	1.022	6.5	7.6	8
CMNS2					New	1	0.866 (0.017)								0.596	55.6	22.5	8
						2	0.726 (0.027)	0.933 (0.011)			0.323 (0.048)	0.677 (0.048)			0.981	6.5	7.3	6
						3	0.527 (0.096)	0.784 (0.022)	0.944 (0.011)		0.032 (0.018)	0.401 (0.051)	0.567 (0.051)		1.002	6.2	7.0	4
						4	0.527 (0.095)	0.784 (0.022)	0.944 (0.016)	0.944 (0.014)	0.033 (0.018)	0.402 (0.051)	0.234 (0.044)	0.332 (0.049)	1.002	6.2	7.0	2
CMND2					Different	1	0.840 (0.019)								0.493	50.2	31.9	8
						2	0.670 (0.029)	0.942 (0.011)			0.374 (0.054)	0.627 (0.054)			1.003	0.4	0.7	6
						3	0.669 (0.061)	0.670 (0.033)	0.942 (0.011)		0.083 (0.031)	0.290 (0.051)	0.627 (0.054)		1.003	0.4	0.7	4
						4	0.670 (0.029)	0.941 (0.027)	0.942 (0.019)	0.942 (0.016)	0.374 (0.054)	0.109 (0.035)	0.212 (0.046)	0.306 (0.052)	1.003	0.4	0.7	2

Z-scores for Penn State and Cooper-Union Males' and Females' Theta Estimates

	PFOS1	PFOD1	PFNS1	PFND1	PFOS2	PFOD2	PFNS2	PFND2
PFOS1		4.40*	4.04*	6.10*	2.15*	4.29*	3.83*	3.05*
PFOD1	3.01*		-0.94	2.45*	-1.32	0.23	-0.60	-0.55
PFNS1	2.75*	6.35*		3.47*	-0.78	1.09	0.25	0.09
PFND1	-0.01	-0.61	1.94		-3.10*	-2.08*	-2.94*	-2.48*
PFOS2	2.96*	0.45	4.08*	-2.85*		1.43	0.88	0.69
PFOD2	-3.79*	-0.79	0.61	3.11*	-2.67*		-0.77	-0.70
PFNS2	-3.29*	-1.05	0.13	-6.16*	-6.97*	-4.48*		-0.09
PFND2	-3.02*	-0.56	1.85	2.86*	2.57*	3.52*	1.21	

	PMOS1	PMOD1	PMNS1	PMND1	PMOS2	PMOD2	PMNS2	PMND2
PMOS1		7.76*	8.90*	11.20*	8.44*	11.90*	11.00*	5.43*
PMOD1	6.66*		0.09	4.83*	3.68*	6.22*	4.30*	-1.35
PMNS1	7.45*	11.10*		5.16*	3.79*	6.56*	4.63*	-1.56
PMND1	3.33*	7.21*	10.30*		0.06	1.68	-0.68	-5.56*
PMOS2	5.84*	1.97	6.29*	-2.59*		1.32	-0.59	-4.39*
PMOD2	0.78	4.65*	1.24	4.18*	-4.07*		-2.36*	-6.82*
PMNS2	-1.25	2.35*	-0.41	-7.97*	-5.61*	-2.13*		-5.10*
PMND2	-4.18*	3.24*	6.64*	3.11*	3.87*	0.64	-2.50*	

	CFOS1	CFOD1	CFNS1	CFND1	CFOS2	CFOD2	CFNS2	CFND2
CFOS1		0.03	-0.49	0.29	-1.34	-0.06	-0.10	0.86
CFOD1	-0.99		-1.09	0.49	-2.37*	-0.15	-0.26	1.49
CFNS1	0.10	-1.39		1.85	-1.84	0.88	1.13	2.94*
CFND1	-1.05	-1.49	-1.57		-3.00*	-0.64	-0.90	1.11
CFOS2	-0.42	0.77	-0.67	-0.20		2.20*	2.55*	3.81*
CFOD2	-0.56	-0.81	0.38	-1.15	-0.86		-0.07	1.62
CFNS2	-1.10	-1.25	-0.41	0.45	0.26	-0.06		2.05*
CFND2	0.86	-0.27	-0.55	0.50	-0.37	0.78	0.98	

	CMOS1	CMOD1	CMNS1	CMND1	CMOS2	CMOD2	CMNS2	CMND2
CMOS1		-1.98	-0.31	1.50	-1.47	-2.82*	-1.04	-1.24
CMOD1	-1.62		3.09*	4.33*	0.12	-1.63	1.86	1.10
CMNS1	1.22	2.46*		2.41*	-1.63	-4.62*	-1.38	-1.57
CMND1	-0.43	-2.47*	-0.86		-3.13*	-5.27*	-3.29*	-3.34*
CMOS2	-0.47	2.92*	4.03*	1.22		-1.02	0.89	0.57
CMOD2	-0.87	0.78	1.02	1.35	-1.68		3.51*	2.47*
CMNS2	-3.83*	-2.13*	-1.63	-2.90*	-4.84*	-3.31*		-0.44
CMND2	-2.79*	-2.09*	-0.43	-0.08	1.65	1.77	0.32	

Note: Comparisons above the diagonal are for the "Low" performing component. Asterisks indicate significant differences at the $\alpha = .05$ level.

Z-scores for Penn State and Cooper-Union Males' and Females' Theta Estimates
within Item-type by Item-status by Time Grouping

OS1					OD1				
	PF	PM	CF	CM		PF	PM	CF	CM
PF		-2.68*	1.18	2.18*	PF		-3.26*	-0.33	-4.47*
PM	-4.28*		-1.95	-3.82*	PM	-4.63*		-1.34	-2.04*
CF	2.50*	-4.87*		-0.01	CF	-0.61	2.40*		-2.29*
CM	-0.14	-4.24*	-2.60*		CM	-5.07*	-1.50	-2.98*	

NS1					ND1				
	PF	PM	CF	CM		PF	PM	CF	CM
PF		-3.00*	-1.66	-0.44	PF		-1.52	-1.47	0.08
PM	-2.72*		0.19	1.73	PM	-4.30*		-0.46	1.09
CF	0.25	1.57		1.10	CF	-3.64*	-1.25		1.23
CM	-1.95	0.81	-1.27		CM	-4.21*	-0.60	0.87	

OS2					OD2				
	PF	PM	CF	CM		PF	PM	CF	CM
PF		1.97	-2.02*	-1.17	PF		2.26*	-0.63	-5.74*
PM	-2.08*		-3.70*	-2.79*	PM	0.434		-2.08*	-8.46*
CF	1.10	2.37*		0.73	CF	1.54	1.42		-2.98*
CM	-0.523	1.77	-1.47		CM	-2.01*	-3.43*	-2.96*	

NS2					ND2				
	PF	PM	CF	CM		PF	PM	CF	CM
PF		1.19	-0.36	-2.06*	PF		-2.45*	1.81	-1.68
PM	-0.41		-1.36	-3.50*	PM	-3.08*		3.91*	0.52
CF	-0.83	-0.68		-1.35	CF	-0.36	1.5		-3.22*
CM	-2.92*	-3.69*	-0.90		CM	-3.37*	-0.81	-1.80	

Note: Comparisons above the diagonal are for the "Low" performing component, while those below the diagonal are for the "High" performing component. Asterisks indicate significant differences at the alpha = 0.05 level.

