Title

Application of Higher-order IRT models and Hierarchical IRT models to Computerized Adaptive Testing

Author

Lee, Moonsoo

Publication Date

2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORMIA

Los Angeles

Application of Higher-Order IRT Models and Hierarchical IRT Models to Computerized Adaptive Testing

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Moonsoo Lee

2014

ABSTRACT OF THE DISSERTATION

Application of Higher-Order IRT Models and Hierarchical IRT Models to Computerized Adaptive Testing

by

Moonsoo Lee

Doctor of Philosophy in Education

University of California, Los Angeles, 2014

Professor Li Cai, Chair

In recent years, the importance of formative assessments has been emphasized within educational measurement. This type of assessment often includes multiple correlated sub-domains and a hierarchical structure among the proficiencies. In this dissertation, several multidimensional CAT procedures are investigated to improve the measurement aspects of diagnostic testing and to better match the psychometric models to the test structure.

Five factors are manipulated with higher-order IRT models and hierarchical IRT models: (1) the different correlation conditions between two primary factors (low, medium, and high), (2) the number of group factors per primary factor (two and four), (3) the number of items (40, 80 and 160), (4) the item selection method (MFI and Bayesian), and (5) the proficiency score estimation method (MLE and EAP). Three outcome measures, including correlations between true and estimated proficiency scores, Root Mean Square Error (RMSE) of estimated proficiency scores, and Standard Errors (SE) are computed totaling 192 different conditions.

As expected, the correlation between true and estimated proficiency scores increase while RMSE and SE decrease when the test length correlation between two primary factors increase under different correlations among the factors, different item selection methods and different scoring methods. In overall, the higher-order IRT model CAT has an advantage over the hierarchical IRT model CAT when we need scores for the primary factors. On the other hand, if test designers are interested in more specific group factors, hierarchical IRT models outperformed the higher-order IRT models.

This study undertakes a comprehensive comparison of item selection methods and proficiency scores estimation in several multidimensional IRT models in conjunction with a CAT. The item selection and proficiency score estimation methods are negligible across the four multidimensional IRT CAT algorithms. However, the Bayesian item selection method has smaller RMSEs and SEs than the MFI method in specific cases and the EAP scoring method outperforms the MLE method, especially for short test length in this study.

The dissertation of Moonsoo Lee is approved.

Michael Seltzer

José-Felipe Martínez-Fernández

Steven Reise

Li Cai, Committee Chair

University of California, Los Angeles

2014

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work could not have been completed without the extensive support and guidance of my advisor and committee chair, Professor Li Cai. His patient encouragement greatly helped me to establish confidence in my ability to conduct research. I could never thank him enough and I deeply appreciate his being in my academic life.

I would like to express my gratitude to the members of my dissertation committee, Professor Michael Seltzer, Professor José-Felipe Martínez-Fernández, and Professor Steve Reise, for their thoughtful suggestions and comments. They are tremendous researchers and exceptional teachers.

A special thank to CTB/McGraw-Hill company, who recognized the value of my work, provided with monetary support and allowed me to access to the real testing data. I would especially like to thank the scientists at CTB, Dr. Wim van der Linden and Dr. Seung Choi for providing perspective comments on this work.

Many fellows have provided me with a great deal of support and assistance of these past few years. I am especially grateful to Ji Seung Yang, Mark Hansen, Larry Thomas, Scott Monroe, Megan Kuhfeld, Zhen Li, Carl Falk, Taehun Lee, and Nami Shin. Through every conversation and project that I had with them I was able to enhance my knowledge in the field of educational measurement and psychometrics.

Finally, I am tremendously grateful to my family - especially to my parents and sister. They have always inspired me with endless love while I was weak and vulnerable.

VITA

EDUCATION

2007	Bachelor of Arts, Education College of Sciences in Education Yonsei University, Seoul, Korea
2009	Master of Arts, Education (Educational measurement and evaluation) College of Sciences in Education Yonsei University, Seoul, Korea

WORK

2007-2008	Teaching Assistant and Research Assistant Department of Education Yonsei University, Seoul, Korea
2009-2010	Assistant Researcher Korea Institute for Curriculum and Evaluation (KICE) Seoul, Korea
2010-Present	Graduate Student Researcher National Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California, Los Angeles, Los Angeles, CA.
2013	Research Intern CTB/McGraw-Hill Monterey, CA.

PUBLICATIONS

Lee, M., Lee, G., & Kang, S. J. (2009). A simulation study for the comparison of IRT scale transformation and true score equating under the different conditions of mixed-format tests. *Korean Journal of Educational Evaluation. 22(3),* 805-826.

PRESENTATIONS

Lee, M., Hansen, M., & Cai, L. (2011, July). Calibration, scaling, DIF, and projection: A common framework using multidimensional IRT. *Paper presented at the 2011 annual meeting of the International Meeting of the Psychometric Society, Hong Kong, China.*

Lee, M. (2012, April). A comparison of Generalizability Theory and Many Facet Rasch Model approaches with focusing on interaction effects. *Paper presented at the 2012 meeting of the National Council of Measurement in Education, Vancouver, Canada.*

Lee, M., & Shin, N. (2013, April). Analysis of Differential Item Functioning of English language Learners and English only students using multi-level models. *Poster presented at the 2013 annual meeting of the American Educational Research Association, San Francisco, CA.*

Lee, M. (2014, April). CAT Using a Bifactor and Two-tier IRT Models. *Paper presented at the 2014 annual meeting of the National Council of Measurement in Education, Philadelphia, PA.*

Lee, M., Xiong, X., & Choi, S. (2014, April). Subscore in CAT Using Higher-order and Hierarchical IRT Models. *Paper presented at the 2014 annual meeting of the National Council of Measurement in Education, Philadelphia, PA.*

CHAPTER 1

Introduction

1.1 Background

Over the last ten years, summative assessments have increased in importance because of the emphasis of the States on student and school accountability. Summative assessments are used to evaluate students' learning, skill acquisition, and academic achievement at the conclusion of a defined instructional period—typically at the end of a project, unit, course, semester, program, or school year. Some of the most well-known and widely discussed examples of summative assessments are the standardized tests administered by States and testing organizations, usually in math, reading, writing, and science. As a vital component of educational systems and policies such as the No Child Left Behind Act of 2001 (NCLB), States or districts were required to assess student learning in relation to well-defined educational achievement standards. However, the purposes of summative assessments is confined to gauging student learning at particular points in time and to help evaluate the effectiveness of programs, school improvement, and student placement. Since educators realize that assessing student achievement according to the accountability requirements of NCLB is not sufficient, formative assessments or diagnostic tests will likely become more popular to monitor and improve student progress in the post-NCLB era. Formative assessments refer to a wide variety of methods that teachers use to conduct in-process

evaluations of students' comprehension, learning needs, and academic progress during a lesson, unit, or course (The Glossary of Education Reform, 2013). These types of assessments help teachers identify concepts that students are struggling to understand, skills they are having difficulty acquiring, or learning standards they have not yet achieved so that adjustments can be made to lessons, instructional techniques, and academic support. The general goal of formative assessments is to achieve an understanding of what students know and can do in order to help teachers make responsive adjustments in teaching and learning. While summative assessments for accountability provide a picture of a student's performance on a given day under standardized test conditions, formative assessments or diagnostic testing provide feedback to teachers, students, and principals or superintendents about specific elements of the measured content domain over the course of instruction through repeated test administrations.

In recent years, there has been an increase attention to improve formative assessments in educational measurement: (a) integrating summative and formative assessments such as Cognitively Based Assessment of, for, and as Learning (CBAL) (Bennett & Gitomer, 2009), (b) developing cognitively robust psychometric models in the context of diagnostic assessment such as Cognitive Diagnostic Assessment (CDA) (de la Torre & Douglas, 2004; de la Torre, 2009; DiBello et al., 2007; Junker & Sijtsma, 2001; Rupp et al., 2010) and (c) transition from paper-and-pencil testing to computer adaptive formative assessments such as Smarter Balanced Assessment (U.S. Department of Education, 2012).

Innovations such as Item Response Theory (IRT) and information technology have made it possible for Computerized Adaptive Testing (CAT) to be widely implemented in educational and psychological measurement over recent decades. The advantage of CAT over traditional paper-and pencil tests is the potential of either increased precision for a given length or a shorter test

for equal precision (Weiss, 1982; Wainer, 2000). Even for a CAT test of short length, the reliability of scores can be similar to that of longer fixed form tests because a CAT is more efficient in selecting those items that are expected to be most informative. Specifically, when items are targeted to the proficiency level of the examinee using IRT and a CAT algorithm, the standard error of measurement is minimized, and test length can be reduced without loss of precision. CAT also benefits from the same advantages of non-adaptive computer-based testing (CBT). For example, tests delivered by computer (whether adaptive or not) can easily utilize multimedia such as audio and video files and provide immediate feedback for examinees (Green, 1983).

As described above, because CAT has various advantages, its use has gradually become more common in the fields of education and psychology. Large-scale CAT implementations include several prominent cases (e.g., Graduate Record Examination [GRE], Graduate Management Admissions Test [GMAT], Armed Services Vocational Aptitude Battery [ASVAB], and Patient Reported Outcomes Measurement Information System [PROMIS]). Another system using CAT is the Smarter Balanced assessment. The Smarter Balanced assessment system capitalizes on the precision and efficiency of CAT for both the mandatory summative assessment and the optional interim assessments (U.S. Department of Education, 2012). The Smarter Balanced Assessment Consortium is one of two multistate consortia awarded funding from the U.S. Department of Education to develop an assessment system based on the new Common Core State Standards (CCSS). CCSS was developed voluntarily and cooperatively by more than 40 states for providing clear, consistent standards in English language arts/literacy and mathematics for grades 3-8 and high school students (U.S. Department of Education, 2012). This assessment system will be fully implemented in the spring of 2015. The core components of

Smarter Balanced are: (1) Summative assessments, (2) Interim assessments, and (3) Formative assessments. This assessment system provides online, tailored reports that link to instructional and professional development resources using CAT. In other words, it can provide students with information about progress toward proficiency and teachers with classroom results in key aspects of the CCSS.

In general formative assessments are often composed of several subtests that measure different proficiencies. These proficiencies are usually not independent of one another, but rather the knowledge/skill of one dimension could assist the students to correctly answer items from another dimension. Since most of the current computerized assessments are utilized by unidimensional IRT models, it is common to assume a test has these two properties - unidimensionality and local item independence. Unidimensionality means that a single ability or trait is measured by the set of items that make up the test (Hambleton & Swaminathan, 1991). Local independence means that conditional on an examinee's proficiency, the probability of correctly responding to an item is statistically independent of the probability of responding correctly to any other item (Hambleton & Swaminathan, 1991). However, as mentioned above, item responses within formative assessments are not entirely independent, and multidimensionality always exists to a lesser or greater extent in the item responses. Therefore, the use of formative assessments poses a challenge to standard IRT models because of the fundamental assumptions of unidimensionality and local independence.

In order to overcome this constraint, a number of approaches have been suggested such as bifactor modeling (Gibbons, & Hedeker, 1992, Gibbons et al., 2007) and higher-order modeling (Mulaik, & Quartetti, 1997, de la Torre, & Douglas, 2004). For example, bifactor modeling was applied to CAT that (1) measures more than one latent trait, (2) yields readily

interpretable latent traits, and (3) estimates directly item and person parameters jointly by Gibbons et al. (2008) and Gibbons et al. (2012). Some of these models will be discussed in subsequent chapters.

1.2 Statement of the Problem

Many studies have explored CAT-related issues over the last three decades. However, most of the studies on CAT have been conducted in the framework of Unidimensional Item Response Theory (UIRT). Because many educational and psychological constructs are multidimensional, a CAT system can utilize Multidimensional Item Response Theory (MIRT) models to measure constructs such as achievements and attitudes (Dodd, De Ayala, & Koch, 1995). In addition, MIRT has been adopted to use the correlation between dimensions to improve the measurement efficiency for individual subscales (Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007). The ability to "borrow strength" from other parts of the MIRT model leads to smaller standard errors of measurement for the scale scores (Cai, 2010). Here, multidimensional item response modeling in CAT is required not only as a last solution when violations of unidimensionality become problematic, but also as an appropriate choice when performance-based testing or diagnostic assessments are administered to measure complex skills in a real-world context.

A few studies have demonstrated the application of MIRT to CAT (Segall, 1996, Weiss & Gibbons, 2007, Gibbons et al., 2008). Segall (1996) proposed a Bayesian approach to the multidimensional CAT item selection process by incorporating prior knowledge of the joint distribution of proficiencies. Segall showed that for realistic item pools, multidimensional CAT can provide equal or higher precision with about one third fewer items than are required by

unidimensional CAT applied over multiple dimensions. In addition to Segall, Weiss and Gibbons (2007) and Gibbons et al. (2008) implemented a CAT with the bifactor model for mental health data. Weiss and Gibbons (2007) developed an algorithm that implements CAT for the bifactor model with dichotomously scored items. They evaluated and demonstrated the efficiency and precision of the performance of the algorithm in both post-hoc simulation data and real-testing data. Gibbons et al. (2008) also administered a CAT with the bifactor model to the Mood and Anxiety Spectrum Scales (MASS) data like Weiss and Gibbons' study (2007) in post-hoc simulation. Although Weiss and Gibbons (2007) and Gibbons et al. (2008) applied the bifactor model to CAT, their methods for item selection and ability / proficiency estimation were based on unidimensional scales. Specifically, the bifactor CAT algorithm still operated within a unidimensional system of item selection and ability or proficiency estimation for a general factor and group factors separately. Consequently, the bifactor CAT did not work out as expected because it did not consider cross-information gathered from items by both the general factor and group factors in implementing CAT.

In recent years, Seo (2011) investigated a full-information multidimensional bifactor CAT algorithm for item selection and scoring that was based on multidimensional IRT models. Seo's study demonstrated that the multidimensional bifactor CAT algorithm worked well when latent scores on the secondary dimensions were estimated properly. In his study, although a multidimensional bifactor CAT algorithm did not improve the accuracy and efficiency of the general factor scores compared to two unidimensional CAT algorithms, multidimensional bifactor CAT did show an improvement in the accuracy and efficiency of the specific factor scores. Huang et al. (2012) proposed a higher-order CAT algorithm in which latent traits have a hierarchical structure. Consistent with previous studies, their simulation results indicated that the

longer the test, the larger the item pool, and the larger the factor loadings, the better the measurement precision.

As described earlier, diagnostic assessments are often composed of several subtests that measure different proficiencies. These proficiencies are usually not independent of one another, but rather the knowledge or skill of one dimension could have an effect on a student's answers on the items of another dimension. In addition, many of these assessments have a hierarchical structure. For example, cognitive abilities can be classified into three strata (Carroll, 1993): Stratum I, which consists of narrow abilities; Stratum II, which consists of broad abilities; and Stratum III, which is Spearman's general ability. The latent traits measured by the Test of English as a Foreign Language (TOEFL) is also example of hierarchical structure. Reading, listening, speaking, and writing are the first-order latent traits of TOEFL, and language proficiency is the second-order latent trait (Sawaki, Sticker, & Andreas, 2009). Another example in educational test batteries is the Programme for International Student Assessment (PISA), in which multiple sub-domains (e.g., quantity, space and shape, change and relationship, and uncertainty) are measured in a subject (e.g., mathematics), and multiple subjects (e.g., mathematics, reading, and science) constitute a general concept of essential knowledge and skills. The sub-domains, subjects, and general concepts can be viewed as the first-, second-, and third-order latent traits, respectively (Huang et al., 2012). Therefore, MIRT model such as the higher-order model (de la Torre & Douglas, 2004), the bifactor model (Gibbons & Hedeker, 1992), or the two-tier model (Cai, 2010) may be appropriate to model the item-examinee interaction with consideration of both multidimensionality and hierarchical structure. For example, the first-order latent traits can be used as a formative assessment for diagnostic

purposes, and the second-order latent traits can provide overall performance for a summative assessment.

Research on formative assessments using the higher-order model and hierarchical model in CAT is still in its early stages. This research would be timely because there is a great premium on shortening tests in educational contexts, especially tests primarily intended for diagnostic purposes. Capitalizing on the correlational structure of abilities, precise and reliable estimates of the overall ability and domain-specific abilities are obtained simultaneously. Given the utility of CAT and the literature reviewed, this study will investigate the feasibility and effectiveness of the higher-order model and hierarchical model under a variety of conditions.

1.3 Purpose

The main purpose of this study is to investigate how diagnostic test designs could capitalize on the dimensional and hierarchical structure among the proficiencies being measured while using CAT to improve measurement precision and efficiency. To address the purpose of the study, simulation studies were conducted based on real data settings. The data generating item parameters were borrowed from the InView assessment (CTB/McGraw-Hill, 2002). InView measures verbal ability with two subtests: Verbal Reasoning-Words and Verbal Reasoning-Context, and nonverbal cognitive ability with three subtests (Sequences, Analogies, and Quantitative Reasoning). Therefore, the higher-order model with two primary factors and the two-tier model were fit to the data, along with a higher-order model with one primary factor and a bifactor model, to test the hypothesis of two primary ability factors (verbal and nonverbal)

underlying multiple sub-domains. The impact of the correlations of proficiency scores and item selection method on proficiencies were studied under a variety of conditions. The specific research questions are as follows:

1) Can the overall and domain abilities be accurately estimated using the higher-order and the hierarchical IRT models in CAT?
2) How well would the higher-order and the two-tier IRT model in CAT work when there are two primary factors?
3) How are all the above questions related to the CAT affected by correlations between the two primary dimensions, the factor types (primary or group factor), the number of group factors per primary factor, test length, item selection methods, and proficiency score estimation methods?

1.4 Significance of the Research

Along with simulation studies, a complete illustration is provided of how CAT procedures improved the potential benefits of a diagnostic assessment. Also, this dissertation addresses the importance of the correlations of estimated proficiencies with a different numbers of primary and specific dimensions and various test lengths. From the psychometric point of view, it validates various multidimensional IRT approaches in CAT to simultaneous estimation of students' overall and domain-specific proficiencies. Specifically, the higher-order and the two-tier IRT model could better capture the structure of multiple-component tests and provide an efficient estimation of hierarchical abilities. By fully utilizing information from each test administration,

“borrowing strength” from the correlational structure of latent traits, the precision and reliability of the targeted subscale estimates is expected to improve. In addition, higher-order and hierarchical IRT models with a CAT can enhance the validity and usefulness of a given test by providing diagnostic subscale estimates in addition to an overall scale estimate. These findings can be used as a guideline for researchers and practitioners when large numbers of examinees are not available to calibrate operational formative assessments.