APPENDIX G

Two Component Restricted Model Estimates and Fit Statistics

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PFOS1	Penn State	Female	1	Old	Same	2	0.726	0.933		0.323	0.677		0.981	6.5	7.3	6
							(0.027)	(0.011)		(0.048)	(0.048)					
							0.628	0.934		0.268	0.732		1.000	9.5	11.6	8
							(0.032)	(0.010)		(0.046)	(0.046)					
PFOD1					Different	2									4.3	2 *
							0.576	0.887		0.404	0.596		0.885	1.0E+04	25.9	12
							(0.021)	(0.011)		(0.051)	(0.051)					
							0.628	0.934		0.529	0.471		0.870	56900.0	34.7	14
							(0.018)	(0.010)		(0.052)	(0.052)					
															8.8	2
PFNS1				New	Same	2	0.600	0.878		0.732	0.268		0.992	3.1	3.2	12
							(0.015)	(0.017)		(0.046)	(0.046)					
							0.628	0.934		0.803	0.197		0.965	7.9	7.6	14
							(0.014)	(0.015)		(0.041)	(0.041)					
															4.4	2 *
							0.491	0.802		0.401	0.599		1.000	1.5	1.6	6
							(0.027)	(0.018)		(0.051)	(0.051)					
PFND1					Different	2	0.628	0.934		0.754	0.246		0.860	19.0	15.2	8
							(0.019)	(0.017)		(0.045)	(0.045)					
															13.6	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_1$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PFOS2			2	Old	Same	2	0.630	0.933		0.290	0.710		0.983	6.8	9.0	6
							(0.035)	(0.012)		(0.054)	(0.054)					
							0.628	0.934		0.312	0.688		1.000	6.4	9.1	8
							(0.034)	(0.012)		(0.055)	(0.055)					
PFOD2					Different	2									0.1	2 *
							0.568	0.942		0.376	0.624		0.936	21.1	16.8	12
							(0.025)	(0.009)		(0.058)	(0.058)					
							0.628	0.934		0.393	0.607		0.701	65.2	22.8	14
							(0.024)	(0.010)		(0.058)	(0.058)					
															6.0	2 *
PFNS2				New	Same	2	0.593	0.901		0.495	0.505		0.990	8.4	8.7	12
							(0.022)	(0.013)		(0.060)	(0.060)					
							0.628	0.934		0.567	0.433		0.959	16.0	12.9	14
							(0.020)	(0.012)		(0.059)	(0.059)					
															4.2	2 *
							0.594	0.875		0.361	0.639		0.997	4.7	4.9	6
							(0.033)	(0.016)		(0.057)	(0.057)					
PFND2					Different	2	0.628	0.934		0.519	0.481		1.000	9.2	9.6	8
							(0.027)	(0.014)		(0.060)	(0.060)					
															4.7	2 *

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PMOS1		Male	1	Old	Same	2	0.807	0.981		0.365	0.635		1.000	0.7	0.9	6
							(0.014)	(0.004)		(0.030)	(0.300)					
							0.628	0.934		0.137	0.863		1.000	22.3	26.9	8
							(0.027)	(0.006)		(0.021)	(0.021)					
															26.0	2
PMOD1					Different	2	0.657	0.942		0.305	0.695		0.853	1.2E+05	53.80	12
							(0.014)	(0.004)		(0.029)	(0.029)					
							0.628	0.934		0.279	0.721		0.913	3.9E+04	55.8	14
							(0.015)	(0.005)		(0.028)	(0.028)					
															2.0	2 *
PMNS1				New	Same	2	0.655	0.927		0.545	0.455		0.952	17.0	16.5	12
							(0.010)	(0.006)		(0.031)	(0.031)					
							0.628	0.934		0.531	0.469		1.000	20.4	23.1	14
							(0.011)	(0.006)		(0.031)	(0.031)					
															6.6	2
PMND1					Different	2	0.542	0.885		0.28	0.72		0.939	20.2	10.6	6
							(0.019)	(0.008)		(0.028)	(0.028)					
							0.628	0.934		0.464	0.536		0.915	77.6	21.1	8
							(0.015)	(0.007)		(0.031)	(0.031)					
															10.5	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PMOS2			2	Old	Same	2	0.540	0.960		0.169	0.831		1.000	18.4	17.7	6
							(0.029)	(0.005)		(0.026)	(0.026)					
							0.628	0.934		0.185	0.815		0.765	51.9	44.6	8
							(0.026)	(0.007)		(0.027)	(0.027)					
															26.9	2
PMOD2					Different	2	0.492	0.937		0.163	0.837		0.952	28.3	27.3	12
							(0.023)	(0.005)		(0.026)	(0.026)					
							0.628	0.934		0.209	0.792		0.633	86.6	55.3	14
							(0.019)	(0.005)		(0.029)	(0.029)					
															28.0	2
PMNS2				New	Same	2	0.560	0.907		0.260	0.741		0.970	20.5	23.7	12
							(0.018)	(0.006)		(0.031)	(0.031)					
							0.628	0.934		0.355	0.645		0.908	36.2	36.3	14
							(0.015)	(0.006)		(0.034)	(0.034)					
															12.6	2
PMND2					Different	2	0.687	0.931		0.392	0.608		1.000	2.9	3.4	6
							(0.017)	(0.008)		(0.035)	(0.035)					
							0.628	0.934		0.358	0.642		1.000	6.8	8.8	8
							(0.019)	(0.007)		(0.034)	(0.034)					
															5.4	2

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CFOS1	Cooper-Union	Female	1	Old	Same	2	0.596	0.872		0.083	0.917		0.999	3.6	3.8	6
							(0.108)	(0.022)		(0.052)	(0.052)					
							0.628	0.934		0.320	0.680		1.000	11.1	9.9	8
							(0.054)	(0.019)		(0.088)	(0.088)					
															6.1	2
CFOD1					Different	2	0.592	0.900		0.270	0.730		0.878	19.6	12.3	12
							(0.046)	(0.017)		(0.084)	(0.084)					
							0.628	0.934		0.357	0.643		0.951	32.2	14.5	14
							(0.039)	(0.015)		(0.091)	(0.091)					
															2.2	2 *
CFNS1				New	Same	2	0.650	0.868		0.804	0.196		1.000	6.8	7.4	12
							(0.026)	(0.037)		(0.075)	(0.075)					
							0.628	0.934		0.830	0.170		1.000	9.4	10.1	14
							(0.026)	(0.029)		(0.071)	(0.071)					
															2.7	2 *
CFND1					Different	2	0.562	0.927		0.621	0.380		0.969	1.6	1.8	6
							(0.040)	(0.028)		(0.092)	(0.092)					
							0.628	0.934		0.673	0.327		0.774	4.5	3.5	8
							(0.037)	(0.027)		(0.089)	(0.089)					
															1.7	2 *

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CFOS2			2	Old	Same	2	0.755	0.905		0.300	0.701		0.997	1.6	1.8	6
							(0.050)	(0.022)		(0.088)	(0.088)					
							0.628	0.934		0.279	0.721		1.000	4.1	5.5	8
							(0.059)	(0.019)		(0.086)	(0.086)					
															3.7	2 *
CFOD2					Different	2	0.602	0.913		0.260	0.740		0.902	4.2	4.1	12
							(0.047)	(0.016)		(0.084)	(0.084)					
							0.628	0.934		0.337	0.663		0.975	5.3	4.8	14
							(0.041)	(0.015)		(0.091)	(0.091)					
															0.7	2 *
CFNS2				New	Same	2	0.606	0.924		0.700	0.300		0.981	6.4	7.2	12
							(0.029)	(0.024)		(0.088)	(0.088)					
							0.628	0.934		0.717	0.283		0.910	6.9	7.7	14
							(0.028)	(0.023)		(0.087)	(0.087)					
															0.5	2 *
CFND2					Different	2	0.495	0.887		0.489	0.511		0.970	0.7	1.0	6
							(0.046)	(0.028)		(0.096)	(0.096)					
							0.628	0.934		0.647	0.353		0.682	7.9	5.5	8
							(0.039)	(0.027)		(0.092)	(0.092)					
															4.5	2 *

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CMOS1		Male	1	Old	Same	2	0.594	0.935		0.122	0.878		0.963	4.7	5.8	6
							(0.054)	(0.010)		(0.038)	(0.038)					
							0.628	0.934		0.172	0.828		1.000	5.3	6.6	8
							(0.045)	(0.011)		(0.044)	(0.044)					
															0.8	2 *
CMOD1					Different	2	0.709	0.956		0.404	0.596		0.850	559.7	19.1	12
							(0.021)	(0.008)		(0.057)	(0.057)					
							0.628	0.934		0.301	0.699		1.000	70.5	24.3	14
							(0.026)	(0.009)		(0.053)	(0.053)					
															5.2	2 *
CMNS1				New	Same	2	0.612	0.917		0.401	0.599		0.894	190.4	14.9	12
							(0.023)	(0.011)		(0.057)	(0.057)					
							0.628	0.934		0.449	0.551		0.910	281.2	15.9	14
							(0.022)	(0.010)		(0.057)	(0.057)					
															1.0	2 *
CMND1					Different	2	0.487	0.895		0.170	0.830		0.936	11.5	11.1	6
							(0.047)	(0.013)		(0.043)	(0.043)					
							0.628	0.934		0.352	0.648		0.884	26.8	15.9	8
							(0.031)	(0.012)		(0.055)	(0.055)					
															4.8	2 *

Group	Curriculum	Sex	Time	Type	Status	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CMOS2			2	Old	Same	2	0.702	0.941		0.114	0.887		0.942	6.1	6.9	6
							(0.051)	(0.009)		(0.035)	(0.035)					
							0.628	0.934		0.124	0.876		1.000	5.5	8.6	8
							(0.051)	(0.010)		(0.037)	(0.037)					
CMOD2					Different	2									1.7	2 *
							0.759	0.965		0.318	0.682		0.916	79.2	12.8	12
							(0.022)	(0.006)		(0.052)	(0.052)					
							0.628	0.934		0.175	0.826		1.000	23.4	24.2	14
							(0.033)	(0.008)		(0.042)	(0.042)					
															11.4	2
CMNS2				New	Same	2	0.654	0.947		0.464	0.536		0.908	76.0	12.5	12
							(0.020)	(0.009)		(0.056)	(0.056)					
							0.628	0.934		0.430	0.570		0.963	43.3	13.9	14
							(0.021)	(0.010)		(0.055)	(0.055)					
															1.4	2 *
							0.670	0.942		0.374	0.626		1.000	0.4	0.7	6
							(0.029)	(0.011)		(0.054)	(0.054)					
CMND2					Different	2										
							0.628	0.934		0.340	0.660		1.000	0.9	1.4	8
							(0.031)	(0.011)		(0.053)	(0.053)					
															0.7	2 *

Z-scores for Penn State and Cooper-Union Males' and Females' Pi Estimates

	PFOS1	PFOD1	PFNS1	PFND1	PFOS2	PFOD2	PFNS2	PFND2
PFOS1		-1.16	-6.13*	-1.11	0.45	-0.71	-2.24*	-0.51
PFOD1	-1.16		-4.78*	0.05	1.54	0.36	-1.16	0.56
PFNS1	-6.13*	-4.78*		4.83*	6.24*	4.81*	3.14*	5.04*
PFND1	-1.11	0.05	4.83*		1.49	0.32	-1.20	0.51
PFOS2	0.45	1.54	6.24*	1.49		-1.09	-2.54*	-0.90
PFOD2	-0.71	0.36	4.81*	0.32	-1.09		-1.43	0.18
PFNS2	-2.24*	-1.16	3.14*	-1.20	-2.54*	-1.43		1.61
PFND2	-0.51	0.56	5.04*	0.51	-0.90	0.18	1.61	
	PMOS1	PMOD1	PMNS1	PMND1	PMOS2	PMOD2	PMNS2	PMND2
PMOS1		1.45	-4.18*	1.94	4.89*	5.08*	2.44*	-0.59
PMOD1	1.45		-5.69*	0.50	3.48*	3.67*	1.07	-1.94
PMNS1	-4.18*	-5.69*		6.22*	9.22*	9.43*	6.51*	3.30*
PMND1	1.94	0.50	6.22*		3.00*	3.19*	0.60	-2.40*
PMOS2	4.89*	3.48*	9.22*	3.00*		0.17	-2.22*	-5.12*
PMOD2	5.08*	3.67*	9.43*	3.19*	0.17		-2.39*	-5.30*
PMNS2	2.44*	1.07	6.51*	0.60	-2.22*	-2.39*		-2.85*
PMND2	-0.59	-1.94	3.30*	-2.40*	-5.12*	-5.30*	-2.85*	
	CFOS1	CFOD1	CFNS1	CFND1	CFOS2	CFOD2	CFNS2	CFND2
CFOS1		-1.89	-7.89*	-5.10*	-2.12*	-1.78	-6.03*	-3.71*
CFOD1	-1.89		-4.74*	-2.82*	-0.25	0.08	-3.54*	-1.72
CFNS1	-7.89*	-4.74*		1.55	4.35*	4.82*	0.89	2.58*
CFND1	-5.10*	-2.82*	1.55		2.52*	2.90*	-0.63	0.99
CFOS2	-2.12*	-0.25	4.35*	2.52*		0.33	-3.22*	-1.45
CFOD2	-1.78	0.08	4.82*	2.90*	0.33		-3.61*	-1.79
CFNS2	-6.03*	-3.54*		-0.63	-3.22*	-3.61*		1.62
CFND2	-3.71*	-1.72	2.58*	0.99	-1.45	-1.79	1.62	
	CMOS1	CMOD1	CMNS1	CMND1	CMOS2	CMOD2	CMNS2	CMND2
CMOS1		-4.14*	-4.10*	-0.84	0.17	-3.04*	-5.08*	-3.81*
CMOD1	-4.14*		0.03	3.28*	4.35*	1.12	-0.76	0.39
CMNS1	-4.10*	0.03		3.24*	4.31*	1.08	-0.79	0.36
CMND1	-0.84	3.28*	3.24*		1.01	-2.18*	-4.16*	-2.93*
CMOS2	0.17	4.35*	4.31*	1.01		-3.25*	-5.31*	-4.02*
CMOD2	-3.04*	1.12	1.08	-2.18*	-3.25*		-1.92	-0.74
CMNS2	-5.08*	-0.76	-0.79	-4.16*	-5.31*	-1.92		1.17
CMND2	-3.81*	0.39	0.36	-2.93*	-4.02*	-0.74	1.17	