CHAPTER 2

Multidimensional IRT CAT

2.1 Multidimensional IRT (MIRT)

Reckase (1985) proposed a multidimensional IRT model as an extension of the 3PL model. In his original formulation, a single item can measure two or more traits. For example, a mathematics word problem might measure both an examinee's mathematical and reading comprehension skills. Extending the 3PL model to a multidimensional context, Reckase (1997) formulated a linear logistic multidimensional model as

$$P\left(X_{ij}=1|\,\boldsymbol{\theta}_j,\boldsymbol{a}_i,d_i,c_i\right)=c_i+(1-c_i)\frac{\exp\left(\boldsymbol{a}_i'\boldsymbol{\theta}_j+d_i\right)}{1+\exp\left(\boldsymbol{a}_i'\boldsymbol{\theta}_j+d_i\right)}\ , \qquad (2.1)$$

where

$P\left(X_{ij}=1|\,\boldsymbol{\theta}_j,\boldsymbol{a}_i,d_i,c_i\right)$ is the probability of examinee *j* responding to item *i* correctly;

$\boldsymbol{\theta}_j$ is a vector of abilities for examinee *j*;

$\boldsymbol{a}_i$ is a vector of parameters related to the discriminating power of item *i* ;

d_i is a parameter related to the difficulty of item *i* (intercept);

c_i is the pseudo-guessing parameter of item *i*

In MIRT models, an overall discrimination index is defined as the multidimensional discrimination index (MDISC; Reckase, 1985); item difficulty, b_{ik} on dimension *k*, is defined with d_i.

$$MDISC_i = \left(\sum_{k=1}^{m} a_{ik}^2 \right)^{1/2} . \tag{2.2}$$

and

$$b_{ik} = \frac{-d_i}{a_{ik}} . \tag{2.3}$$

Ability estimation in MIRT models is relatively more challenging than in UIRT models. Therefore, MIRT models have been developed through the connection between the normal ogive model and item factor analysis to accommodate correlations among latent traits. When the classical linear factor model is applied to binary items, it is called item factor analysis (Bock, Gibbons, & Muraki, 1988). A basic estimation procedure for compensatory MIRT models involves full information factor analysis procedures performed on TESTFACT (Wilson, Wood, & Gibbons, 1991). More recently the flexMIRT® item response modeling software (Cai, 2013) has the capability to fit hierarchical IRT models (including item bifactor and two-tier item factor analysis models; Cai, 2010; Cai et al., 2011) using dimension reduction. Other useful alternatives for estimation of multidimensional compensatory models include the application of Markov chain Monte Carlo techniques (e.g., Béguin & Glas, 2001; Bolt & Lall, 2003).

2.2 Full-Information Item Factor Analysis

Bock, Gibbons, and Muraki (1988) introduced an IRT-based item factor analysis called "full-information item factor analysis (FIIFA)" that does not require calculation of inter-item correlation coefficients and is not strongly limited by the number of items. In addition full-information methods are no longer limited in applications by the number of factors and the total number of response patterns increasing exponentially (Bock, Gibbons, & Muraki, 1988).

In the FIIFA model, the conditional probability of an item score $u_{ij} = 1$, a correct response to item *i* by examinee *j* with trait vector $\boldsymbol{\theta}_j$, can be described as

$$P\left(u_{ij} = 1 \middle| \boldsymbol{\theta}_j, \boldsymbol{\lambda}_i, \tau_i\right) = \frac{1}{\sqrt{2\pi}\sigma_i} \int_{\tau_i}^{\infty} exp\left(-\frac{1}{2}\left(\frac{X_i - \boldsymbol{\lambda}_i \boldsymbol{\theta}_j}{\sigma_i}\right)^2\right) dX_i \,, \qquad (2.4)$$

where $\boldsymbol{\lambda}_i$ is factor loading vector of item *i* and τ_i is the threshold of item *i*. X_i is assumed to follow *N*(*0,1*), and the data are assumed to be sampled from a population of people whose θ follow a particular multivariate distribution. Estimates of factor loadings and thresholds are obtained from slope and intercept values estimated in the framework of IRT modeling. Takane and De Leeuw (1987) showed the formal equivalence of the marginal likelihood of the multidimensional two-parameter normal ogive model and the marginal likelihood of the item factor analysis for dichotomous variables, the parameters of the MIRT model can be linearly transformed to those of the FIIFA model. The parameters of MIRT $\boldsymbol{a}_i$ and d_i can be expressed

by the parameters of the FIIFA model, $\boldsymbol{\lambda}_i, \tau_i, \sigma_i$ and vice-versa. Given factor loadings and thresholds, the item slope and intercept of the *k*th dimension can be obtained by

$$a_{ik} = \frac{\lambda_{ik}}{\sigma_i}\,, \qquad d_i = -\frac{\tau_i}{\sigma_i}\,, \tag{2.5}$$

where $\sigma_i = \sqrt{1 - \boldsymbol{\lambda}_i \Phi \boldsymbol{\lambda}_i'}$. Factor loadings and the thresholds of *k* dimensions can be transformed from item slopes and intercepts as

$$\lambda_{ik} = \frac{a_{ik}}{\sqrt{1 + \boldsymbol{a}_i \Phi \boldsymbol{a}_i'}}\,, \qquad \tau_i = \frac{-d_i}{\sqrt{1 + \boldsymbol{a}_i \Phi \boldsymbol{a}_i'}}\,, \tag{2.6}$$

2.3 Higher-Order IRT Model

As described in previous sections, the higher-order factor analysis has been widely used in the behavioral sciences because many latent traits have a hierarchical structure (Matin & Adkins, 1954). In the higher-order IRT model, a test is viewed as consisting of several unidimensional subtest domains. That is, a single domain-specific ability $\theta_j^{(k)}$ accounts for examinee *j* 's performance on domain *k*, where *k* = 1,2, ... , K. The correlations between different domain abilities can be accounted for by positing a higher-order ability θ_j that is viewed as the examinee's overall ability. Specifically, the domain abilities are expressed as linear functions of the overall ability,

$$\theta_j^{(k)} = \lambda^{(k)} \theta_j + \varepsilon_{jk}\,, \tag{2.7}$$

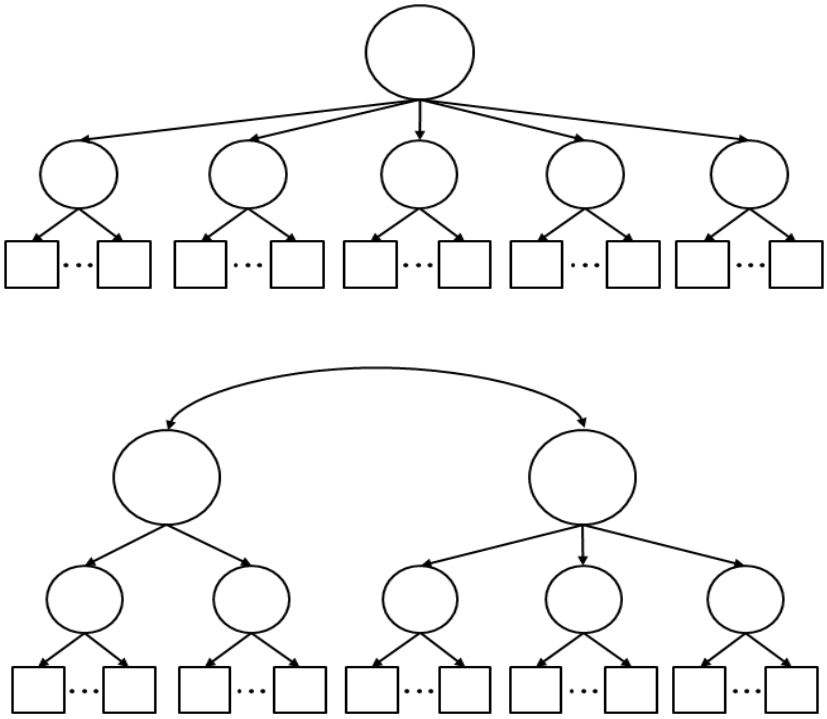
where ε_{jk} is assumed to be normally distributed with a mean of zero and independent of other εs and θs, and $\lambda^{(k)}$ is a measure of association (correlation) between the second-order latent trait and the kth first-order latent trait (Huang et al., 2012). In the first order, an item response function is imposed, which in theory can be any kind of model. For example, the three parameter logistic model is defined as

$$P_{j|k} = c_k + (1 - c_k)\frac{\exp\left[a_k\left(\theta_j^{(k)} - b_k\right)\right]}{1 + \exp\left[a_k\left(\theta_j^{(k)} - b_k\right)\right]}, \tag{2.8}$$

where $P_{j|k}$ is the probability of scoring 1 in domain k for person j, a_k is the discrimination parameter, b_k is the difficulty parameter, and c_k is the pseudo-guessing parameter, of domain k. Combining Equations (2.7) and (2.8) leads to the three parameter higher-order IRT:

$$P_{j|k} = c_k + (1 - c_k)\frac{\exp\left[a_k\left(\lambda^{(k)}\theta_j - b_k + \varepsilon_{jk}\right)\right]}{1 + \exp\left[a_k\left(\lambda^{(k)}\theta_j - b_k + \varepsilon_{jk}\right)\right]}. \tag{2.9}$$

The diagrammatic representation of the HO-IRT model is given in Figure 2.1.



2.1: Factor structures for higher-order models with one and two primary factors

2.4 Two-Tier IRT Model

The two-tier item factor analysis model (Cai, 2010) is a restricted confirmatory item factor model with special features: items may load on any number of primary dimensions but at most one specific dimension, the primary dimensions may be correlated, but specific dimensions are uncorrelated with one another and with the primary dimensions. This model generalizes and unifies correlated-traits multidimensional IRT models (Reckase, 2009), bifactor IRT models (Gibbons & Hedeker, 1992), and testlet response models (Wainer et al., 2007) in a single

modeling framework (Cai, 2010). As an example, Equation 2.10 shows a two-tier factor pattern for 500 items:

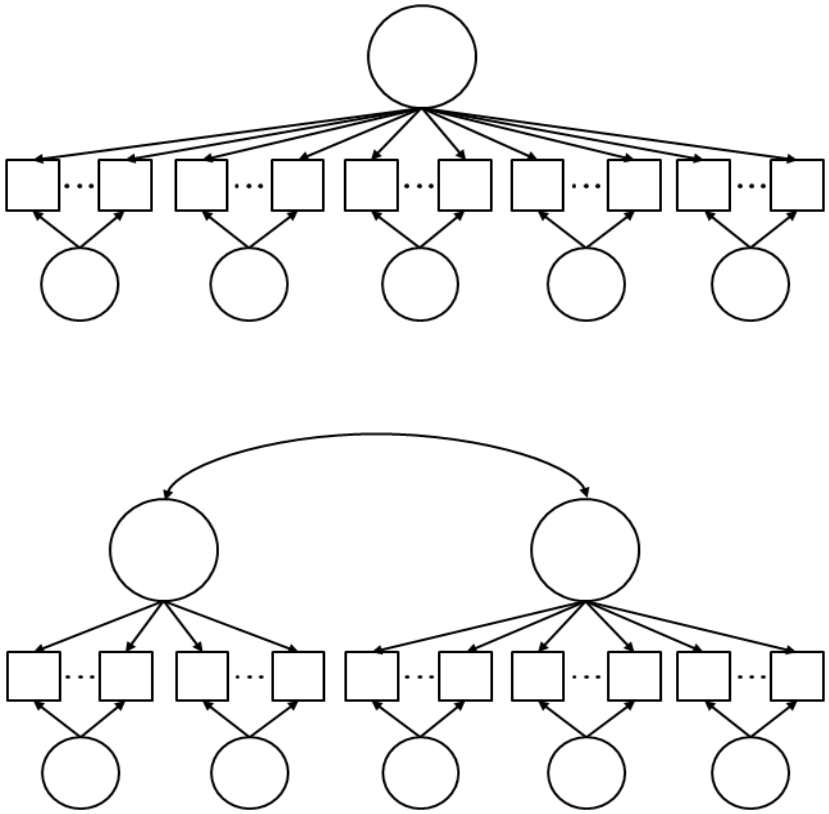
$$\left(\begin{array}{cc|ccccc} \alpha_{1\,1} & & \alpha_{1\,3} & & & & \\ \vdots & & \vdots & & & & \\ & & \alpha_{100\,3} & & & & \\ & & & \alpha_{101\,4} & & & \\ & & & \vdots & & & \\ \alpha_{200\,1} & & & \alpha_{200\,4} & & & \\ & \alpha_{201\,2} & & & \alpha_{201\,5} & & \\ & & & & \vdots & & \\ & & & & \alpha_{300\,5} & & \\ & \vdots & & & & \alpha_{301\,6} & \\ & \vdots & & & & \vdots & \\ & \vdots & & & & \alpha_{400\,6} & \\ & & & & & & \alpha_{401\,7} \\ & & & & & & \vdots \\ & \alpha_{500\,2} & & & & & \alpha_{500\,7} \end{array}\right) \tag{2.10}$$

where the α's denote nonzero item slopes. The first two columns of the matrix presents the primary dimensions and the rest five columns are the specific dimensions. Each item has a vector of slope parameters, corresponding to each primary and specific factors. Each item is allowed to load on the general factor and a single specific factor. The two-tier model also imposes a specific kind of factor covariance structure:

$$\begin{pmatrix} \boldsymbol{\Sigma} & \\ \mathbf{0} & \text{diag}(\boldsymbol{\tau}) \end{pmatrix}. \tag{2.11}$$

where $\boldsymbol{\Sigma}$ can be of any type and diag($\boldsymbol{\tau}$) is diagonal, with $\boldsymbol{\tau}$ as the diagonal elements. The two primary factors have unit variances, and are correlated. As a result, the ability to "borrow strength" from other parts of the model to enhance statistical prediction is an essential benefit of the two-tier model over separate bifactor analyses that would ignore the correlations among the

primary factors (Cai, 2010). Figure 3.2 presents the path diagram for the bifactor model and the two-tier model.



2.2: Hierarchical factor structure for the bifactor model and two-tier model.

The three parameter model can be extended to cover the two-tier case. Denote the primary latent variables for respondent j as a p×1 vector $\boldsymbol{\eta}_j = \left(\eta_{j1},\ \ldots\ ,\eta_{jp}\right)'$, and the specific latent variables as an S×1 vector $\boldsymbol{\xi}_j = \left(\xi_{j1},\ \ldots\ ,\xi_{jS}\right)'$.

$$P_1(\boldsymbol{\eta}, \xi_s, \boldsymbol{\theta}) = c(\boldsymbol{\theta}) + \frac{1-c(\boldsymbol{\theta})}{1+(\exp\{-[d(\boldsymbol{\theta})+[\boldsymbol{\alpha}(\boldsymbol{\theta})]'\boldsymbol{\eta}+\alpha_s(\boldsymbol{\theta})\xi_s]\})} \quad . \qquad (2.12)$$

where d is the intercept, $\boldsymbol{\alpha}$ is the $p\times1$ vector of item slopes on the primary factor, and αs is the item slope on specific factor s. The conditional probability for the incorrect or nonendorsement response is

$$P_0(\eta, \xi_s, \theta) = 1 - P_1(\eta, \xi_s, \theta)\,. \tag{2.13}$$

The higher-order and bifactor or two-tier model are alternative approaches for representing primary constructs comprised of several highly related domains. Yung et al.'s (1999) demonstration that the second-order models are nested within corresponding bifactor models made it possible to directly compare the two models. Chen et al. (2006) compared the results of the second-order model and the bifactor model using a health-related quality of life data. Their study indicated that the bifactor model had several advantages over the higher-order model when researchers are interested in both the primary latent variable and the specific latent variables. Specifically, the bifactor model fit the data significantly better than the second-order model and allowed for easier interpretation of the relationship between the domain specific factors and external variables, over and above the general factor. However, the bifactor models may not be preferable to higher-order models under all conditions. If the primary factor is the main focus of the study, the higher-order model may be more parsimonious, given that the higher-order model fits the data equally well as the bifactor model. Moreover, the bifactor and second-order representations are not mutually exclusive, and they can coexist in different parts of the same complex model (Chen et al., 2006)

2.5 Computerized Adaptive Testing (CAT)

Advancements in IRT over the past several decades have opened ways of powerful data analysis in the field of educational and psychological measurement, such as differential item functioning and test score linking/equating, and CAT. CAT is a process of test administration in which test items are selected for administration on the basis of the examinee's responses to previously administered items (Weiss & Kingsbury, 1984). Characteristics of CAT include a pre-calibrated item bank, use of differential entry item, a procedure of item selection, a proficiency or ability estimation method, and a stopping rule for terminating the test (Reckase, 1989).

Pre-calibrated item bank

The implementation of CAT requires developing a large bank of test items. For a CAT, each examinee gets an individualized test consisting of varying sets of items drawn from the item bank. Thus, the quality of the item bank has a significant effect on the performance of the adaptive algorithm in a CAT. A bank might contain thousands of items, and all items are assumed to measure identical latent traits on the same scale. It is usually necessary to link subsets of items administered to different groups onto a target matrix using a reference group to create a large item bank (e.g., Hambleton, Swaminathan, & Rogers, 1991; Kolen & Brennan, 2004). Various IRT models offer pre-calibrated item parameters and a reasonable method for linking subsets of test items, owing to the invariance properties of parameters in items and examinees.

Entry item

An entry item should be determined before implementing CAT. Usually an entry item in CAT is assigned based on a θ of 0 (this is typically the assumed population mean) because it is difficult to obtain valid prior information about the θ level of examinees. In practice, since CAT begins with an item of median difficulty level, the item would be readily overexposed. Therefore, several possible methods are proposed to reduce the item exposure rate. One possible method is to use random selection (combining IRT and Bayesian statistical methods) of the first few items from a subset of the item bank (e.g., Baker, 1992; Weiss & McBride, 1984).

Item selection rule

One of the key factors characterizing CAT is the item selection rule, which is essential to continue the adaptive testing process after an entry item is given to examinees. Two commonly used item selection algorithms are: maximum information and Bayesian selection (Thissen & Mislevy, 2000). The maximum information (MI) selection procedure chooses, at each step of the CAT, an item from the pool that provides the maximum amount of item information, $I(\theta)$, given the provisional estimate of the examinee's ability, θ (Lord, 1977). By maximizing the incremental information provided with each item, the MI procedure is also maximizing the expected precision of θ and doing so with substantially less items than traditional non-adaptive tests. A Bayesian counterpart to the MI procedure is known as the maximum posterior precision selection procedure (Owen, 1969, 1975). This procedure at each step chooses the item that is expected to maximize the precision of the posterior ability distribution. This procedure overcomes the issue of large errors in the provisional ability estimates, especially at the beginning of a CAT, by selecting items based on the entire posterior ability distribution instead of a single point estimate. Thus, while the selected item may not provide maximum information

at the provisional ability estimate, it is the most informative on average across the high density region of the posterior distribution (Parshall et al., 2002). The disadvantages of this procedure, however, is that it can be far more computationally intensive than MI and that the ability estimate is sensitive to the order in which items are administered (Thissen & Mislevy, 2000).

Scoring procedure

In most CAT systems, the parameter values for items in the item bank are assumed to have been pre-tested and calibrated before the items are administered operationally. Thus, the only parameter that requires estimation during CAT administration is the examinee's proficiency or ability level, θ. The first step in ability estimation process involves determining an initial ability estimate. One way to determine the initial ability estimate is to use prior information known about the examinee, such as the examinee's previous test scores in the same subject area. Or, it can simply be set to the mean of the assumed distribution, which would be zero, if θ is assumed to be from the standard normal distribution. After each item is given, interim or provisional estimates of θ are typically needed by the CAT algorithm to choose the next item. The final ability estimation is then performed at the end of the test based on the examinee's entire set of responses (Thissen & Mislevy, 2000). The provisional and final ability estimates do not have to be obtained using the same method (Chang, Ansley & Lin, 2000). The final ability estimate may also be transformed to a different ability metric (Parshall et al., 2002). Two common approaches to ability estimation in CAT are maximum likelihood estimation and Bayesian estimation. A likelihood function, $L(\theta)$, describes the probability of observing the set of item responses. If item parameters are assumed known, examinees' θ level can be estimated from the likelihood function, which is the product of all item response functions. Usually, the local maximum value of the likelihood function given a θ value can be obtained by setting the first derivative of the

natural log of the likelihood function at zero. However, maximum likelihood methods can be used only when there is a mixed response pattern (not all 0 or 1 responses). On the other hand, Bayesian methods can be used for any response pattern because they are based on Bayes' rule that is proportional to the product of the likelihood and prior probability. Usually, a prior probability distribution of θ assumes a standard normal distribution. In Bayesian estimation methods, the Bayesian modal estimator is to find the maximum value of a posterior distribution of θ (MAP). The expected a posteriori (EAP) method is to find the mean of the posterior distribution of θ (Owen, 1975).

Stopping rule

Every CAT needs a stopping rule that determines when the item administration should terminate. CAT stopping rules fall generally into two categories resulting in two types of adaptive tests: fixed-length test and variable-length tests. Fixed-length CATs administer items until a predetermined number of items have been given. Thus, each examinee receives the same number of items on the test. Fixed length CATs have the advantage of being easier to implement and better prediction of item pool usage rates (Thissen & Mislevy, 2000). As such fixed-length CATs are very popular and have been implemented for CATs in a variety of assessments. Examples include the CAT version of the GRE (Mills, 1999) and the CAT version of the ASVAB (Segall & Moreno, 1999). On the other hand, a variable-length CAT tests each examinee until a pre-specified level of measurement precision is reached. The criterion for stopping can be a target standard error (SE) of measurement for MI selection or a target posterior precision under Bayesian selection (Thissen & Mislevy, 2000). The main advantage of variable-length CATs is that every examinee is measured with approximately the same degree of precision. Examinees with ability well targeted by the items in the pool generally receive shorter tests than those with

ability levels in the extremes (Parshall et al., 2002). Examples of variable-length CATs include several national licensure and certification tests such as the National Certification Examination (NCE) for registered nurse anesthetists and the National Council Licensure Examination of Registered Nurses (Bergstrom & Lunz, 1999).

2.6 Multidimensional CAT (MCAT)

If a multidimensional test data set is assumed to be unidimensional, the invariant feature of IRT models may be jeopardized because of model misfit (Ackerman, 1991; Reckase, 1985). Because of the necessity of much larger numbers of items required to show a satisfactory fit to the response model, van der Linden (2008) expected the problem of multidimensionality to be more influential for adaptive testing. Segall (2010) pointed out that "when the dimensions measured by a test or battery are correlated, responses to items measuring one dimension provide clues about the examinee's standing along other dimensions" (p. 57). Such a unique feature might make MCAT more appealing, by increasing the accuracy for an examinee's proficiency estimates (e.g., Luecht, 1996; Segall, 1996). The change from unidimensional to multidimensional adaptive testing involves an important adjustment to the item selection and scoring procedures. Since the item information functions are substituted by item information matrices, the presence of more than one person parameter to be estimated during the test complicates the item selection process extensively, which not only reflects the accuracy of the estimates but also their correlations.

The process of extending unidimensional CAT to MCAT methods has been explored (e.g., Bloxom & Vale, 1987; Luecht, 1996; Segall, 1996). Bloxom and Vale (1987) proposed an approximate scoring procedure to item selection based on a multivariate extension of Owen's

(1975) sequential updating procedure. Segall (1996) proposed a Bayesian approach to the MCAT item selection process by incorporating the prior knowledge of the joint distribution of proficiency. Segall showed that for realistic item pools, MCAT can provide equal or higher precision with about one third fewer items than required by unidimensional CAT applied over dimensions. Luecht (1996) extended Segall's approach by imposing a more complex set of content-balancing constraints within a licensing/certification context. Results from both studies indicated that a shorter of test could achieve a similar subscore reliability as its longer unidimensional counterpart.

2.6.1 MCAT Item Selection

Maximum Fisher Information method

MCAT can select an item that provides Maximum Fisher Information (MFI) at current $\hat{\theta}$ as described in Equation 2.14. The Fisher information matrix is a convenient measure of the information in the observable response variables on the vector of proficiency parameters θ (Mulder & van der Linden, 2009).In MCAT, the provisional trait estimate vector, $\hat{\theta}^{(n)}$, obtained after responding to the *n*th item, is used to evaluate the item information function (Lord, 1980):

$$I(\theta, u_i) = \frac{\left[\frac{\partial P_i(\hat{\theta}^{(n)})}{\partial \theta}\right]^2}{P_i(\hat{\theta}^{(n)})Q_i(\hat{\theta}^{(n)})} \ , \qquad (2.14)$$

where u_i is the candidate item response among items in the item bank, $P_i(\hat{\theta}^{(n)})$ is the item response function with candidate item *i* at $\hat{\theta}^{(n)}$, and $Q_i(\hat{\theta}^{(n)}) = 1 - P_i(\hat{\theta}^{(n)})$. The test

information matrix of a set of *S* items is equal to the sum of the information matrices for the single items,

$$\boldsymbol{I}_s(\boldsymbol{\theta}) = \sum_{i \in S} \boldsymbol{I}_i(\boldsymbol{\theta}) \ . \tag{2.15}$$

When evaluating the selection of the n^{th} item in the CAT, the amount of information can be expressed as the sum of the test information matrices for the *n-1*items already administered and the matrix for candidate item i_n.

$$I_{S_{n-1}}\left(\hat{\theta}^{(n-1)}\right) + I_{i_n}\left(\hat{\theta}^{(n-1)}\right) . \tag{2.16}$$

Bayesian method

In a Bayesian approach, the posterior distribution of θ is updated after each observed response. Owen (1969, 1975) was the first to use a Bayesian approach to adaptive testing. His method had the format of a sequential Bayes procedure in which at each stage the previous posterior distribution of the unknown parameter served as its new prior distribution. According to Bayes' theorem, the posterior density function of θ is described as

$$f(\boldsymbol{\theta}|\mathbf{u}) = \frac{L(\mathbf{u}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{u})} \ , \tag{2.17}$$

where $L(\mathbf{u}|\boldsymbol{\theta})$ is the likelihood function; $f(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, MVN($\mathbf{0}$,$\boldsymbol{\Phi}$); and $f(\mathbf{u})$ is the marginal probability of $\mathbf{u}$. Usual choices of point estimates of θ are the mean and the mode of its posterior distribution known as the expected (EAP) and maximum a posteriori (MAP) estimates, respectively. The former requires numerical integration, (e.g. the Gauss–Hermite

formulas from Glas (1992)). The latter can be determined using a Newton–Raphson procedure, for instance, Segall (1996). For example, the Maximum a posteriori (MAP) estimates of $\boldsymbol{\theta}$ can be approximated by setting the partial derivative of the log of the posterior distribution in Equation 2.17 at zero. The Bayesian item selection method adjusts the maximum-likelihood (ML) item selection method like the Bayesian θ estimation method for selecting candidate item i by maximizing the determinant of the posterior information matrix as

$$\left|\mathbf{I}_{i|S_{n-1}}\left(\widehat{\boldsymbol{\theta}}^{(\mathrm{n})}, u_i\right) + \boldsymbol{\Phi}^{-1}\right|, \tag{2.18}$$

where $\boldsymbol{\Phi}^{-1}$ is the inverse of the covariance matrix of the prior distribution of vector $\boldsymbol{\theta}$.

2.6.2 MCAT Proficiency Estimation

ML proficiency estimation was first applied in MCAT by Segall (1996, 2010). The estimator is defined as the maximizer of the likelihood function over the range of possible θ values:

$$\hat{\theta}^{ML}_{u_{i_1} \ldots u_{i_{k-1}}} \equiv \arg max_{\theta} \left\{L\left(\theta | u_{i_1} \ldots u_{i_{k-1}}\right) : \theta \in (-\infty, \infty)\right\}. \tag{2.19}$$

and

$$L\left(\theta | u_{i_1}, \ldots, u_{i_{k-1}}\right) = \prod_{i=1}^{n-1} P(\theta|u_i)^{u_i} \, Q(\theta|u_i)^{1-u_i}, \tag{2.20}$$

where $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$ is the proficiency estimator after the responses to the first k-1 items, and $L\left(\theta|u_{i_1},\ldots,u_{i_{k-1}}\right)$ is the likelihood of response U_i . The ML estimates are the solution to set of m simultaneous equations given by :

$$\frac{\partial}{\partial \boldsymbol{\theta}} ln\, L(\boldsymbol{u}|\boldsymbol{\theta}) = \mathbf{0}\,, \tag{2.21}$$

Segall (1996, 2000) suggested using an iterative numerical procedure, (e.g. Newton-Raphson procedure), to obtain the estimates. A more detailed description of the method can be found in Segall (1996, 2000).

In a Bayesian approach, a point estimator of θ can be based on its posterior distribution in (2.22). A prior for the unknown value of the ability parameter g(*θ)*, is assumed. Together, the likelihood and prior yield the posterior distribution of *θ*:

$$\mathrm{g}\left(\theta|u_{i_1},\ldots,u_{i_{k-1}}\right) = \frac{L\left(\theta|u_{i_1},\ldots,u_{i_{k-1}}\right)\mathrm{g}(\theta)}{\int L\left(\theta|u_{i_1},\ldots,u_{i_{k-1}}\right)\mathrm{g}(\theta)\,d\theta}\,. \tag{2.22}$$

Typically, this density is assumed to be uniform or, if the examinees can be taken to be exchangeable, to be an empirical estimate of the ability distribution in the population of examinees. The population distribution is often modeled to be normal (van der Linden & Pashley, 2010). Posterior-based estimators used in adaptive testing are the Bayes modal (BM) or maximum a posteriori (MAP) estimator and the expected a posteriori (EAP) estimator. The former is defined as the maximizer of the posterior of *θ*,

$$\hat{\theta}^{MAP}_{u_{i_1} \ldots u_{i_{k-1}}} \equiv \arg max_{\theta} \left\{ g\left(\theta | u_{i_1} \ldots u_{i_{k-1}}\right) : \theta \in (-\infty, \infty) \right\}; \qquad (2.23)$$

the latter as its expected value:

$$\hat{\theta}^{EAP}_{u_{i_1} \ldots u_{i_{k-1}}} \equiv \int \theta \, g\left(\theta | u_{i_1} \ldots u_{i_{k-1}}\right) d\theta \, . \qquad (2.24)$$

The MAP estimator was introduced in IRT by Lord (1986) and Mislevy (1986). Use of the EAP estimator in adaptive testing is discussed extensively in Bock and Mislevy (1988). For a uniform prior, the posterior distribution in (2.22) becomes proportional to the likelihood function over the support of the prior. On the other hand, for nonuniform prior distributions, the small-sample properties of the MAP estimator depend not only on the likelihood but also on the shape of the prior distribution. Depending on the choice of prior distribution, the posterior distribution may be multimodal. For a proper prior distribution, the EAP estimator always exists and it is easy to calculate. No iterative procedures are required; one round of numerical integration generally suffices (van der Linden & Pashley, 2010).

CHAPTER 3

Methodology

3.1 Experimental Design

The purpose of this study is to investigate the accuracy and precision of the higher-order and the hierarchical IRT model CAT algorithms. Six facets that reflect realistic testing situations and that could affect the precision of CAT were considered: (1) CAT algorithms (IRT models: bifactor, two-tier, higher-order with one primary factor, and higher-order with two primary factors), (2) correlation conditions between two primary factors (low, medium, and high), (3) the number of group factors per primary factor (two and four), (4) test length (40, 80 and 160), (5) item selection method (MFI and Bayesian), and (6) θ estimation methods (MLE and EAP). A total of 192 conditions were simulated to gather a comprehensive understanding of higher-order and hierarchical IRT models in CAT.

The comparison was based on three criteria, including the correlation between true θ (θ) and estimated θ ($\hat{\theta}$), Root Mean Square Error (RMSE), and Standard Error (SE). In order to minimize the sample variance and increase the power to detect the effects of interest, 10 replications were used to compare the differences between θ and $\hat{\theta}$.

Table 3.1: All 192 simulated conditions

MIRT models	Correlation b/t primary factors	# of group factors per primary factor	Test length (# of items)	Item selection methods	Proficiency estimation methods
Bifactor	N/A	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
Two-tier	0.1	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
	0.4	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
	0.7	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP

Continued, next page.

Table 3.1: (Continued)

MIRT models	Correlation b/t primary factors	# of group factors per primary factor	Test length (# of items)	Item selection methods	Proficiency estimation methods
Higher-order (One primary factor)	N/A	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
Higher-order (Two primary factors)	0.1	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
	0.4	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
	0.7	2	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP
		4	40	MFI/ Bayesian	MLE/EAP
			80	MFI/ Bayesian	MLE/EAP
			160	MFI/ Bayesian	MLE/EAP

3.2 Data Generation

In this study, an item bank was simulated based on real data from the InView assessment (CTB/McGraw-Hill, 2002). InView provides cognitive ability and anticipated achievement information of students in grade 2 through 12. This real data set included 6481 examinees (Grade 8) and 100 multiple choice items (Level 4). The item parameters from InView were estimated using both a higher-order and hierarchical IRT model. From the confirmatory factor analysis results of CTB/McGraw-Hill (2002), InView measures verbal ability with two subtests: Verbal Reasoning-Words and Verbal Reasoning-Context, and nonverbal cognitive ability with three subtests: Sequences, Analogies, and Quantitative Reasoning. In Table 3.2, all correlations among the domains were about 0.5. Among the five subsets, Verbal Reasoning-Words correlates more highly with Verbal Reasoning-Context than with the rest of the three subsets.

Table 3.2: InView domain correlations

	VW	VC	SQ	AN	QR	V	NV
Verbal Reasoning-Words (VW)	1						
Verbal Reasoning-Context (VC)	0.63	1					
Sequences (SQ)	0.51	0.52	1				
Analogies (AN)	0.54	0.51	0.60	1			
Quantitative Reasoning (QR)	0.46	0.51	0.57	0.48	1		
Verbal (V)	0.89	0.89	0.58	0.59	0.54	1	
Nonverbal (NV)	0.68	0.61	0.88	0.84	0.80	0.68	1
InView Total	0.79	0.80	0.82	0.80	0.75	0.89	0.94

For the Monte-Carlo simulation study, IRT parameters were specified that could be transformed into factor analytic parameters. The equation for the probability of a correct response for a 2PL MIRT model is

$$P\left(X_{ij}=1|\,\boldsymbol{\theta}_j,\boldsymbol{a}_i,d_i\right)=\frac{\exp\left(\boldsymbol{a}_i'\boldsymbol{\theta}_j+d_i\right)}{1+\exp\left(\boldsymbol{a}_i'\boldsymbol{\theta}_j+d_i\right)}\ , \tag{3.1}$$

where

$P\left(X_{ij}=1|\,\boldsymbol{\theta}_j,\boldsymbol{a}_i,d_i\right)$ is the probability of examinee *j* responding to item *i* correctly;

$\boldsymbol{\theta}_j$ is a vector of the general and group factor latent traits of examinee *j*;

$\boldsymbol{a}_i$ is a vector of discrimination parameters of item *i*;

d_i is a parameter related to the difficulty of item *i* (intercept).

The item responses for this study were generated given the true θs and item parameters using Equation 3.1. In this study, each item bank contained 500 dichotomous items with true item parameters. Total of 1,000 examinees were generated of response matrices according to each of four models using the FlexMIRT program (Cai, 2013). For this study it was assumed that an examinee was sampled from a population with a known multivariate distribution. In the population of examinees, on each dimension, proficiency was normally distributed (M=0.0, SD=1.0). For each item bank, the item difficulty parameters were randomly generated from a uniform distribution from -3 to 3. The discrimination parameters for two general dimensions were generated from N(1, 0.2) and each of specific dimension was generated from N(0.5, 0.2) based on InView's item parameters. The correlations between two primary factors were 0.1, 0.4, and 0.7. The correlation of 0.1 was chosen to represent a low correlation and 0.7 was chosen to represent high correlation between two primary dimensions.