Note: Comparisons above the diagonal are for the "Low" performing component. Asterisks indicate significant differences at the $\alpha = 0.05$ level

APPENDIX I

One, Two, Three, and Four Component Model Parameter Estimates and Fit Statistics for the Summary-level Groups

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
PMO1	1	3	0.879 (0.006)								0.258	5.5+E5	389	23	1552	1554
	2	36	0.719 (0.010)	0.948 (0.003)			0.304 (0.029)	0.696 (0.029)			0.877	79.4	43.8	21	1176	1182
	3	136	0.618 (0.018)	0.850 (0.007)	0.971 (0.009)		0.123 (0.020)	0.398 (0.030)	0.479 (0.031)		0.990	10.2	10.8	19	1143	1153
	4	521	0.598 (0.019)	0.822 (0.009)	0.947 (0.004)	1.000 (0.001)	0.106 (0.019)	0.301 (0.029)	0.469 (0.031)	0.123 (0.020)	1.000	8.0	8.8	17	1141	1155
PFO1	1	3	0.801 (0.012)								0.321	2271	112.2	23	576	578
	2	41	0.697 (0.013)	0.924 (0.008)			0.543 (0.052)	0.457 (0.052)			0.911	29.8	12.6	21	476	482
	3	368	0.617 (0.023)	0.773 (0.013)	0.941 (0.008)		0.207 (0.042)	0.436 (0.051)	0.358 (0.050)		0.983	8.7	8.3	19	472	482
	4	741	0.489 (0.071)	0.671 (0.017)	0.808 (0.015)	0.945 (0.008)	0.022 (0.015)	0.328 (0.049)	0.324 (0.049)	0.326 (0.049)	0.991	7.7	8.1	17	472	486
PMN1	1	3	0.782 (0.007)								0.268	5631	433.9	23	1778	1780
	2	43	0.645 (0.009)	0.907 (0.005)			0.476 (0.031)	0.524 (0.031)			0.885	68.5	59.6	21	1404	1410
	3	191	0.577 (0.012)	0.805 (0.007)	0.955 (0.005)		0.275 (0.028)	0.458 (0.031)	0.267 (0.028)		0.985	23.2	23.1	19	1367	1377
	4	178	0.556 (0.014)	0.756 (0.009)	0.906 (0.006)	1.000 (0.001)	0.217 (0.026)	0.366 (0.030)	0.341 (0.030)	0.076 (0.016)	0.998	16.1	16.7	17	1361	1375
PFN1	1	3	0.676 (0.014)								0.383	692.6	87.0	23	576	578
	2	54	0.592 (0.013)	0.839 (0.013)			0.661 (0.049)	0.339 (0.049)			0.934	18.8	16.9	21	506	512
	3	190	0.564 (0.015)	0.757 (0.015)	0.930 (0.017)		0.513 (0.052)	0.386 (0.050)	0.101 (0.031)		1.000	10.4	11.1	19	500	510
	4	113	0.564 (0.015)	0.753 (0.015)	0.915 (0.019)	1.000 (0.001)	0.509 (0.052)	0.380 (0.050)	0.102 (0.031)	0.009 (0.010)	1.000	10.3	10.9	17	500	514
PMO2	1	3	0.874 (0.007)								0.148	1.4+E7	710.2	23	1531	1533
	2	10	0.473 (0.019)	0.938 (0.004)			0.139 (0.024)	0.861 (0.024)			0.949	74.9	60.4	21	882	888
	3	149	0.458 (0.020)	0.884 (0.008)	0.975 (0.003)		0.131 (0.024)	0.372 (0.034)	0.497 (0.035)		1.000	23.6	24.9	19	846	856
	4	1631	0.451 (0.020)	0.816 (0.020)	0.913 (0.006)	0.982 (0.003)	0.126 (0.023)	0.773 (0.019)	0.410 (0.035)	0.386 (0.034)	1.000	21.2	24.3	17	845	859