3.3 MCAT Scoring and Item Selection Methods

In MCAT, proficiency estimation and item selection are conducted for all dimensions simultaneously. As a consequence, the MCAT might administer an unequal number of items from each of the group factor scales, which would result in θ estimates based on different mixtures of group factor scales. Thus, the MCAT algorithm alternated items that loaded on each group factor, which functioned as content balancing in the MCAT (Weiss & Gibbons, 2007). In the example of this study for a bifactor model with two group factors, the MCAT was terminated after 40 items total were administered, with 20 items selected from the item bank measuring the first group factor scale and 20 items selected from the item bank measuring the second group factor scale in the case of a short length test. In the two-tier model with four group factors, the MCAT was terminated after 40 items total were administered with 5 items loading from each group factor scale in the case of a short length test. For selecting the initial item, θ is fixed at 0, which was the midpoint of the scale for all dimensions.

Multidimensional θ estimation

Multidimensional MLE and EAP methods were used for estimation of θ for each examinee in the CAT. As described in the previous sections, the Newton-Raphson procedure was used to obtain the MLE of θ. The Newton-Raphson procedure approximated the maximum of the likelihood by using an iterative procedure. The Newton-Raphson iterations were repeated until the incremental change in $\hat{\theta}$ became less than the criterion of .001. Early in the CAT procedure, it was necessary for MLE to specify an alternative for all correct response patterns or all incorrect response patterns that did not result in likelihood with a maximum; $\hat{\theta}$ was decreased or increased by 1 for

each incorrect response and for each correct response, until $\hat{\theta}$ reached 4 in absolute value. This procedure was employed until the response pattern became mixed (van der Linden, 1999). The EAP approach can obtain finite θ estimates for non-mixed response patterns because it used a standard normal distribution as the prior. In the EAP method, since integration of the distribution for an assumed general factor and group factor scores was not in a closed form, 15 quadrature points were used from –3 to 3 on the standard normal distribution for both the general factor and group factors.

Multidimensional item selection

Two possible item selection methods were considered. Item selection proceeded by computing and maximizing the determinant of either (a) the Fisher information matrix evaluated at the vector of current proficiency estimates, or (b) the posterior covariance matrix of the proficiencies (Segall, 1996). Both procedures are known as the criteria of D-optimality (see Silvey, 1980). The first method is maximizing the determinant of the Fisher information matrix, and the second method is generating the largest decrement in the volume of the Bayesian credibility ellipsoid for the estimates of the dimension scores of each examinee. For the second method, Segall (1996) developed a Bayesian version of the D-optimality criterion that evaluates the determinant of the posterior covariance matrix at the posterior modes of the proficiencies. Assuming the prior distribution for the proficiency estimates was multivariate normal with variance-covariance matrix, $\boldsymbol{\Phi}$, he showed that the volume of the Bayesian credibility ellipsoid for the estimates of proficiency is related to the following expression:

$$\arg\max_{i_n \in R_n} \det\left(\mathbf{I}_{s_{n-1}}\left(\widetilde{\boldsymbol{\theta}}_{n-1}\right) + \mathbf{I}_{i_n}\left(\widetilde{\boldsymbol{\theta}}_{n-1}\right) + \boldsymbol{\Phi}^{-1}\right), \tag{3.2}$$

where $\widetilde{\boldsymbol{\theta}}_{n-1}$ is the posterior mode after n–1 items have been administered. By applying the Bayesian principles to MCAT, both item-selection and scoring algorithms can be specified to enhance the precision of the adaptive test scores (Segall, 2010). The primary difference between these two approaches is that the Bayesian-based item selection method uses the posterior distribution instead of the maximum likelihood function. The criterion for the maximum Fisher information item selection and the criterion for the Bayesian item selection differ only by the term which consists of the inverse of the covariance matrix of the prior distribution of proficiencies, $\boldsymbol{\Phi^{-1}}$.

3.4 MCAT Procedure

A program in R (R Core Development Team, 2008) was utilized to implement the MACT algorithm. Figure 3 shows the flow chart of the MCAT algorithm. The higher-order IRT model CAT and hierarchical IRT model CAT algorithms implemented the following steps:

Step 1: Generate true item parameters and θ based on pre-specified distributions.

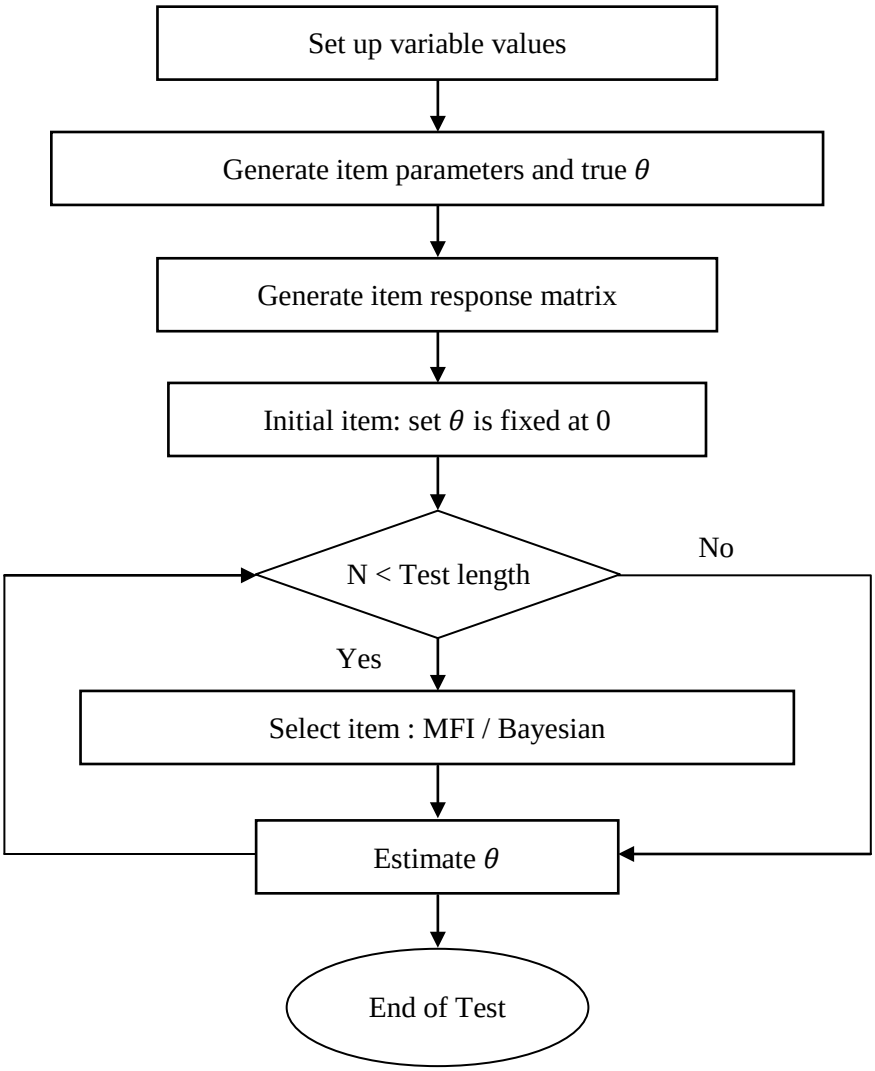
Step 2: Generate item responses based on the true item parameters and θ parameters.
Using the bifactor, the two-tier model and true parameters, the probability of each response for an item was calculated. Random numbers from U[0, 1] were generated and compared to the probabilities of responding at each score response.

Step 3: Simulate CAT.
The MCAT algorithm proceeded for a general factor and group factor scales simultaneously. For the first item, $\hat{\theta}$ for each examinee was set to a vector of zeros. The

examinees' $\hat{\theta}$ on all factors were estimated at the same time. This step is an iterative process.

Step 4: Obtain final estimated $\hat{\theta}$s.

After all "fixed length" items were administered, dependent variables (Correlation, RMSE, and OSE) were calculated and saved.



3.1: Flow chart for MCAT algorithm

Reprinted from "*Application of the Bifactor Model to Computerized Adaptive Testing*," by D. Seo, 2011, Unpublished doctoral dissertation, University of Minnesota .

3.5 Composite score for hierarchical IRT model

Scoring sub-domains using MIRT models are often designed such that each item measures the primary trait and one additional secondary trait. The secondary traits may reflect different content categories in the test, or different tests within a battery of tests. In the bifactor and the two-tier IRT model, all items are specified to load on the primary factor. Additionally, each item may load on one additional factor and the factors are orthogonal (Gibbons & Hedeker, 1992). In a hierarchical model, the specific factors represent variation above and beyond the primary factor (i.e., the variation that cannot be explained by the primary factor). On the other hand, in a higher-order model, the first-order factors represent both that part of the observed variance that can be explained by the higher-order factors, as well as the specific variance that cannot be explained by the higher-order factor.

With hierarchical IRT models, scores can be estimated for the primary trait and each secondary trait. On a battery of tests, though, it would seem desirable for each subtest score to be a measure of the overall construct covered in the subtest, not just the part of the construct not covered by the primary factor. In other words, the score should be a combination of the primary trait and the secondary trait, not just the secondary trait. To quantify the relative weights of the factors contributing to an item response, Reckase (1985; 1997; Reckase & McKinley,1991) defined the direction of greatest slope for item *i* as

$$\alpha_{ik} = arccos \frac{a_{ik}}{\sum_{k=1}^{m} a_{ik}^2}, \tag{3.3}$$

where α_{ik} is the angle with axis *k* and a_{ik} is the discrimination parameter for trait *k*. In the hierarchical IRT models, it is simplest to view the angle for each item relative to the primary factor. If an item measured only the primary trait, a would be 0; if an item measured the primary trait and secondary trait equally, a would be 45º. In this study, the average direction cosines were around $\cos \alpha_{i1} = .326$ and $\cos \alpha_{i2} = .875$ corresponding to angles of 71º and 29º with the θ_1 and θ_2 axes, respectively. Based on these angles, the sub-domain ability was calculated as a weighted linear composite.

3.5 Evaluation

The goal of this dissertation was to compare measurement accuracy and precision of the different multidimensional IRT models across several manipulated test conditions. Measurement accuracy and precision were assessed by the degree to which each test design recovered the known examinee θ values. This included computing and comparing the Root Mean Squared Error (RMSE) and Standard Error (SE) of the final θ estimates, and the correlation between the true θ and estimated θ values for each replication and grand means calculated across the 10 replications.

In this research the θ estimates obtained from administering a fixed number of items were used for evaluation of the performance of the CAT. The correlation was computed as a Pearson product-moment correlation, $r(\theta_j , \hat{\theta}_j)$. The RMSE of examinee θ estimates was calculated by computing the square root of the mean squared difference between the examinee true and

estimated θ for a given dimension. These RMSE statistics were averaged over 1000 examinees at the dimension level and reported. The RMSE and SE statistics for each examinee on each factor and across replications were computed by the following formulas:

$$RMSE(\theta_{jk}) = \sqrt{\frac{1}{\mathrm{N}}\sum_{j=1}^{\mathrm{N}}(\hat{\theta}_{jk} - \theta_{jk})^2}\ , \tag{3.4}$$

and

$$SE(\widehat{\boldsymbol{\theta}}_{jk}) = \frac{1}{\sqrt{I(\theta_{jk})}}\ , \tag{3.5}$$

where *j* is an examinee, *k* is each factor, and *N* is the number of examinees.

CHAPTER 4

Results

The results for the CAT designs of different multidimensional IRT models are presented in this chapter. The four designs, (1) Bifactor IRT model, (2) Higher-order IRT model, (3) Two-tier IRT model, and (4) Higher-order IRT model with two primary factors, were compared on measurement accuracy and precision; these findings are presented first. Each multidimensional IRT CAT design was simulated across five manipulated test conditions – different correlation conditions between two primary factors (low, medium, and high), number of group factors per primary factor (two and four), test length (40, 80 and 160 items), item selection method (MFI and Bayesian), and θ estimation methods (MLE and EAP). To save space, only Tables and Figures that list notable results from particular conditions are included in this chapter. Tables and Figures for conditions omitted from this chapter are given in Appendixes A, B and C. All results were averaged across 10 replications in each study condition.

4.1 Correlation between True and Estimated Proficiency Scores

Tables 4.1 - 4.6 and Figures 4.1 - 4.6 show average correlations between true and estimated θ*s* for each of 10 replications of the multidimensional IRT models. The high correlation between

Table 4.1: Correlation between True and Estimated Proficiency Scores (Two group factors, 40 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.848	N/A	0.841	0.843	N/A	
		EAP	0.858		0.842	0.853		
	Bayes	MLE	0.847		0.839	0.836		
		EAP	0.859		0.847	0.850		
Higher-order								
	MFI	MLE	0.861	N/A	0.833	0.835	N/A	
		EAP	0.871		0.836	0.842		
	Bayes	MLE	0.866		0.832	0.846		
		EAP	0.878		0.839	0.846		
Two-tier								
	MFI	MLE	0.802	0.807	0.875	0.879	0.896	0.898
		EAP	0.816	0.818	0.849	0.847	0.857	0.887
	Bayes	MLE	0.808	0.815	0.878	0.882	0.897	0.901
		EAP	0.834	0.824	0.859	0.853	0.879	0.891
Higher-order (2 general factors)								
	MFI	MLE	0.850	0.822	0.827	0.884	0.836	0.897
		EAP	0.853	0.838	0.829	0.859	0.832	0.880
	Bayes	MLE	0.845	0.829	0.834	0.889	0.835	0.904
		EAP	0.856	0.830	0.829	0.861	0.828	0.877

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.2: Correlation between True and Estimated Proficiency Scores (Two group factors, 80 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.871	N/A	0.833	0.836	N/A	
		EAP	0.884		0.852	0.862		
	Bayes	MLE	0.877		0.842	0.831		
		EAP	0.884		0.857	0.867		
Higher-order								
	MFI	MLE	0.888	N/A	0.865	0.863	N/A	
		EAP	0.886		0.865	0.866		
	Bayes	MLE	0.882		0.871	0.866		
		EAP	0.891		0.869	0.867		
Two-tier								
	MFI	MLE	0.868	0.868	0.839	0.850	0.867	0.848
		EAP	0.869	0.866	0.841	0.843	0.852	0.846
	Bayes	MLE	0.876	0.860	0.854	0.859	0.865	0.866
		EAP	0.879	0.876	0.836	0.840	0.850	0.858
Higher-order (2 general factors)								
	MFI	MLE	0.877	0.909	0.815	0.870	0.815	0.860
		EAP	0.895	0.908	0.831	0.865	0.837	0.861
	Bayes	MLE	0.890	0.907	0.819	0.875	0.813	0.874
		EAP	0.899	0.908	0.835	0.858	0.841	0.864

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.3: Correlation between True and Estimated Proficiency Scores (Two group factors, 160 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.896	N/A	0.864	0.867	N/A	
		EAP	0.904		0.868	0.875		
	Bayes	MLE	0.899		0.859	0.855		
		EAP	0.907		0.871	0.877		
Higher-order								
	MFI	MLE	0.915	N/A	0.874	0.868	N/A	
		EAP	0.911		0.874	0.877		
	Bayes	MLE	0.916		0.881	0.870		
		EAP	0.926		0.877	0.870		
Two-tier								
	MFI	MLE	0.893	0.887	0.828	0.855	0.824	0.848
		EAP	0.904	0.905	0.856	0.852	0.856	0.847
	Bayes	MLE	0.903	0.911	0.840	0.849	0.846	0.847
		EAP	0.911	0.910	0.837	0.850	0.851	0.842
Higher-order (2 general factors)								
	MFI	MLE	0.923	0.940	0.855	0.853	0.848	0.889
		EAP	0.928	0.942	0.857	0.884	0.859	0.878
	Bayes	MLE	0.923	0.942	0.850	0.863	0.857	0.887
		EAP	0.932	0.939	0.869	0.872	0.864	0.886

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.4: Correlation between True and Estimated Proficiency Scores (Four group factors, 40 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.844	N/A	0.852	0.849	0.857	0.855	N/A			
		EAP	0.861		0.858	0.856	0.858	0.857				
	Bayes	MLE	0.852		0.854	0.847	0.859	0.850				
		EAP	0.863		0.858	0.850	0.858	0.852				
Higher-order												
	MFI	MLE	0.875	N/A	0.835	0.842	0.846	0.834	N/A			
		EAP	0.878		0.840	0.852	0.851	0.841				
	Bayes	MLE	0.877		0.845	0.849	0.862	0.838				
		EAP	0.870		0.858	0.850	0.853	0.845				
Two-tier												
	MFI	MLE	0.791	0.780	0.904	0.900	0.909	0.902	0.898	0.908	0.918	0.905
		EAP	0.799	0.803	0.910	0.876	0.915	0.846	0.842	0.890	0.927	0.844
	Bayes	MLE	0.799	0.809	0.904	0.906	0.910	0.896	0.895	0.909	0.915	0.905
		EAP	0.804	0.811	0.899	0.905	0.907	0.864	0.895	0.905	0.921	0.848
Higher-order (2 general factors)												
	MFI	MLE	0.802	0.791	0.852	0.848	0.857	0.830	0.846	0.856	0.866	0.853
		EAP	0.801	0.814	0.858	0.824	0.853	0.814	0.790	0.838	0.875	0.802
	Bayes	MLE	0.810	0.820	0.854	0.856	0.850	0.826	0.845	0.859	0.863	0.813
		EAP	0.815	0.821	0.847	0.853	0.855	0.842	0.843	0.853	0.869	0.796

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.5: Correlation between True and Estimated Proficiency Scores (Four group factors, 80 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.858	N/A	0.858	0.860	0.851	0.860	N/A			
		EAP	0.867		0.864	0.850	0.844	0.866				
	Bayes	MLE	0.857		0.863	0.854	0.851	0.862				
		EAP	0.871		0.864	0.865	0.844	0.864				
Higher-order												
	MFI	MLE	0.876	N/A	0.845	0.841	0.856	0.842	N/A			
		EAP	0.888		0.840	0.852	0.851	0.841				
	Bayes	MLE	0.879		0.844	0.831	0.854	0.837				
		EAP	0.880		0.858	0.850	0.853	0.845				
Two-tier												
	MFI	MLE	0.819	0.831	0.892	0.882	0.937	0.924	0.906	0.921	0.941	0.919
		EAP	0.835	0.829	0.904	0.858	0.934	0.860	0.881	0.910	0.947	0.864
	Bayes	MLE	0.831	0.837	0.902	0.895	0.933	0.923	0.908	0.926	0.937	0.918
		EAP	0.844	0.843	0.901	0.905	0.916	0.882	0.905	0.925	0.930	0.868
Higher-order (2 general factors)												
	MFI	MLE	0.828	0.842	0.840	0.830	0.875	0.862	0.854	0.869	0.889	0.867
		EAP	0.846	0.840	0.852	0.806	0.879	0.840	0.831	0.860	0.897	0.812
	Bayes	MLE	0.842	0.838	0.850	0.843	0.881	0.851	0.856	0.874	0.885	0.866
		EAP	0.855	0.854	0.849	0.853	0.864	0.830	0.862	0.882	0.887	0.825

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.6: Correlation between True and Estimated Proficiency Scores (Four group factors, 160 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.886	N/A	0.859	0.858	0.850	0.851	N/A			
		EAP	0.906		0.859	0.861	0.858	0.864				
	Bayes	MLE	0.907		0.858	0.861	0.854	0.853				
		EAP	0.916		0.859	0.865	0.862	0.864				
Higher-order												
	MFI	MLE	0.920	N/A	0.840	0.846	0.850	0.855	N/A			
		EAP	0.927		0.873	0.865	0.868	0.856				
	Bayes	MLE	0.926		0.850	0.854	0.863	0.859				
		EAP	0.923		0.879	0.872	0.870	0.859				
Two-tier												
	MFI	MLE	0.885	0.891	0.904	0.897	0.898	0.875	0.909	0.913	0.929	0.907
		EAP	0.896	0.881	0.896	0.903	0.905	0.879	0.885	0.894	0.933	0.853
	Bayes	MLE	0.884	0.876	0.904	0.891	0.890	0.885	0.905	0.911	0.937	0.909
		EAP	0.896	0.880	0.901	0.904	0.911	0.901	0.917	0.912	0.937	0.881
Higher-order (2 general factors)												
	MFI	MLE	0.896	0.902	0.852	0.845	0.866	0.823	0.857	0.861	0.877	0.855
		EAP	0.907	0.892	0.844	0.851	0.873	0.827	0.833	0.842	0.881	0.801
	Bayes	MLE	0.895	0.887	0.854	0.841	0.910	0.835	0.855	0.861	0.885	0.857
		EAP	0.907	0.891	0.849	0.852	0.889	0.849	0.865	0.860	0.885	0.829

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

the two primary factors ($\rho = 0.7$) is represented in this chapter because it closely resembled the real testing parameters (InView: $\rho = 0.89$, two or three group factors per each primary factor). Tables and figures for low and medium correlations between two primary factors are given in Appendixes A.

As shown in Tables 4.1 - 4.6, the correlation between true and estimated proficiency scores increased when the test length increased for different correlations among the factors, both item selection methods and, both scoring methods. The higher-order IRT model algorithms provided slightly higher correlations than the bifactor and the two-tier IRT models in primary factors. On the other hand, the hierarchical IRT models showed higher average correlation than the higher-order IRT models in group factors.

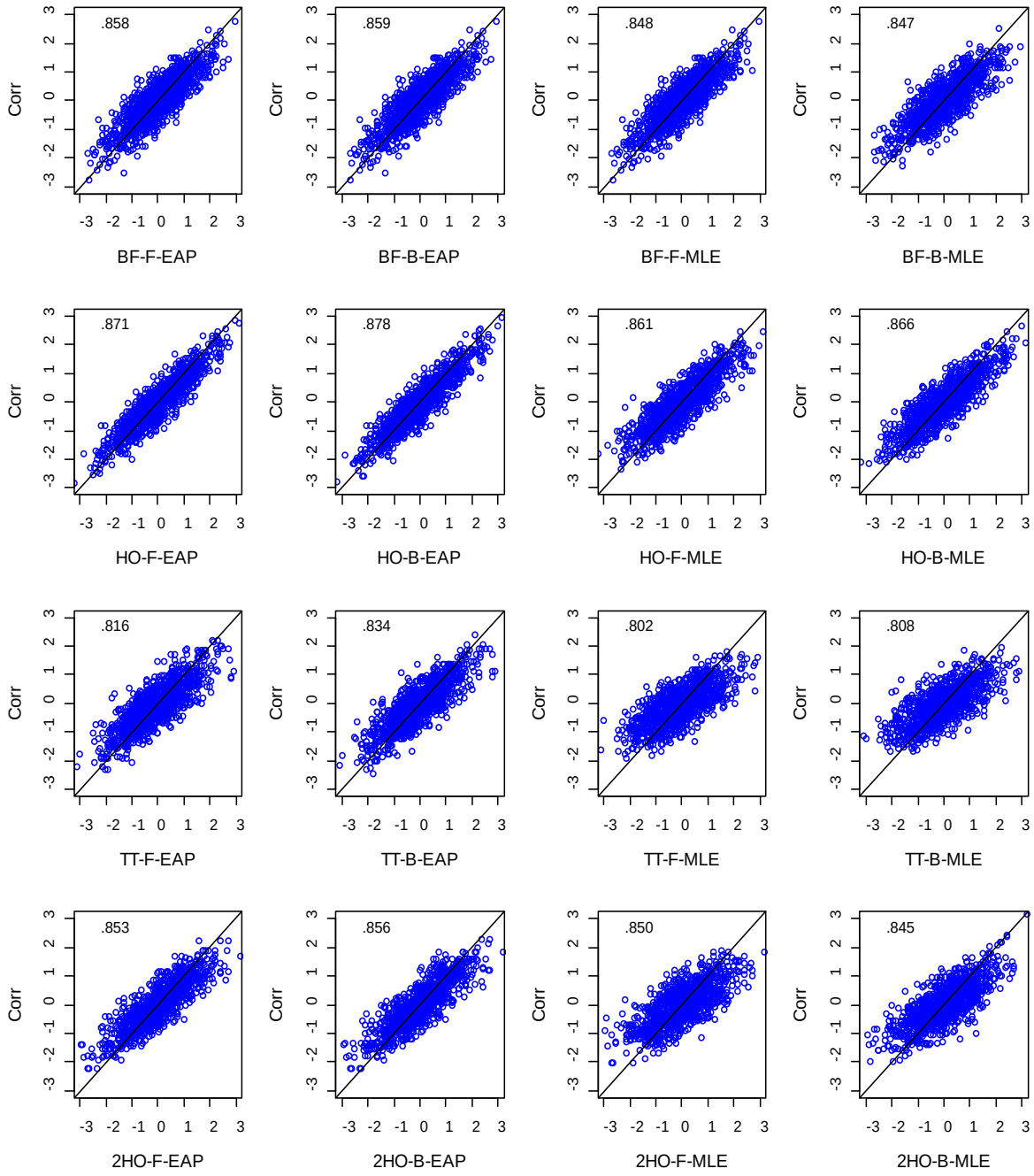
Results for the bifactor IRT model CAT algorithm showed that the correlations for the primary factor were above .844 in all conditions. The correlations for the group factors ranged from 0.836 to 0.877. There were no large differences for the correlations for the primary factor and the group factors both with two and four group factors. There were also no large differences between two item selections. The EAP method provided more accurate proficiency scores for the bifactor IRT model over all other conditions.

Results for the higher-order model showed that the correlations for the primary factor were above 0.861 in both two group factors and four group factors. The correlations for the group factors ranged from 0.832 to 0.881. There were no large correlation differences between the higher-order model with two group factors and the higher-order model with four group factors in case of a general factor. On the other hand, slightly lower correlations were found for the higher-order model with four group factors because the number of items administered on the group factor model with two group factors was twice the number of items administered on the

group factor model with four group factors. While there were no large differences between the two item selections approaches, the EAP method provided more accurate proficiency scores than the MLE method.

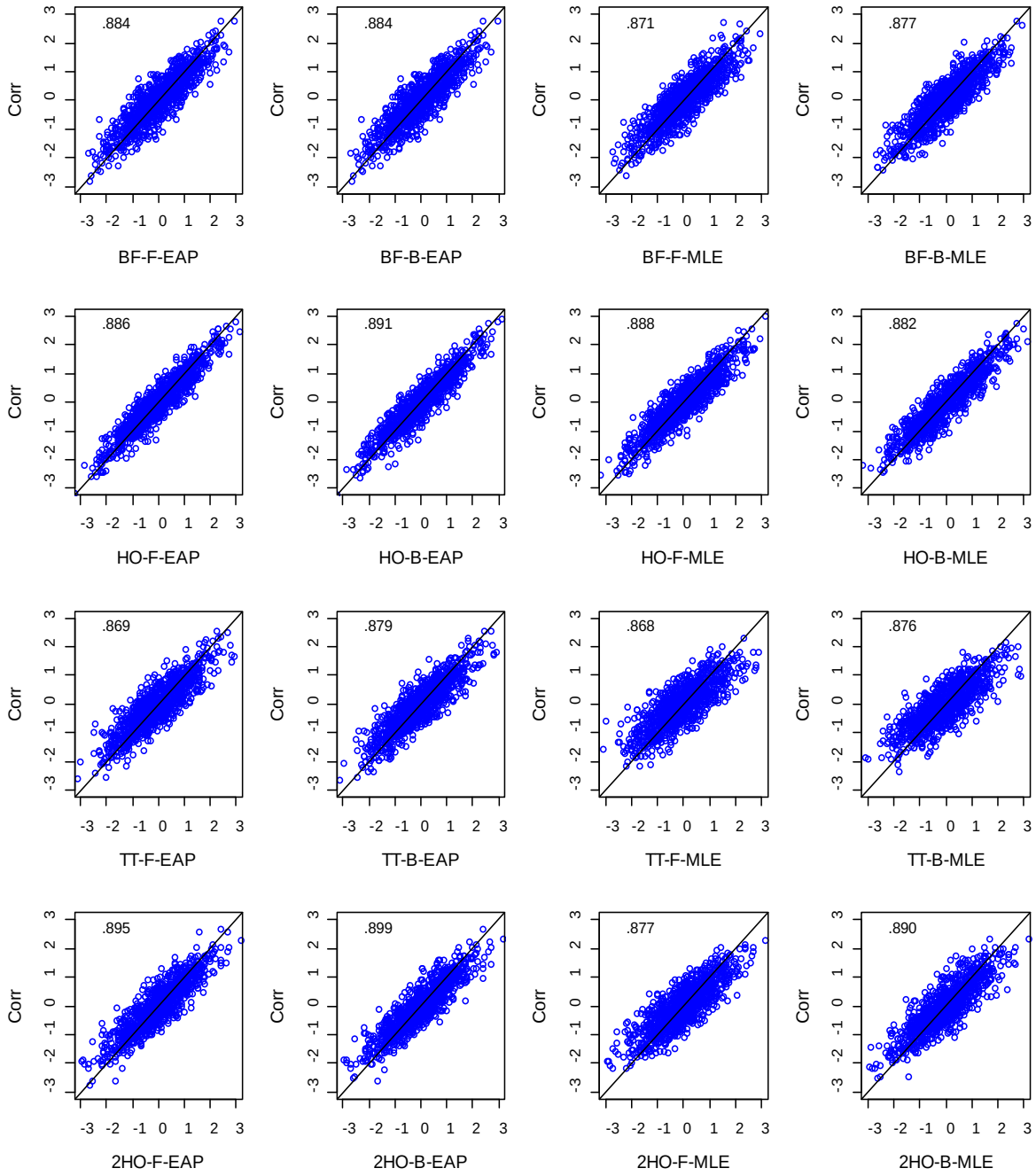
As presented in Tables 4.1 - 4.6, the two-tier IRT model CAT algorithm showed that the correlations ranged from 0.791 to 0.911 for the first primary factor and 0.780 to 0.911 for the second primary factor. The correlations for the group factors ranged from 0.842 to 0.937. The primary factor correlations for the models with two group factors were slightly larger than the primary factor correlations for the model with four group factors. However, group factor correlations did not differ between models with two factors or four factors. The correlation between true and estimated proficiency scores gradually increased when the test length increased and the correlation between the two primary factors was high. There were no large differences between the two item selection approaches, and the EAP method provided more accurate proficiency scores than MLE scoring method in the two-tier model algorithm.

Results for the higher-order IRT model with two primary factors showed that the correlations were from 0.801 to 0.932 for the first primary factor and 0.791 to 0.942 for the second primary factor. The correlations for the group factors ranged from 0.790 to 0.910. The primary factor correlations for the models with two group factors were slightly larger than the primary factor correlations for the model with four group factors. However, group factor correlations did not differ between models with two factors or four factors. The correlation between true and estimated proficiency scores gradually increased when the test length increased and the correlation between the two primary factors was high. There were no large differences between both the two item selection approaches and the EAP and MLE scoring methods in the higher-order IRT model with two primary factors.



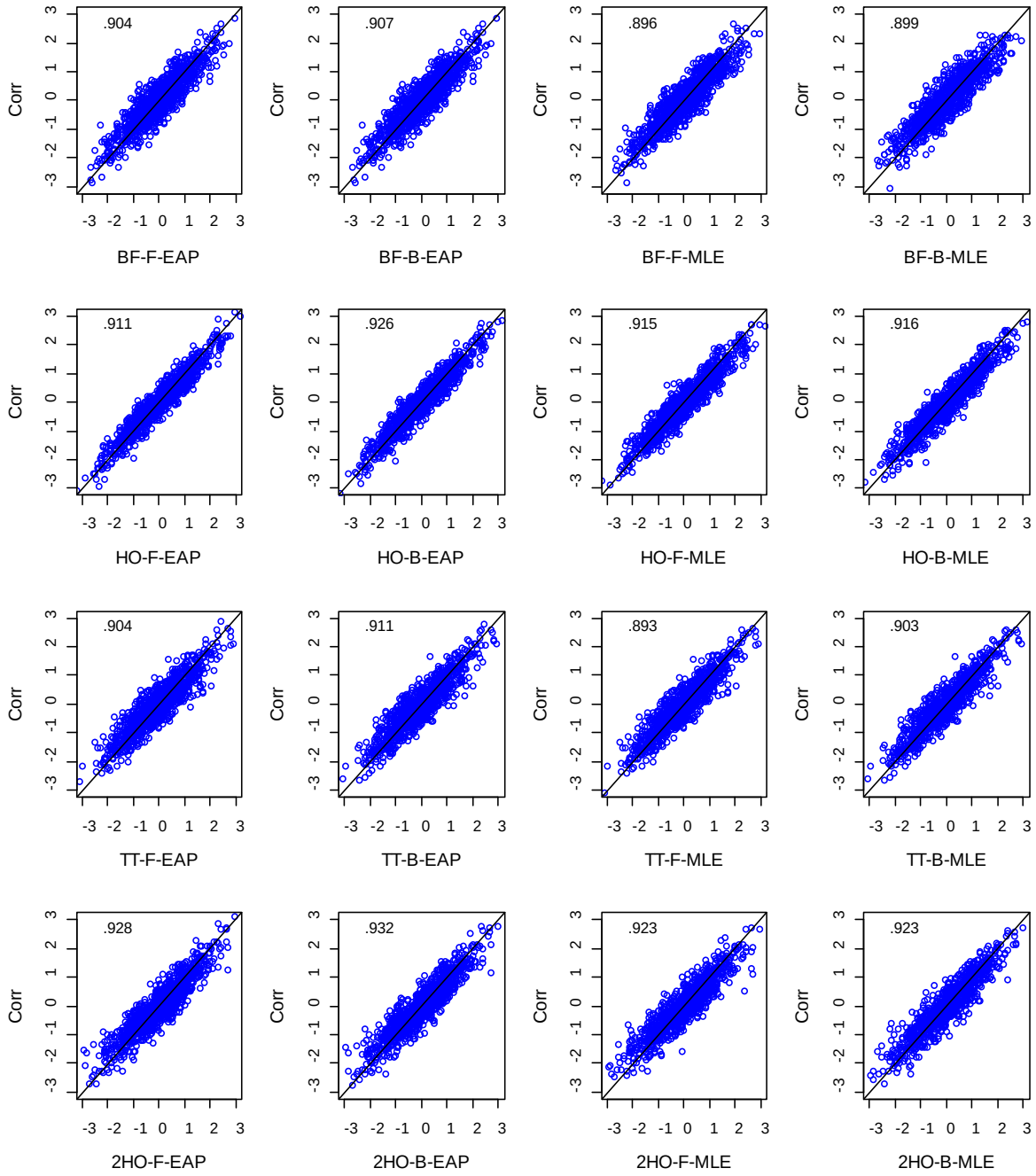
4.1: Correlation between True and Estimated Proficiency Scores (First primary factor with two group factors (40 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two primary factors)