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
PFO2	1	3	0.820 (0.013)								0.189	4.2+E4	214.5	23	548	550
	2	22	0.621 (0.020)	0.936 (0.008)			0.367 (0.058)	0.633 (0.058)			0.870	44.0	36.8	21	370	376
	3	74	0.505 (0.029)	0.797 (0.017)	0.961 (0.007)		0.181 (0.046)	0.353 (0.057)	0.465 (0.060)		0.996	16.8	19.9	19	353	363
	4	76	0.500 (0.029)	0.765 (0.020)	0.926 (0.010)	1.000 (0.001)	0.174 (0.045)	0.265 (0.053)	0.411 (0.059)	0.150 (0.043)	1.000	13.7	16.4	17	350	364
PMN2	1	3	0.824 (0.008)								0.238	1.1+E4	398.1	23	1370	1372
	2	30	0.605 (0.013)	0.906 (0.005)			0.274 (0.032)	0.726 (0.032)			0.914	59.2	53.1	21	1025	1031
	3	116	0.557 (0.016)	0.835 (0.008)	0.958 (0.005)		0.198 (0.028)	0.445 (0.035)	0.357 (0.034)		1.000	21.0	20.7	19	992	1002
	4	721	0.557 (0.016)	0.835 (0.008)	0.956 (0.007)	0.960 (0.007)	0.198 (0.028)	0.445 (0.035)	0.166 (0.026)	0.190 (0.028)	1.000	21.0	20.7	17	992	1006
PFN2	1	3	0.758 (0.015)								0.285	414.9	114.1	23	476	478
	2	32	0.634 (0.016)	0.904 (0.011)			0.540 (0.060)	0.460 (0.060)			0.932	11.80	12.30	21	374	380
	3	148	0.587 (0.020)	0.769 (0.019)	0.926 (0.011)		0.354 (0.057)	0.303 (0.055)	0.343 (0.057)		1.000	8.0	9.0	19	371	381
	4	132	0.587 (0.020)	0.769 (0.019)	0.926 (0.012)	0.927 (0.026)	0.354 (0.057)	0.303 (0.055)	0.283 (0.054)	0.060 (0.028)	1.000	8.0	9.0	17	371	385
CMO1	1	3	0.870 (0.011)								0.281	8.3+E6	108.4	23	433	435
	2	65	0.738 (0.017)	0.945 (0.007)			0.363 (0.056)	0.637 (0.056)			0.847	1157	25.5	21	350	356
	3	49	0.338 (0.096)	0.775 (0.015)	0.956 (0.006)		0.013 (0.013)	0.426 (0.057)	0.560 (0.057)		0.955	15.5	15.0	19	340	350
	4	147	0.337 (0.096)	0.727 (0.021)	0.879 (0.013)	0.971 (0.006)	0.013 (0.013)	0.246 (0.050)	0.355 (0.055)	0.385 (0.056)	1.000	12.9	12.5	17	337	351
CFO1	1	3	0.829 (0.021)								0.288	1269	40.6	23	169	171
	2	35	0.678 (0.030)	0.913 (0.014)			0.357 (0.091)	0.643 (0.091)			0.879	12.0	10.0	21	139	145
	3	67	0.523 (0.064)	0.770 (0.026)	0.930 (0.014)		0.090 (0.054)	0.401 (0.093)	0.509 (0.094)		0.990	6.1	7.3	19	136	146
	4	154	0.523 (0.064)	0.770 (0.026)	0.930 (0.016)	0.930 (0.026)	0.090 (0.054)	0.401 (0.093)	0.361 (0.091)	0.148 (0.067)	0.990	6.1	7.3	17	136	150

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
CMN1	1	3	0.806 (0.013)								0.269	1277	127.6	23	502	504
	2	34	0.664 (0.017)	0.921 (0.009)			0.448 (0.057)	0.552 (0.057)			0.914	16.1	14.7	21	390	396
	3	213	0.634 (0.019)	0.870 (0.012)	0.982 (0.007)		0.355 (0.055)	0.466 (0.058)	0.179 (0.044)		0.980	6.0	7.5	19	382	392
	4	1023	0.627 (0.020)	0.833 (0.016)	0.926 (0.011)	0.999 (0.002)	0.329 (0.054)	0.294 (0.053)	0.292 (0.052)	0.085 (0.032)	0.987	5.6	7.3	17	382	396
CFN1	1	3	0.695 (0.025)								0.443	245.7	26.1	23	160	162
	2	59	0.623 (0.023)	0.839 (0.025)			0.667 (0.089)	0.333 (0.089)			0.939	14.1	12.8	21	147	153
	3	55	0.608 (0.025)	0.792 (0.025)	1.000 (0.000)		0.566 (0.094)	0.400 (0.093)	0.034 (0.034)		0.980	8.9	9.6	19	144	154
	4	78	0.608 (0.025)	0.792 (0.029)	0.792 (0.046)	1.000 (0.000)	0.566 (0.094)	0.286 (0.085)	0.114 (0.060)	0.034 (0.034)	0.980	8.9	9.6	17	144	158
CMO2	1	3	0.905 (0.009)								0.315	2.8+E5	90.8	23	398	400
	2	28	0.781 (0.017)	0.964 (0.005)			0.325 (0.052)	0.675 (0.052)			0.932	92.4	15.0	21	322	328
	3	31	0.482 (0.097)	0.804 (0.016)	0.966 (0.005)		0.014 (0.013)	0.337 (0.053)	0.649 (0.053)		1.000	10.1	10.5	19	318	328
	4	153	0.482 (0.097)	0.804 (0.016)	0.966 (0.008)	0.966 (0.007)	0.014 (0.013)	0.337 (0.053)	0.275 (0.050)	0.374 (0.054)	1.000	10.1	10.5	17	318	332
CFO2	1	3	0.843 (0.020)								0.285	421	42.1	23	161	163
	2	21	0.627 (0.040)	0.906 (0.013)			0.228 (0.081)	0.773 (0.081)			0.963	7.2	8.2	21	127	133
	3	68	0.615 (0.042)	0.886 (0.016)	0.974 (0.016)		0.210 (0.078)	0.635 (0.093)	0.154 (0.070)		1.000	6.1	7.1	19	126	136
	4	106	0.615 (0.042)	0.886 (0.026)	0.886 (0.020)	0.974 (0.016)	0.210 (0.078)	0.240 (0.082)	0.395 (0.094)	0.155 (0.070)	1.000	6.1	7.1	17	126	140
CMN2	1	3	0.822 (0.012)								0.225	7.8+E5	172.8	23	567	569
	2	22	0.645 (0.018)	0.929 (0.007)			0.377 (0.054)	0.622 (0.054)			0.900	98.0	17.1	21	411	417
	3	192	0.514 (0.035)	0.721 (0.018)	0.939 (0.007)		0.105 (0.034)	0.332 (0.053)	0.563 (0.055)		0.967	13.0	11.1	19	405	415
	4	381	0.360 (0.080)	0.649 (0.019)	0.884 (0.012)	0.969 (0.007)	0.019 (0.015)	0.326 (0.052)	0.366 (0.054)	0.289 (0.051)	0.983	8.9	8.6	17	403	417