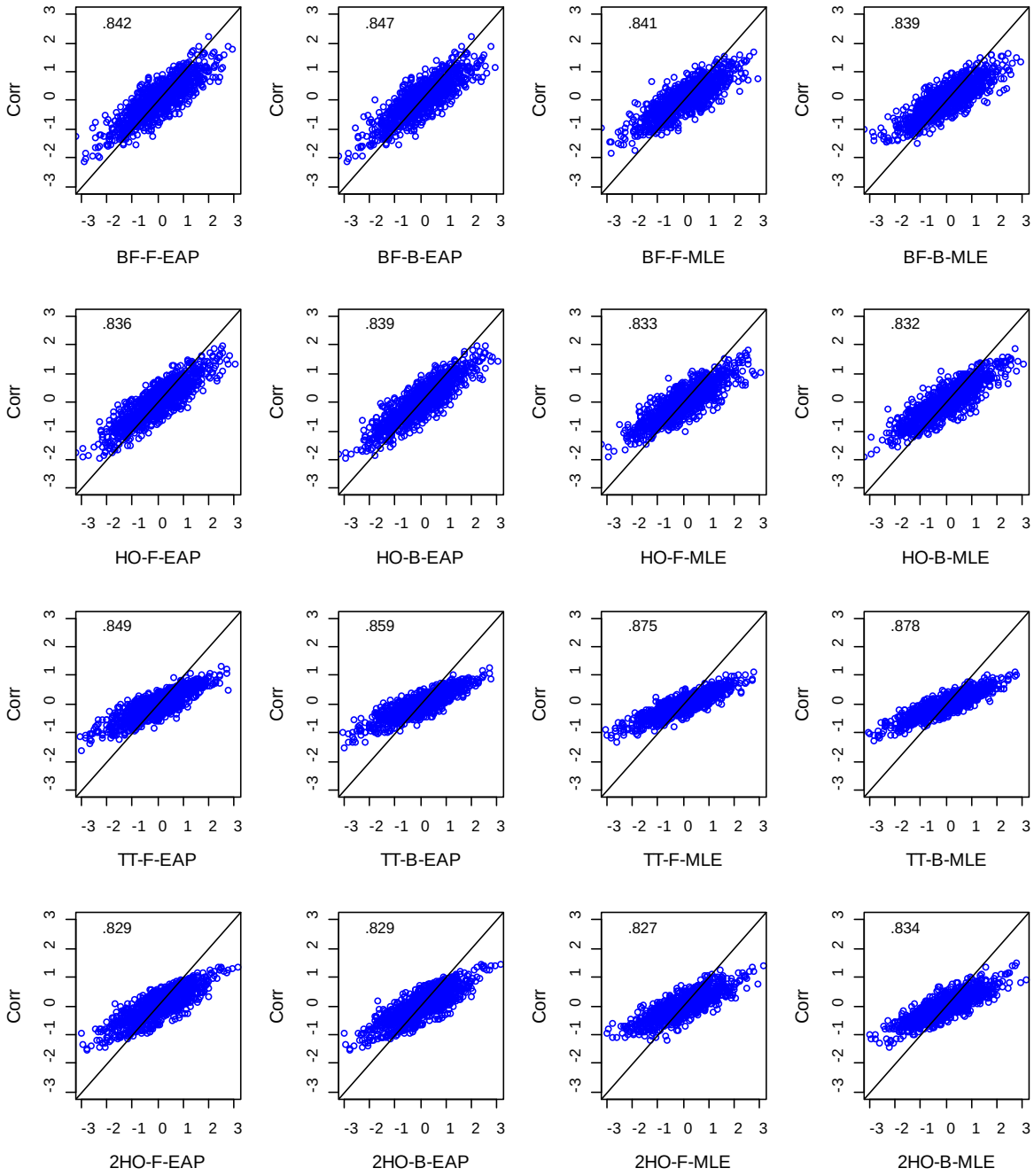
4.2: Correlation between True and Estimated Proficiency Scores (First primary factor with two group factors (80 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two primary factors)



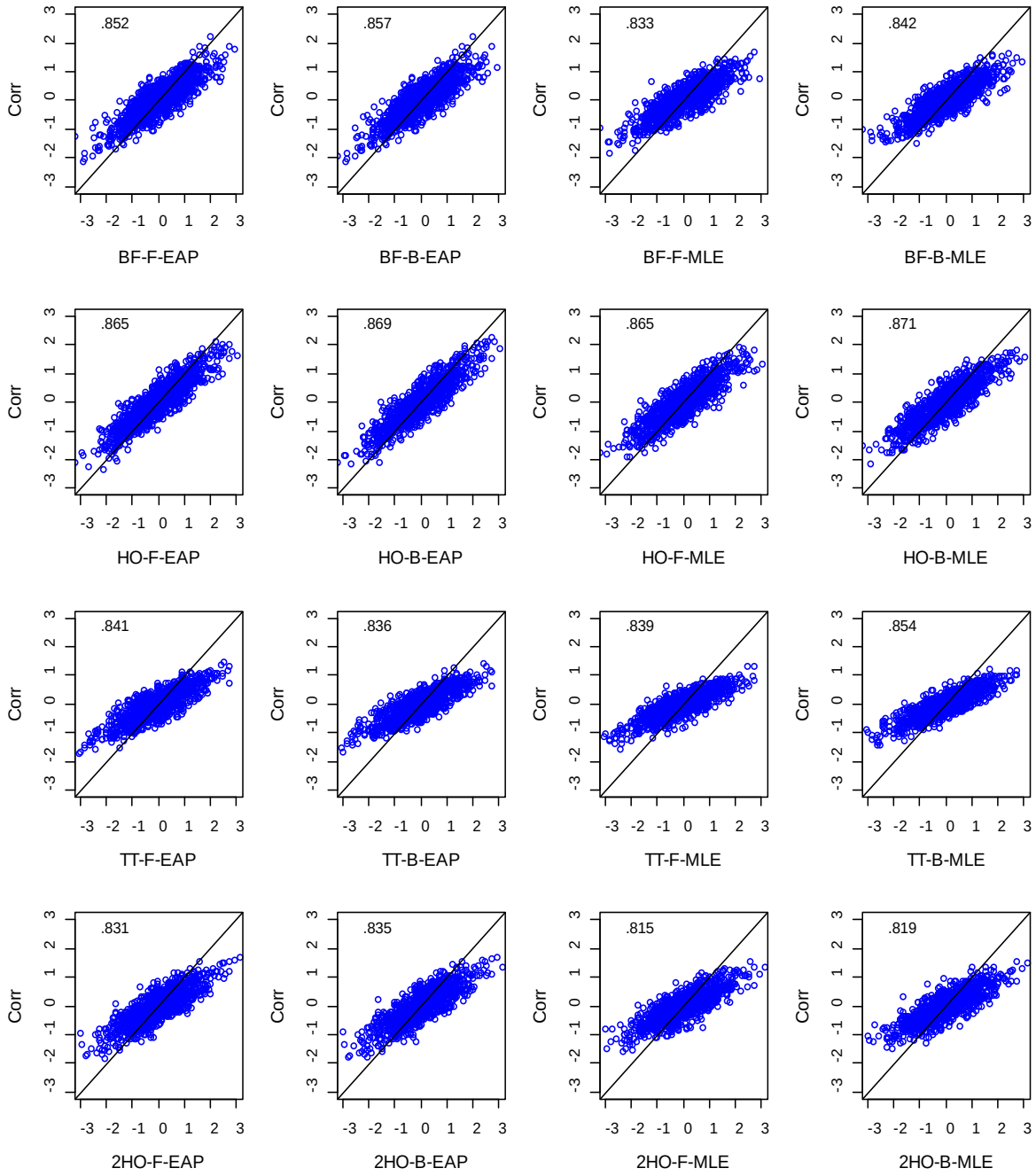
4.3: Correlation between True and Estimated Proficiency Scores (First primary factor with two group factors (160 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two primary factors)



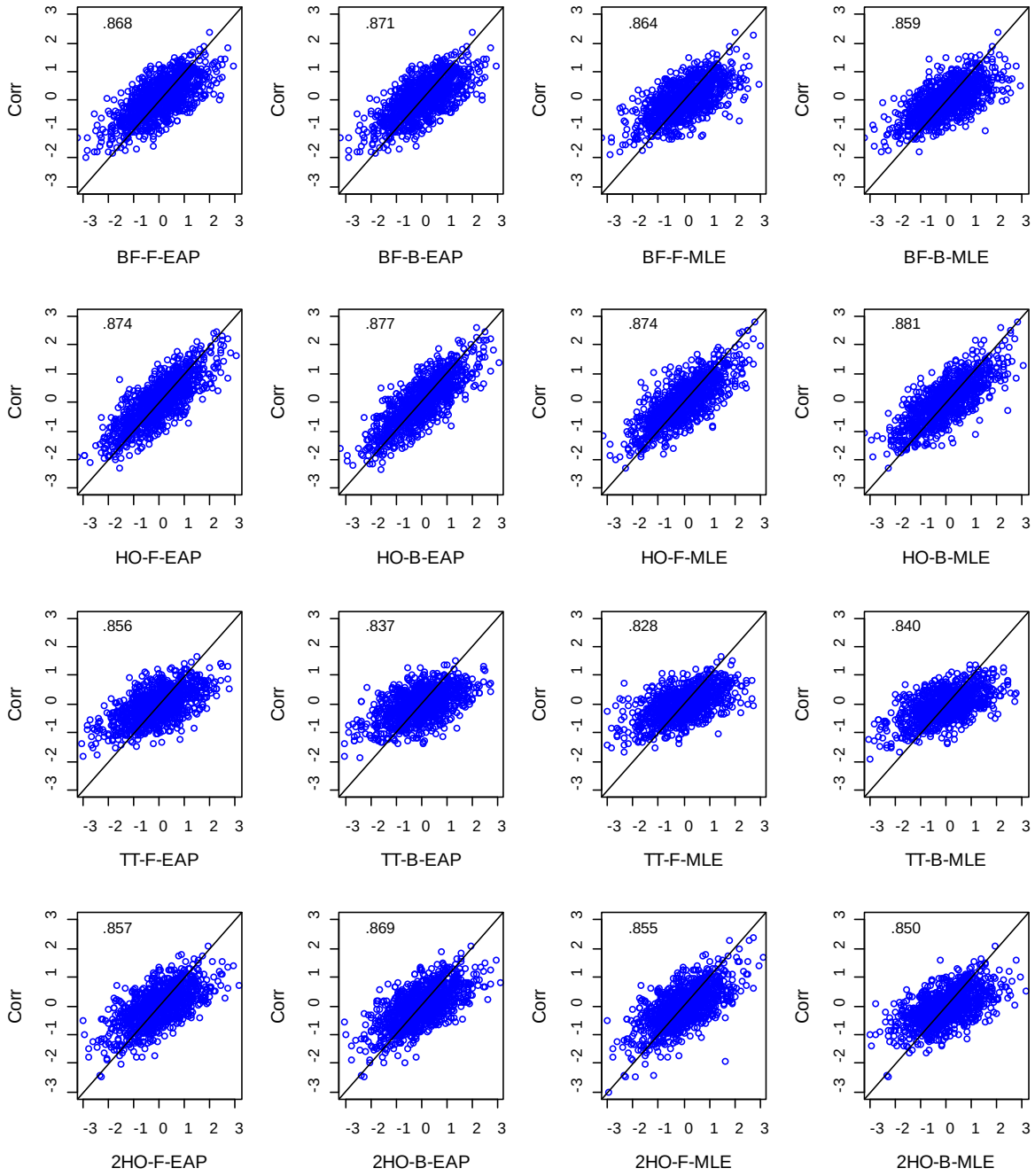
4.4: Correlation between True and Estimated Proficiency Scores (First group factor with two group factors (40 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two primary factors)



4.5: Correlation between True and Estimated Proficiency Scores (First group factor with two group factors (80 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two primary factors)



4.6: Correlation between True and Estimated Proficiency Scores (First group factor with two group factors (160 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two primary factors)

4.2 Average RMSEs

Tables 4.7 - 4.12 and Figures 4.7 - 4.9 show the results of average RMSEs for each of the 10 replications of the multidimensional IRT models. The high correlation between two primary factors ($\rho = 0.7$) is chosen to represent the overall results again. Tables and figures for low and medium correlations between the two primary factors are given in Appendixes B.

As shown in Tables 4.7 - 4.12, average RMSEs decreased when the test length increased for different correlations among the factors, item selection methods and scoring methods. The higher-order IRT model algorithm provided lower RMSEs for primary factors than the other three multidimensional IRT models. However, there were no large differences for group factors among the models. Smaller RMSEs for the primary factor were observed for the bifactor model and the higher-order model with four group factors than those for the bifactor model and the higher-order model with two group factors. In contrast, the RMSEs for the group factor from the bifactor model and the higher-order model with two group factors were slightly smaller than those from the bifactor model and the higher-order model with four group factors. For the conditions with two primary factors, the RMSEs from the higher-order IRT model with two primary factors and the two-tier IRT model with two group factors were slightly smaller than those from two models with four group factors in both of primary and group factors.

As described in Tables 4.7 - 4.12, average RMSEs for the bifactor IRT model for the primary factor ranged from 0.327 to 0.416 and the RMSEs for the group factors ranged from 0.414 to 0.547. Smaller RMSEs for the group factors were observed for the bifactor IRT model with two group factors than those for the bifactor IRT model with four group factors. There were

Table 4.7: Average RMSE (Two group factors, 40 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.415	N/A	0.444	0.472	N/A	
		EAP	0.405		0.452	0.468		
	Bayes	MLE	0.415		0.466	0.470		
		EAP	0.409		0.444	0.470		
Higher-order								
	MFI	MLE	0.414	N/A	0.458	0.468	N/A	
		EAP	0.418		0.450	0.449		
	Bayes	MLE	0.413		0.446	0.463		
		EAP	0.415		0.469	0.482		
Two-tier								
	MFI	MLE	0.460	0.465	0.529	0.558	0.530	0.535
		EAP	0.464	0.458	0.504	0.531	0.496	0.529
	Bayes	MLE	0.473	0.452	0.527	0.558	0.537	0.534
		EAP	0.453	0.458	0.517	0.536	0.513	0.529
Higher-order (2 general factors)								
	MFI	MLE	0.423	0.433	0.502	0.533	0.517	0.561
		EAP	0.421	0.430	0.473	0.505	0.499	0.523
	Bayes	MLE	0.425	0.425	0.508	0.540	0.511	0.558
		EAP	0.415	0.421	0.466	0.526	0.499	0.530

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.8: Average RMSE (Two group factors, 80 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.385	N/A	0.422	0.455	N/A	
		EAP	0.366		0.431	0.441		
	Bayes	MLE	0.374		0.419	0.444		
		EAP	0.372		0.416	0.445		
Higher-order								
	MFI	MLE	0.371	N/A	0.445	0.449	N/A	
		EAP	0.370		0.472	0.475		
	Bayes	MLE	0.365		0.457	0.467		
		EAP	0.367		0.465	0.473		
Two-tier								
	MFI	MLE	0.411	0.424	0.524	0.561	0.506	0.526
		EAP	0.413	0.424	0.476	0.514	0.466	0.504
	Bayes	MLE	0.417	0.416	0.507	0.536	0.501	0.521
		EAP	0.409	0.400	0.499	0.517	0.490	0.491
Higher-order (2 general factors)								
	MFI	MLE	0.391	0.404	0.493	0.512	0.498	0.520
		EAP	0.390	0.396	0.455	0.468	0.473	0.498
	Bayes	MLE	0.382	0.404	0.482	0.502	0.495	0.523
		EAP	0.384	0.390	0.447	0.497	0.464	0.502

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.9: Average RMSE (Two group factors, 160 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.334	N/A	0.417	0.448	N/A	
		EAP	0.334		0.424	0.432		
	Bayes	MLE	0.343		0.416	0.426		
		EAP	0.337		0.414	0.432		
Higher-order								
	MFI	MLE	0.324	N/A	0.437	0.441	N/A	
		EAP	0.324		0.431	0.434		
	Bayes	MLE	0.323		0.428	0.441		
		EAP	0.317		0.439	0.432		
Two-tier								
	MFI	MLE	0.358	0.368	0.486	0.510	0.503	0.490
		EAP	0.357	0.355	0.445	0.477	0.449	0.464
	Bayes	MLE	0.354	0.364	0.486	0.519	0.483	0.490
		EAP	0.362	0.351	0.468	0.486	0.457	0.481
Higher-order (2 general factors)								
	MFI	MLE	0.357	0.369	0.466	0.492	0.469	0.485
		EAP	0.348	0.357	0.426	0.413	0.433	0.464
	Bayes	MLE	0.351	0.361	0.457	0.476	0.464	0.500
		EAP	0.352	0.361	0.422	0.442	0.426	0.478

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.10: Average RMSE (Four group factors, 40 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.416	N/A	0.526	0.525	0.528	0.541	N/A			
		EAP	0.411		0.509	0.500	0.498	0.510				
	Bayes	MLE	0.415		0.523	0.521	0.528	0.547				
		EAP	0.411		0.509	0.502	0.497	0.508				
Higher-order												
	MFI	MLE	0.399	N/A	0.495	0.487	0.523	0.503	N/A			
		EAP	0.397		0.471	0.454	0.474	0.465				
	Bayes	MLE	0.388		0.479	0.477	0.511	0.490				
		EAP	0.399		0.469	0.449	0.463	0.449				
Two-tier												
	MFI	MLE	0.504	0.500	0.558	0.558	0.558	0.529	0.561	0.603	0.528	0.542
		EAP	0.502	0.497	0.550	0.548	0.563	0.526	0.546	0.594	0.527	0.534
	Bayes	MLE	0.500	0.505	0.555	0.556	0.559	0.526	0.569	0.605	0.530	0.542
		EAP	0.506	0.489	0.567	0.557	0.562	0.529	0.571	0.603	0.523	0.541
Higher-order (2 general factors)												
	MFI	MLE	0.488	0.486	0.558	0.603	0.569	0.574	0.556	0.577	0.526	0.546
		EAP	0.486	0.481	0.566	0.609	0.561	0.553	0.552	0.592	0.542	0.566
	Bayes	MLE	0.484	0.489	0.557	0.609	0.558	0.576	0.569	0.568	0.525	0.553
		EAP	0.490	0.483	0.564	0.609	0.569	0.549	0.547	0.593	0.538	0.565