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
CFN2	1	3	0.699 (0.025)								0.263	525.4	52.9	23	197	199
	2	28	0.604 (0.023)	0.915 (0.020)			0.694 (0.089)	0.306 (0.089)			0.855	14.4	8.4	21	152	158
	3	59	0.471 (0.043)	0.669 (0.026)	0.926 (0.020)		0.204 (0.077)	0.523 (0.096)	0.274 (0.086)		0.986	4.6	5.0	19	149	159
	4	78	0.471 (0.043)	0.669 (0.026)	0.926 (0.022)	0.926 (0.044)	0.204 (0.076)	0.523 (0.096)	0.219 (0.080)	0.054 (0.044)	0.986	4.6	5.0	17	149	163
MO1	1	3	0.876 (0.005)								0.264	4.4+E6	478.3	23	1955	1957
	2	40	0.723 (0.009)	0.947 (0.003)			0.315 (0.025)	0.685 (0.025)			0.873	213.6	50.6	21	1528	1534
	3	149	0.620 (0.015)	0.848 (0.006)	0.970 (0.003)		0.128 (0.018)	0.399 (0.027)	0.474 (0.027)		0.983	13.0	10.5	19	1488	1498
	4	969	0.483 (0.040)	0.677 (0.014)	0.866 (0.006)	0.973 (0.003)	0.020 (0.008)	0.149 (0.020)	0.399 (0.027)	0.433 (0.027)	0.997	6.9	8.2	17	1485	1499
FO1	1	3	0.807 (0.010)								0.314	2809	146.3	23	748	750
	2	40	0.693 (0.012)	0.92 (0.007)			0.499 (0.045)	0.502 (0.045)			0.912	27.3	15.1	21	617	623
	3	218	0.590 (0.024)	0.763 (0.012)	0.936 (0.007)		0.148 (0.032)	0.45 (0.045)	0.402 (0.045)		0.989	8.1	8.6	19	610	620
	4	1751	0.576 (0.027)	0.746 (0.012)	0.889 (0.014)	0.947 (0.008)	0.119 (0.029)	0.424 (0.045)	0.176 (0.035)	0.281 (0.041)	0.993	8.0	8.5	17	610	624
MN1	1	3	0.788 (0.006)								0.268	6739	552.8	23	2285	2287
	2	40	0.65 (0.008)	0.911 (0.004)			0.472 (0.027)	0.529 (0.027)			0.893	72.8	62.8	21	1795	1801
	3	285	0.594 (0.010)	0.824 (0.006)	0.963 (0.004)		0.304 (0.025)	0.456 (0.027)	0.240 (0.023)		0.981	21.0	20.7	19	1753	1763
	4	225	0.568 (0.012)	0.757 (0.008)	0.907 (0.005)	1.000 (0.007)	0.224 (0.023)	0.34 (0.026)	0.357 (0.026)	0.079 (0.015)	0.997	13.0	13.3	17	1746	1760
FN1	1	3	0.680 (0.012)								0.397	917	100.5	23	737	739
	2	56	0.599 (0.011)	0.839 (0.012)			0.662 (0.043)	0.338 (0.043)			0.939	21.4	17.2	21	654	660
	3	367	0.586 (0.012)	0.804 (0.012)	0.986 (0.012)		0.594 (0.045)	0.375 (0.044)	0.031 (0.016)		0.98	9.8	10.6	19	647	657
	4	252	0.573 (0.013)	0.751 (0.014)	0.875 (0.017)	1.000 (0.000)	0.512 (0.045)	0.345 (0.043)	0.124 (0.030)	0.019 (0.013)	0.998	8.8	9.7	17	646	660

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
MO2	1	3	0.883 (0.006)								0.169	2.8+E7	779.6	23	1942	1944
	2	20	0.519 (0.017)	0.936 (0.003)			0.127 (0.020)	0.873 (0.020)			0.893	133.7	99.3	21	1262	1268
	3	73	0.454 (0.020)	0.843 (0.008)	0.968 (0.003)		0.095 (0.018)	0.291 (0.027)	0.615 (0.029)		1.000	18.0	19.9	19	1183	1193
	4	1494	0.451 (0.020)	0.829 (0.009)	0.954 (0.003)	0.998 (0.002)	0.094 (0.017)	0.236 (0.025)	0.554 (0.030)	0.117 (0.019)	1.000	16.9	18.7	17	1181	1195
FO2	1	3	0.827 (0.011)								0.208	4.8+E4	240	23	710	712
	2	25	0.621 (0.018)	0.926 (0.007)			0.327 (0.048)	0.673 (0.048)			0.891	37.8	31.1	21	502	508
	3	202	0.547 (0.023)	0.837 (0.012)	0.967 (0.006)		0.204 (0.041)	0.424 (0.050)	0.373 (0.049)		0.982	13.1	14.9	19	485	495
	4	243	0.529 (0.025)	0.772 (0.018)	0.920 (0.008)	1.000 (0.001)	0.175 (0.039)	0.235 (0.043)	0.466 (0.051)	0.124 (0.033)	1.000	10.1	11.8	17	482	496
MN2	1	3	0.823 (0.007)								0.235	2.6+E5	555.1	23	1937	1939
	2	30	0.625 (0.011)	0.915 (0.004)			0.316 (0.028)	0.684 (0.028)			0.906	73.0	58.9	21	1441	1447
	3	123	0.566 (0.013)	0.827 (0.007)	0.956 (0.004)		0.208 (0.024)	0.402 (0.029)	0.390 (0.029)		1.000	18.8	19.1	19	1401	1411
	4	490	0.563 (0.013)	0.817 (0.008)	0.927 (0.007)	0.969 (0.005)	0.203 (0.024)	0.351 (0.029)	0.239 (0.025)	0.208 (0.024)	1.000	18.6	19.0	17	1401	1415
FN2	1	3	0.742 (0.013)								0.275	1053	166.7	23	681	683
	2	30	0.624 (0.013)	0.906 (0.009)			0.581 (0.050)	0.419 (0.050)			0.917	17.5	15.2	21	529	535
	3	161	0.557 (0.019)	0.728 (0.016)	0.923 (0.009)		0.309 (0.047)	0.349 (0.048)	0.342 (0.048)		0.998	8.1	9.4	19	523	533
	4	136	0.557 (0.019)	0.728 (0.016)	0.923 (0.010)	0.923 (0.022)	0.309 (0.047)	0.349 (0.048)	0.281 (0.046)	0.061 (0.024)	0.998	8.1	9.4	17	523	537
O1	1	3	0.858 (0.005)								0.268	5.3+E5	670.5	23	2783	2785
	2	38	0.713 (0.007)	0.942 (0.003)			0.367 (0.023)	0.633 (0.023)			0.888	152.5	59.5	21	2172	2178
	3	175	0.620 (0.012)	0.830 (0.006)	0.964 (0.003)		0.156 (0.017)	0.389 (0.023)	0.455 (0.023)		0.985	14.2	14.3	19	2127	2137
	4	1235	0.531 (0.023)	0.706 (0.010)	0.868 (0.005)	0.970 (0.003)	0.044 (0.010)	0.211 (0.019)	0.361 (0.023)	0.385 (0.023)	0.997	9.2	11.4	17	2124	2138