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.11: Average RMSE (Four group factors, 80 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.343		0.507	0.516	0.512	0.523				
		EAP	0.330	N/A	0.486	0.483	0.465	0.485		N/A		
	Bayes	MLE	0.353		0.497	0.523	0.512	0.517				
		EAP	0.338		0.486	0.489	0.475	0.475				
Higher-order												
	MFI	MLE	0.328		0.470	0.489	0.514	0.487				
		EAP	0.327	N/A	0.466	0.454	0.466	0.460		N/A		
	Bayes	MLE	0.333		0.477	0.491	0.502	0.491				
		EAP	0.319		0.464	0.447	0.461	0.439				
Two-tier												
	MFI	MLE	0.420	0.440	0.519	0.515	0.529	0.480	0.525	0.571	0.485	0.512
		EAP	0.424	0.434	0.502	0.502	0.525	0.489	0.509	0.556	0.489	0.497
	Bayes	MLE	0.425	0.439	0.509	0.511	0.524	0.490	0.529	0.570	0.493	0.511
		EAP	0.421	0.435	0.504	0.510	0.524	0.493	0.534	0.566	0.486	0.504
Higher-order (2 general factors)												
	MFI	MLE	0.413	0.423	0.545	0.596	0.566	0.574	0.558	0.577	0.523	0.542
		EAP	0.417	0.417	0.576	0.599	0.556	0.553	0.552	0.606	0.544	0.561
	Bayes	MLE	0.408	0.422	0.548	0.615	0.564	0.586	0.573	0.573	0.527	0.557
		EAP	0.404	0.418	0.564	0.604	0.565	0.552	0.547	0.605	0.541	0.561

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.12: Average RMSE (Four group factors, 160 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.334	N/A	0.487	0.490	0.484	0.492	N/A			
		EAP	0.327		0.462	0.463	0.470	0.484				
	Bayes	MLE	0.340		0.486	0.490	0.487	0.474				
		EAP	0.332		0.482	0.483	0.471	0.469				
Higher-order												
	MFI	MLE	0.319	N/A	0.469	0.460	0.478	0.450	N/A			
		EAP	0.314		0.456	0.449	0.455	0.438				
	Bayes	MLE	0.312		0.459	0.446	0.480	0.460				
		EAP	0.323		0.434	0.441	0.458	0.429				
Two-tier												
	MFI	MLE	0.389	0.383	0.492	0.493	0.511	0.495	0.513	0.539	0.489	0.505
		EAP	0.377	0.384	0.499	0.497	0.510	0.480	0.508	0.537	0.478	0.499
	Bayes	MLE	0.386	0.390	0.487	0.504	0.516	0.487	0.514	0.521	0.484	0.507
		EAP	0.385	0.384	0.489	0.491	0.502	0.478	0.502	0.525	0.479	0.502
Higher-order (2 general factors)												
	MFI	MLE	0.370	0.364	0.537	0.611	0.553	0.561	0.561	0.589	0.526	0.527
		EAP	0.364	0.373	0.568	0.623	0.541	0.551	0.543	0.614	0.544	0.574
	Bayes	MLE	0.367	0.371	0.547	0.620	0.564	0.564	0.552	0.568	0.534	0.542
		EAP	0.366	0.365	0.562	0.622	0.526	0.546	0.534	0.612	0.537	0.574

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

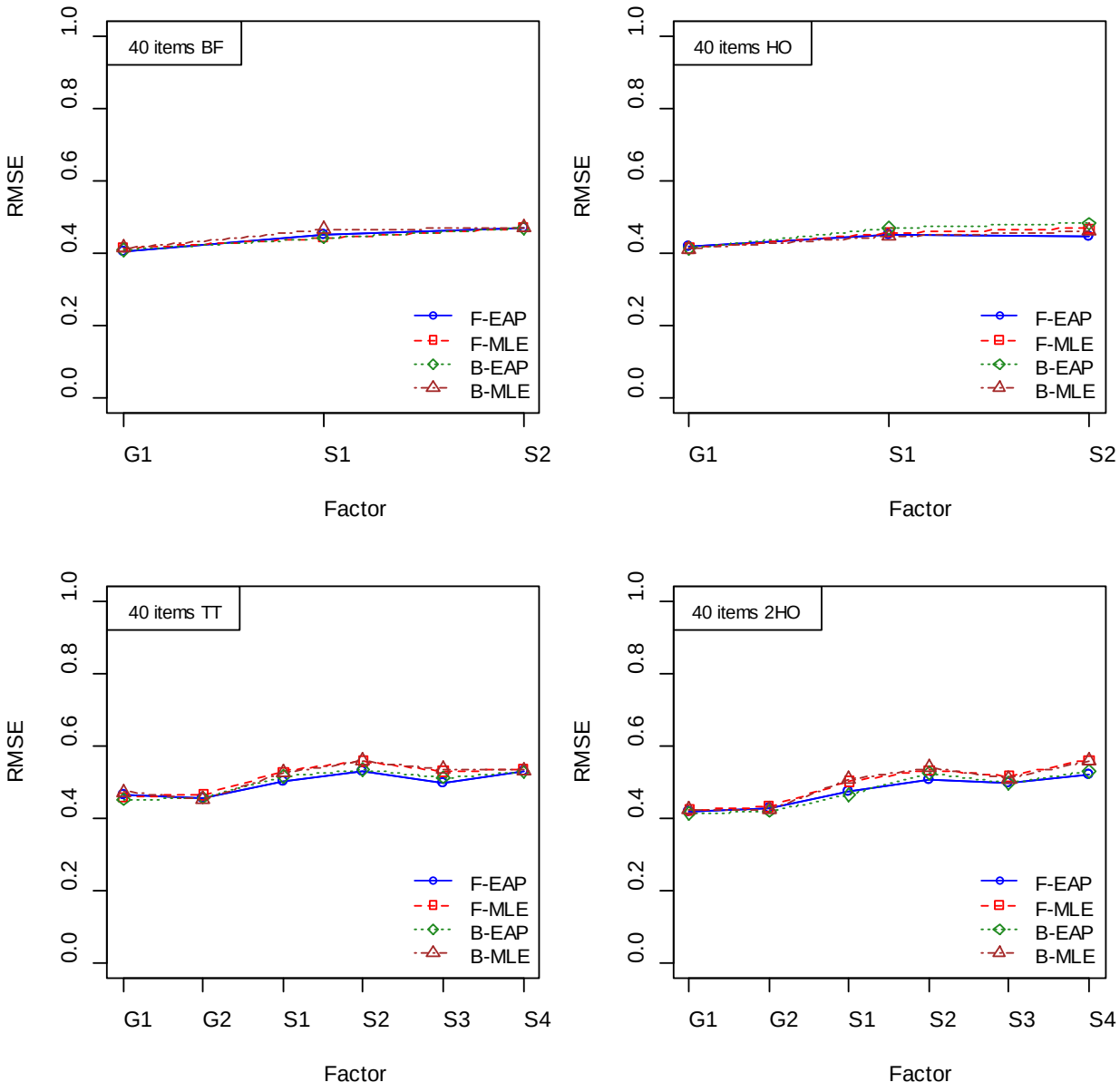
no large differences for the primary factor between two group factors and four group factors. There was slight a tendency for the EAP method to have lower RMSEs, but the differences were not large between the two scoring methods.

Average RMSEs for the higher-order IRT model for the primary factor ranged from 0.312 to 0.418 and the RMSEs for the group factors ranged from 0.428 to 0.511. Smaller RMSEs for the primary factor were observed for the higher-order IRT model with four group factors than those for the higher-order IRT model with two group factors. In contrast, larger RMSEs for the group factors were founded for this model with four group factors than two group factors because of the difference in the number of items administered in the group factors. There were no large differences between MFI and Bayesian item selection methods and scoring methods in the higher-order IRT model algorithm.

Average RMSEs for the two-tier IRT model for the first and the second primary factors ranged from 0.354 to 0.506 and from 0.351 to 0.505 respectively. The RMSEs for the two-tier model for the group factors ranged from 0.445 to 0.605. Smaller RMSEs in primary factors were observed for the two-tier IRT model with two group factors than those for the two-tier IRT model with four group factors, but not in group factors. Average RMSEs decreased when the test length and the correlation between two primary factors increased. There were no large differences between the two item selection methods and EAP and MLE scoring methods in this multidimensional IRT model algorithm.

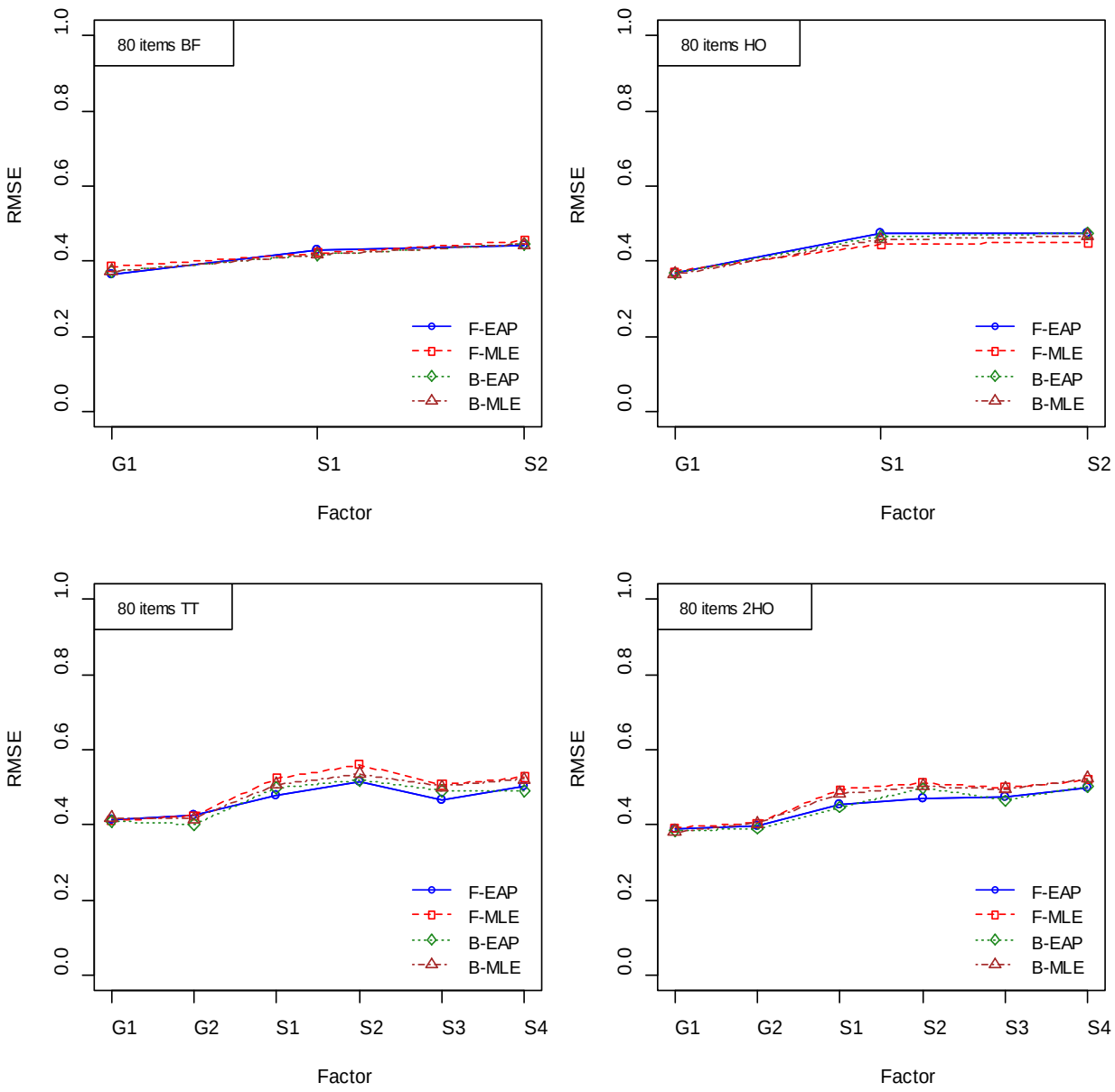
As presented in Tables 4.7 - 4.12, average RMSEs for the higher-order IRT model with two primary factors for the first and the second primary factors ranged from 0.348 to 0.490 and from 0.357 to 0.489 respectively. The RMSEs for the higher-order IRT model with two primary

factors for the group factors ranged from 0.413 to 0.609. Smaller RMSEs for the primary factors and the group factors were observed for the higher-order IRT model with two primary factors with two group factors than those for this model with four group factors. Like the two-tier IRT model case, average RMSEs decreased when the test length and the correlation between the two primary factors increased. There were no large differences between MFI and Bayesian item selection methods and EAP and MLE scoring methods in this IRT model algorithm.



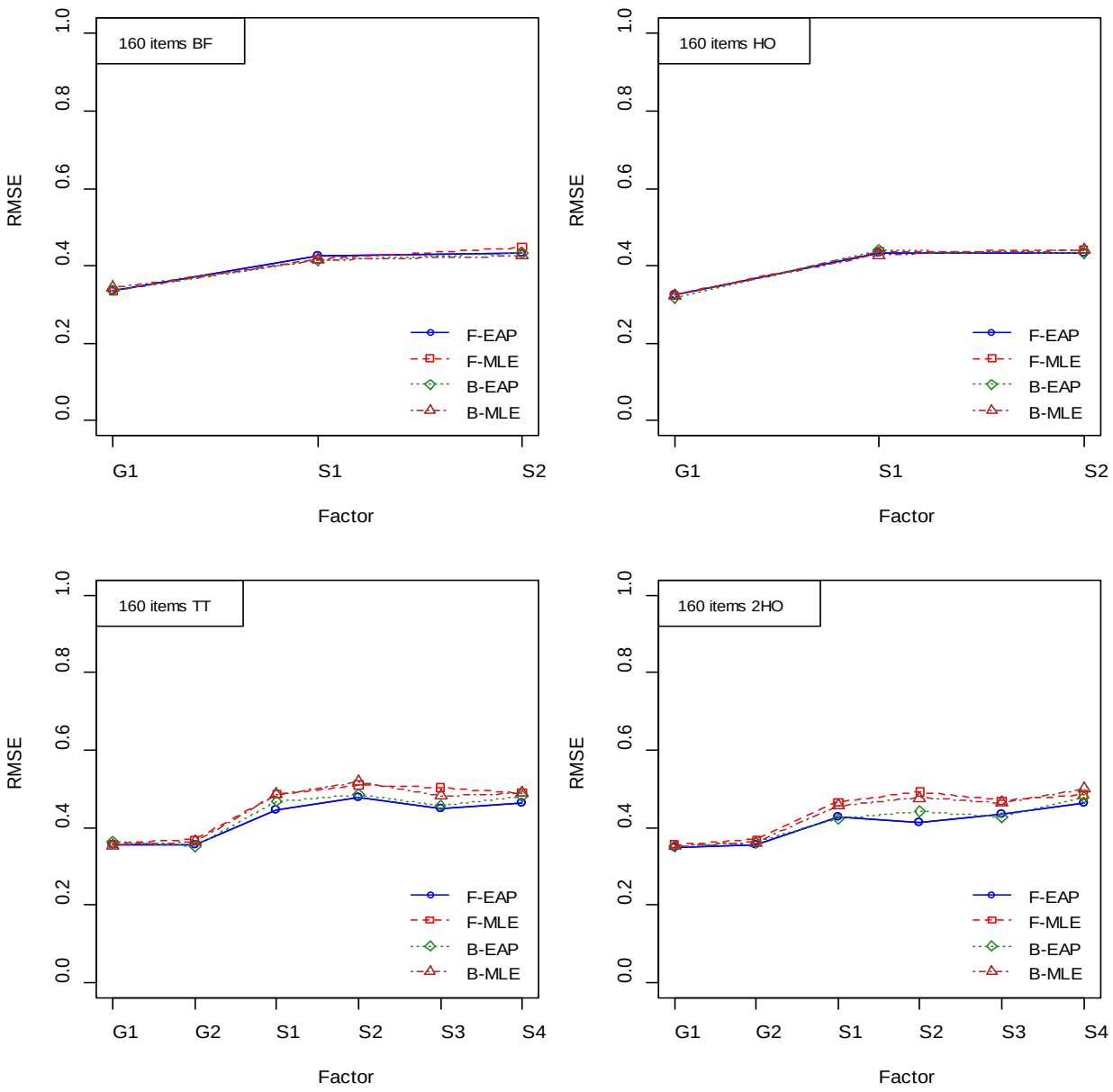
4.7: Average RMSE (Two group factors (40 items))

Note, F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method
(BF: Bifactor IRT model, HO: Higher-order IRT model, TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



4.8: Average RMSE (Two group factors (80 items))

Note, F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method
(BF: Bifactor IRT model, HO: Higher-order IRT model, TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



4.9: Average RMSE (Two group factors (160 items))

Note, F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method
(BF: Bifactor IRT model, HO: Higher-order IRT model, TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)

4.3 Average SEs

Tables 4.13 - 4.18 and Figures 4.10 - 4.15 show the SE results for each of 10 replications of the bifactor, the higher-order, the two-tier, and the higher-order IRT model with two primary factors. The high correlation between two primary factors ($\rho = 0.7$) is chosen to represent the overall results again. Tables and figures for low and medium correlation between two general factors are given in Appendixes C.

The standard error of measurement is a critical value for CAT because it is widely used as a stopping rule. As a termination criterion, many researchers used a pre-specified observed standard error (0.3 - 0.5) of the θ estimates (e.g., Gibbons et al., 2008; Immekus, Gibbons, & Rush, 2007; Weiss & Gibbons, 2007).

As shown in Tables 4.13 - 4.18, average SEs decreased as the number of items and correlation between two primary factors increased. The bifactor IRT model and the higher-order IRT model provided smaller SEs for both of primary and group factors than the two-tier and the higher-order IRT models with two primary factors. Smaller SEs for the primary factor were observed for the bifactor model and the higher-order model with four group factors than those for the bifactor model and the higher-order model with two group factors. In contrasts, the SEs for group factors from the bifactor model and the higher-order model with two group factors were slightly smaller than those from the bifactor model and the higher-order model with four group factors. For the generating conditions with two primary factors, the SEs for two group factors from the two-tier IRT model and the higher-order IRT model with two primary factors were slightly smaller than those from two models with four group factors.