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
N1	1	3	0.759 (0.006)								0.278	5173	726.5	23	3152	3154
	2	42	0.634 (0.006)	0.902 (0.004)			0.534 (0.023)	0.467 (0.023)			0.901	87.5	73.1	21	2498	2504
	3	285	0.590 (0.008)	0.817 (0.006)	0.963 (0.004)		0.379 (0.023)	0.430 (0.023)	0.191 (0.018)		0.985	18.3	18.5	19	2444	2454
	4	188	0.571 (0.009)	0.758 (0.007)	0.906 (0.005)	1.000 (0.001)	0.303 (0.022)	0.346 (0.022)	0.288 (0.021)	0.063 (0.011)	1.000	9.5	10.2	17	2436	2450
O2	1	3	0.868 (0.005)								0.177	6.5+E6	1024	23	2698	2700
	2	28	0.575 (0.012)	0.935 (0.003)			0.186 (0.020)	0.814 (0.020)			0.879	147.7	116.1	21	1790	1796
	3	82	0.483 (0.015)	0.834 (0.007)	0.966 (0.003)		0.117 (0.017)	0.314 (0.024)	0.569 (0.026)		1.000	12.9	14.9	19	1689	1699
	4	731	0.477 (0.016)	0.807 (0.009)	0.941 (0.004)	0.994 (0.002)	0.113 (0.016)	0.221 (0.021)	0.489 (0.026)	0.177 (0.020)	1.000	8.8	11.0	17	1685	1699
N2	1	3	0.802 (0.006)								0.237	4.3+E4	756.5	23	2686	2688
	2	29	0.627 (0.008)	0.914 (0.004)			0.389 (0.025)	0.611 (0.025)			0.913	68.7	59.9	21	1990	1996
	3	128	0.568 (0.010)	0.806 (0.007)	0.948 (0.004)		0.251 (0.022)	0.354 (0.025)	0.395 (0.025)		1.000	15.4	17.0	19	1947	1957
	4	1111	0.565 (0.011)	0.795 (0.007)	0.937 (0.004)	1.000 (0.001)	0.243 (0.022)	0.324 (0.024)	0.401 (0.025)	0.033 (0.010)	1.000	14.5	16.2	17	1946	1960
P1	1	3	0.806 (0.004)								0.248	2.9+E4	1320	23	4919	4921
	2	38	0.663 (0.005)	0.929 (0.003)			0.462 (0.019)	0.538 (0.019)			0.885	161.9	123.4	21	3722	3728
	3	143	0.586 (0.008)	0.810 (0.005)	0.960 (0.003)		0.252 (0.016)	0.402 (0.019)	0.347 (0.018)		0.992	22.4	21.9	19	3620	3630
	4	241	0.574 (0.008)	0.779 (0.005)	0.931 (0.003)	1.000 (0.001)	0.218 (0.016)	0.348 (0.018)	0.352 (0.018)	0.082 (0.010)	1.000	13.0	13.0	17	3611	3625
P2	1	3	0.833 (0.005)								0.194	2.5+E5	1424	23	4052	4054
	2	26	0.594 (0.008)	0.925 (0.003)			0.277 (0.019)	0.723 (0.019)			0.899	152.5	132.3	21	2760	2766
	3	98	0.525 (0.010)	0.827 (0.006)	0.961 (0.002)		0.185 (0.017)	0.351 (0.021)	0.465 (0.021)		1.000	18.1	19.9	19	2648	2658
	4	556	0.519 (0.010)	0.805 (0.007)	0.940 (0.003)	1.000 (0.005)	0.177 (0.016)	0.276 (0.019)	0.455 (0.021)	0.091 (0.012)	1.000	13.3	15.1	17	2643	2657

Group	Comp's	Iter's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	VAF	χ^2	L^2	df	$-2LL$	AIC
C1	1	3	0.817 (0.008)								0.267	2.5+E4	333.2	0.023	1362	1364
	2	34	0.671 (0.010)	0.923 (0.005)			0.419 (0.034)	0.581 (0.034)			0.906	49.7	32.1	21	1061	1067
	3	197	0.631 (0.013)	0.850 (0.008)	0.969 (0.005)		0.301 (0.032)	0.420 (0.034)	0.279 (0.031)		0.979	14.9	11.9	19	1041	1051
	4	353	0.501 (0.033)	0.676 (0.012)	0.870 (0.008)	0.973 (0.005)	0.046 (0.015)	0.319 (0.032)	0.399 (0.034)	0.243 (0.030)	1.000	8.6	9.5	17	1039	1053
C2	1	3	0.840 (0.007)								0.225	1.7+E6	438.2	23	1475	1477
	2	28	0.664 (0.011)	0.938 (0.004)			0.358 (0.033)	0.643 (0.033)			0.885	118.6	35.0	21	1072	1078
	3	143	0.557 (0.018)	0.782 (0.010)	0.954 (0.004)		0.146 (0.024)	0.325 (0.032)	0.529 (0.034)		0.981	10.6	9.5	19	1046	1056
	4	860	0.507 (0.025)	0.705 (0.013)	0.876 (0.009)	0.965 (0.004)	0.081 (0.019)	0.245 (0.029)	0.278 (0.031)	0.397 (0.033)	0.994	6.6	7.3	17	1044	1058
P	1	3	0.818 (0.003)								0.222	1.4+E5	2730	23	9008	9010
	2	34	0.639 (0.005)	0.927 (0.002)			0.378 (0.014)	0.622 (0.014)			0.886	306.3	254	21	6532	6538
	3	110	0.559 (0.006)	0.813 (0.004)	0.960 (0.002)		0.215 (0.012)	0.380 (0.014)	0.406 (0.014)		0.996	28.2	30.2	19	6308	6318
	4	330	0.549 (0.007)	0.787 (0.004)	0.935 (0.002)	1.000 (0.001)	0.196 (0.011)	0.317 (0.013)	0.400 (0.014)	0.087 (0.008)	1.000	13.6	16.0	17	6294	6308
C	1	3	0.829 (0.005)								0.245	4.0+E5	767.6	23	2846	2848
	2	31	0.668 (0.008)	0.931 (0.003)			0.390 (0.024)	0.610 (0.024)			0.896	129.1	60.7	21	2139	2145
	3	250	0.606 (0.010)	0.824 (0.006)	0.960 (0.003)		0.230 (0.021)	0.365 (0.023)	0.405 (0.024)		0.977	21.6	17.0	19	2096	2106
	4	372	0.498 (0.021)	0.684 (0.009)	0.872 (0.006)	0.969 (0.003)	0.058 (0.011)	0.276 (0.022)	0.350 (0.023)	0.316 (0.023)	1.000	8.4	10.1	17	2089	2103
ALL	1	3	0.821 (0.003)								0.227	2.2+E5	3473	23	1.2+E4	1.2+E4
	2	33	0.647 (0.004)	0.928 (0.002)			0.381 (0.012)	0.619 (0.012)			0.888	349.2	290.2	21	8677	8683
	3	126	0.567 (0.005)	0.812 (0.003)	0.959 (0.002)		0.212 (0.010)	0.377 (0.012)	0.412 (0.012)		0.992	22.5	23.3	19	8410	8420
	4	1076	0.554 (0.006)	0.780 (0.004)	0.933 (0.002)	0.998 (0.001)	0.186 (0.010)	0.314 (0.011)	0.409 (0.012)	0.091 (0.007)	1.000	7.1	8.0	17	8395	8409

APPENDIX J

Parameter Estimates and Fit Statistics for the Three Component Restricted Models

Group	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
PMO1	3	0.618	0.850	0.971	0.123	0.398	0.479	0.990	10.2	10.8	19
		(0.018)	(0.007)	(0.009)	(0.020)	(0.030)	(0.031)				
		0.628	0.788	0.934	0.116	0.178	0.706	0.872	46.8	40.3	21
		(0.018)	(0.012)	(0.004)	(0.020)	(0.024)	(0.028)				
										29.5	2
PFO1	3	0.617	0.773	0.941	0.207	0.436	0.358	0.983	8.7	8.3	19
		(0.023)	(0.013)	(0.008)	(0.042)	(0.051)	(0.050)				
		0.628	0.769	0.934	0.224	0.393	0.383	0.959	10.1	8.6	21
		(0.022)	(0.014)	(0.009)	(0.043)	(0.051)	(0.050)				
										0.3	2
											*
PMN1	3	0.577	0.805	0.955	0.275	0.458	0.267	0.985	23.2	23.1	19
		(0.012)	(0.007)	(0.005)	(0.028)	(0.031)	(0.028)				
		0.628	0.808	0.934	0.355	0.307	0.338	0.855	38.9	35.3	21
		(0.010)	(0.009)	(0.005)	(0.030)	(0.029)	(0.029)				
										12.2	2
PFN1	3	0.564	0.757	0.930	0.513	0.386	0.101	1.000	10.4	11.1	19
		(0.015)	(0.015)	(0.017)	(0.052)	(0.050)	(0.031)				
		0.628	0.820	0.934	0.723	0.200	0.077	0.804	23.2	18.8	21
		(0.012)	(0.018)	(0.019)	(0.046)	(0.041)	(0.028)				
										7.7	2
PMO2	3	0.458	0.884	0.975	0.131	0.372	0.497	1.000	23.6	24.9	19
		(0.020)	(0.008)	(0.003)	(0.024)	(0.034)	(0.035)				
		0.628	0.835	0.934	0.165	0.004	0.830	0.536	256.4	116.3	21
		(0.017)	(0.081)	(0.004)	(0.026)	(0.005)	(0.027)				
										91.4	2
PFO2	3	0.505	0.797	0.961	0.181	0.353	0.465	0.996	16.8	19.9	19
		(0.029)	(0.017)	(0.007)	(0.046)	(0.057)	(0.060)				
		0.628	0.791	0.934	0.297	0.121	0.583	0.729	51.5	34.6	21
		(0.022)	(0.029)	(0.008)	(0.055)	(0.039)	(0.059)				
										14.7	2
PMN2	3	0.557	0.835	0.958	0.198	0.445	0.357	1.000	21.0	20.7	19
		(0.016)	(0.008)	(0.005)	(0.028)	(0.035)	(0.034)				
		0.628	0.831	0.934	0.251	0.272	0.478	0.808	50.5	38.0	21
		(0.014)	(0.010)	(0.005)	(0.031)	(0.031)	(0.035)				
										17.3	2
PFN2	3	0.587	0.769	0.926	0.354	0.303	0.343	1.000	8.0	9.0	19
		(0.020)	(0.019)	(0.011)	(0.057)	(0.055)	(0.057)				
		0.628	0.821	0.934	0.470	0.237	0.293	0.919	9.9	10.4	21
		(0.017)	(0.019)	(0.011)	(0.060)	(0.051)	(0.054)				
										1.4	2
											*

Group	Comp's	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	VAF	χ^2	L^2	df
CMO1	3	0.338	0.775	0.956	0.013	0.426	0.560	0.955	15.5	15.0	19
		(0.096)	(0.015)	(0.006)	(0.013)	(0.057)	(0.057)				
		0.628	0.765	0.934	0.067	0.274	0.659	0.835	95.3	22.9	21
		(0.044)	(0.019)	(0.007)	(0.029)	(0.051)	(0.055)				
										7.9	2
CFO1	3	0.523	0.770	0.930	0.090	0.401	0.509	0.990	6.1	7.3	19
		(0.064)	(0.026)	(0.014)	(0.054)	(0.093)	(0.094)				
		0.628	0.799	0.934	0.178	0.351	0.472	0.874	8.6	8.3	21
		(0.044)	(0.026)	(0.014)	(0.072)	(0.090)	(0.094)				
										1.0	2
											*
CMN1	3	0.634	0.870	0.982	0.355	0.466	0.179	0.980	6.0	7.5	19
		(0.019)	(0.012)	(0.007)	(0.055)	(0.058)	(0.044)				
		0.628	0.810	0.934	0.319	0.255	0.427	0.943	9.3	10.5	21
		(0.020)	(0.018)	(0.009)	(0.054)	(0.050)	(0.057)				
										3.0	2
											*
CFN1	3	0.608	0.792	1.000	0.566	0.400	0.034	0.980	8.9	9.6	19
		(0.025)	(0.025)	(0.000)	(0.094)	(0.093)	(0.034)				
		0.628	0.810	0.934	0.644	0.313	0.043	0.893	11.2	12.0	21
		(0.023)	(0.027)	(0.046)	(0.091)	(0.088)	(0.038)				
										2.4	2
											*
CMO2	3	0.482	0.804	0.966	0.014	0.337	0.649	1.000	10.1	10.5	19
		(0.097)	(0.016)	(0.005)	(0.013)	(0.053)	(0.053)				
		0.628	0.777	0.934	0.024	0.217	0.760	0.844	33.6	27.8	21
		(0.072)	(0.020)	(0.007)	(0.017)	(0.046)	(0.048)				
										17.3	2
CFO2	3	0.615	0.886	0.974	0.210	0.635	0.154	1.000	6.1	7.1	19
		(0.042)	(0.016)	(0.016)	(0.078)	(0.093)	(0.070)				
		0.628	0.873	0.934	0.217	0.396	0.387	0.951	6.9	7.6	21
		(0.041)	(0.021)	(0.016)	(0.079)	(0.094)	(0.094)				
										0.5	2
											*
CMN2	3	0.514	0.721	0.939	0.105	0.332	0.563	0.967	13.0	11.1	19
		(0.035)	(0.018)	(0.007)	(0.034)	(0.053)	(0.055)				
		0.628	0.789	0.934	0.291	0.152	0.558	0.862	70.4	14.1	21
		(0.020)	(0.024)	(0.008)	(0.051)	(0.040)	(0.056)				
										3.0	2
											*
CFN2	3	0.471	0.669	0.926	0.204	0.523	0.274	0.986	4.6	5.0	19
		(0.043)	(0.026)	(0.020)	(0.077)	(0.096)	(0.086)				
		0.628	0.895	0.934	0.710	0.083	0.207	0.772	26.0	9.4	21
		(0.023)	(0.042)	(0.021)	(0.087)	(0.053)	(0.078)				
										4.4	2
											*