Table 4.13: Average SE (Two group factors, 40 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.503	N/A	0.801	0.800	N/A	
		EAP	0.495		0.788	0.795		
	Bayes	MLE	0.500		0.799	0.802		
		EAP	0.496		0.783	0.789		
Higher-order								
	MFI	MLE	0.481	N/A	0.846	0.787	N/A	
		EAP	0.483		0.762	0.761		
	Bayes	MLE	0.508		0.833	0.838		
		EAP	0.476		0.756	0.708		
Two-tier								
	MFI	MLE	0.525	0.573	0.958	0.931	0.966	0.890
		EAP	0.518	0.565	0.933	0.917	0.903	0.924
	Bayes	MLE	0.525	0.564	0.958	0.923	0.909	0.976
		EAP	0.513	0.558	0.943	0.956	0.902	0.940
Higher-order (2 general factors)								
	MFI	MLE	0.528	0.565	0.827	0.941	0.905	0.982
		EAP	0.525	0.565	0.834	0.871	0.892	0.903
	Bayes	MLE	0.538	0.549	0.835	0.920	0.939	0.977
		EAP	0.523	0.561	0.830	0.903	0.889	0.902

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.14: Average SE (Two group factors, 80 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.381	N/A	0.765	0.772	N/A	
		EAP	0.370		0.742	0.758		
	Bayes	MLE	0.382		0.766	0.772		
		EAP	0.370		0.742	0.758		
Higher-order								
	MFI	MLE	0.423	N/A	0.560	0.560	N/A	
		EAP	0.413		0.542	0.505		
	Bayes	MLE	0.424		0.560	0.560		
		EAP	0.423		0.523	0.496		
Two-tier								
	MFI	MLE	0.402	0.401	0.879	0.881	0.877	0.871
		EAP	0.403	0.389	0.850	0.863	0.853	0.891
	Bayes	MLE	0.402	0.398	0.877	0.882	0.878	0.871
		EAP	0.386	0.377	0.889	0.872	0.884	0.893
Higher-order (2 general factors)								
	MFI	MLE	0.453	0.458	0.629	0.678	0.622	0.689
		EAP	0.456	0.455	0.604	0.636	0.631	0.693
	Bayes	MLE	0.452	0.460	0.629	0.679	0.623	0.690
		EAP	0.447	0.455	0.595	0.675	0.626	0.700

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.15: Average SE (Two group factors, 160 items)

MIRT Model	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
Bifactor								
	MFI	MLE	0.334	N/A	0.735	0.743	N/A	
		EAP	0.331		0.710	0.725		
	Bayes	MLE	0.334		0.734	0.743		
		EAP	0.331		0.710	0.725		
Higher-order								
	MFI	MLE	0.335	N/A	0.425	0.423	N/A	
		EAP	0.330		0.409	0.387		
	Bayes	MLE	0.335		0.424	0.425		
		EAP	0.335		0.402	0.388		
Two-tier								
	MFI	MLE	0.316	0.320	0.830	0.835	0.831	0.820
		EAP	0.335	0.324	0.785	0.803	0.811	0.800
	Bayes	MLE	0.307	0.335	0.827	0.834	0.830	0.821
		EAP	0.319	0.314	0.813	0.812	0.825	0.826
Higher-order (2 general factors)								
	MFI	MLE	0.386	0.396	0.570	0.629	0.577	0.648
		EAP	0.381	0.391	0.546	0.566	0.580	0.644
	Bayes	MLE	0.371	0.375	0.570	0.629	0.576	0.648
		EAP	0.380	0.393	0.533	0.599	0.568	0.657

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.16: Average SE (Four group factors, 40 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.489	N/A	0.904	0.916	0.975	0.968	N/A			
		EAP	0.477		0.897	0.909	0.923	0.970				
	Bayes	MLE	0.510		0.917	0.974	0.953	0.967				
		EAP	0.477		0.899	0.905	0.943	0.959				
Higher-order												
	MFI	MLE	0.459	N/A	0.861	0.820	0.816	0.865	N/A			
		EAP	0.460		0.806	0.793	0.826	0.839				
	Bayes	MLE	0.490		0.841	0.858	0.847	0.880				
		EAP	0.476		0.785	0.794	0.851	0.793				
Two-tier												
	MFI	MLE	0.588	0.570	0.940	0.986	0.988	0.969	0.987	0.976	0.987	0.959
		EAP	0.607	0.579	0.974	0.965	0.941	0.946	0.927	0.969	0.980	0.903
	Bayes	MLE	0.582	0.569	0.975	0.990	0.996	0.993	0.937	0.950	0.968	0.984
		EAP	0.571	0.564	0.970	0.988	0.990	0.957	0.995	0.969	0.982	0.975
Higher-order (2 general factors)												
	MFI	MLE	0.595	0.571	0.958	0.963	0.947	0.985	0.989	0.979	0.981	0.938
		EAP	0.588	0.562	0.951	0.936	0.893	0.917	0.981	0.912	0.976	0.908
	Bayes	MLE	0.591	0.576	0.965	0.972	0.952	0.989	0.983	0.977	0.978	0.937
		EAP	0.577	0.561	0.951	0.935	0.894	0.975	0.982	0.911	0.980	0.908

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.17: Average SE (Four group factors, 80 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.362	N/A	0.872	0.878	0.874	0.879	N/A			
		EAP	0.322		0.871	0.846	0.830	0.847				
	Bayes	MLE	0.351		0.820	0.876	0.875	0.877				
		EAP	0.322		0.871	0.846	0.830	0.847				
Higher-order												
	MFI	MLE	0.392	N/A	0.624	0.622	0.668	0.628	N/A			
		EAP	0.358		0.615	0.612	0.647	0.609				
	Bayes	MLE	0.372		0.624	0.620	0.667	0.629				
		EAP	0.378		0.588	0.585	0.642	0.584				
Two-tier												
	MFI	MLE	0.463	0.517	0.967	0.971	0.983	0.985	0.982	0.982	0.979	0.983
		EAP	0.452	0.496	0.961	0.951	0.982	0.962	0.950	0.973	0.980	0.956
	Bayes	MLE	0.463	0.515	0.968	0.971	0.984	0.985	0.981	0.984	0.979	0.984
		EAP	0.441	0.481	0.970	0.965	0.983	0.972	0.990	0.981	0.981	0.960
Higher-order (2 general factors)												
	MFI	MLE	0.532	0.570	0.700	0.728	0.724	0.752	0.748	0.736	0.749	0.726
		EAP	0.533	0.555	0.686	0.690	0.682	0.706	0.731	0.711	0.738	0.679
	Bayes	MLE	0.534	0.570	0.701	0.728	0.723	0.753	0.749	0.737	0.749	0.726
		EAP	0.529	0.556	0.682	0.687	0.675	0.736	0.730	0.710	0.740	0.679

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

Table 4.18: Average SE (Four group factors, 160 items)

MIRT Model	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Bifactor												
	MFI	MLE	0.245	N/A	0.815	0.823	0.817	0.826	N/A			
		EAP	0.254		0.792	0.769	0.765	0.766				
	Bayes	MLE	0.246		0.761	0.824	0.818	0.822				
		EAP	0.254		0.792	0.769	0.765	0.766				
Higher-order												
	MFI	MLE	0.296	N/A	0.573	0.567	0.617	0.577	N/A			
		EAP	0.297		0.534	0.542	0.553	0.552				
	Bayes	MLE	0.296		0.572	0.566	0.616	0.578				
		EAP	0.315		0.517	0.517	0.552	0.530				
Two-tier												
	MFI	MLE	0.348	0.410	0.952	0.956	0.956	0.959	0.965	0.968	0.959	0.969
		EAP	0.354	0.394	0.934	0.928	0.925	0.912	0.919	0.950	0.954	0.924
	Bayes	MLE	0.347	0.411	0.952	0.957	0.954	0.958	0.966	0.968	0.959	0.971
		EAP	0.347	0.385	0.942	0.940	0.932	0.920	0.943	0.965	0.957	0.930
Higher-order (2 general factors)												
	MFI	MLE	0.419	0.470	0.677	0.706	0.676	0.725	0.731	0.719	0.728	0.700
		EAP	0.434	0.471	0.650	0.662	0.609	0.675	0.691	0.690	0.709	0.647
	Bayes	MLE	0.418	0.469	0.677	0.705	0.675	0.724	0.730	0.719	0.727	0.701
		EAP	0.431	0.471	0.648	0.663	0.608	0.682	0.702	0.691	0.712	0.644

Note. G_1*: First general(primary) factor,* s_1*: First group(specific) factor.*

As described in Tables 4.13 - 4.18 above, average SEs for the bifactor IRT model for the primary factor ranged from 0.245 to 0.503 and the SEs for the group factors ranged from 0.710 to 0.801. Smaller SEs for the primary factor were observed for the bifactor IRT model with four group factors than those for this model with two group factors. In contrast, larger SEs for the group factors were found for this model with four group factors than two group factors because of the difference in the number of items administered in the group factors. There were no large differences between MFI and Bayesian item selection methods in this model. Overall, EAP scoring method showed lower SEs than the MLE method in the bifactor IRT model.

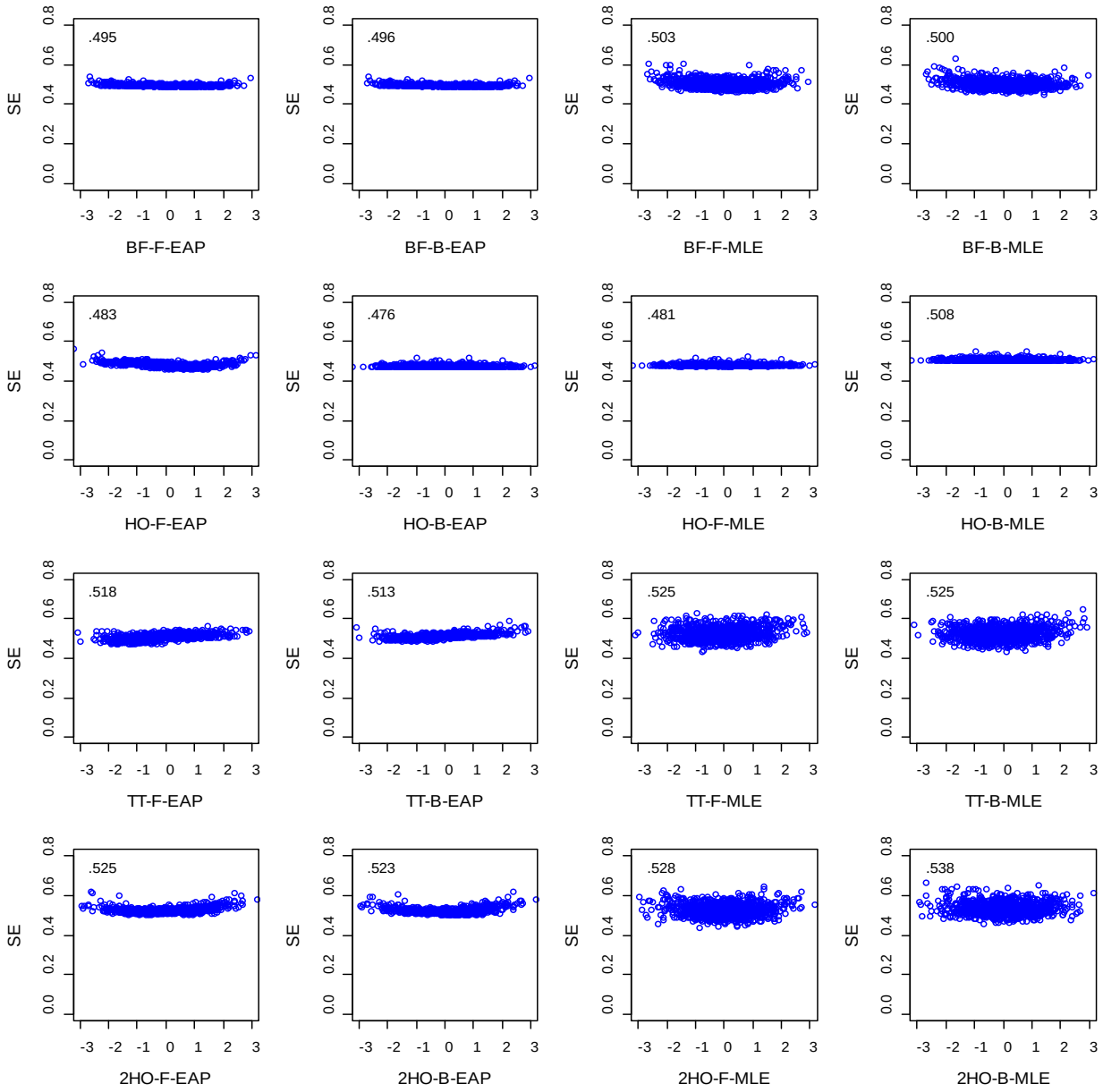
Average SEs for the higher-order IRT model for the general factor ranged from 0.296 to 0.508 and the SEs for the group factors ranged from 0.387 to 0.880. Like in the bifactor IRT model case, smaller SEs for the primary factor were observed for the higher-order IRT model with four group factors than those for the higher-order IRT model with two group factors. In contrast, larger SEs for the group factors were found for this model with four group factors than two group factors because of the difference in the number of items administered in the group factors. There were no large differences between the MFI and Bayesian item selection methods and scoring methods in the higher-order IRT model algorithm.

Average SEs for the two-tier IRT model for the first and the second primary factors ranged from 0.307 to 0.607 and from 0.314 to 0.579 respectively. The SEs for the two-tier model for the group factors ranged from 0.785 to 0.996. Smaller SEs for the primary factors and the group factors were observed for the two-tier IRT model with two group factors than those for the two-tier IRT model with four group factors. Average SEs decreased when the test length and the correlation between the two primary factors increased. There were no large differences between

the MFI and Bayesian item selection methods in this model. Overall, EAP scoring method provided lower SEs than the MLE method in the two-tier IRT model.

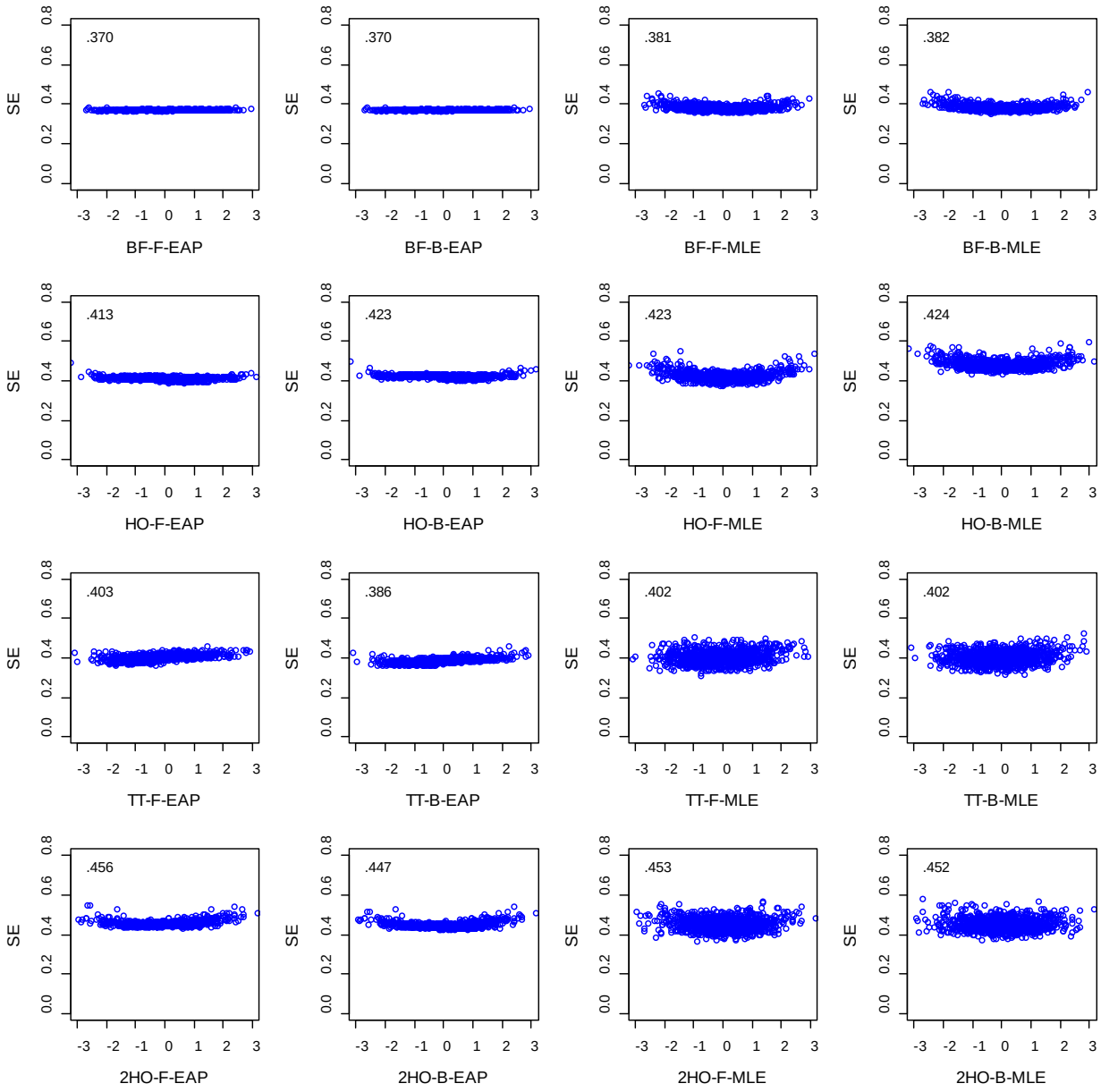
As presented in Tables 4.13 - 4.18, average SEs for the higher-order IRT model with two primary factors for the first and the second primary factors ranged from 0.371 to 0.595 and from 0.375 to 0.576 respectively. The SEs for the higher-order IRT model with two primary factors for the group factors ranged from 0.533 to 0.989. Smaller SEs for the primary factors and the group factors were observed for the higher-order IRT model with two primary factors and two group factors than those for this model with four group factors. Like in the two-tier IRT model case, average SEs decreased when the test length and the correlation between the two primary factors increased. There were no large differences between the MFI and Bayesian item selection methods and the EAP and MLE scoring methods in this multidimensional IRT model algorithm.

Figures 4.10 - 4.15 shows the estimated SEs for the first primary factor and the first group factor using four multidimensional IRT models. It can be found that all of the estimated SEs across the ability levels were very similar. It can also be shown that average SEs for primary factors were acceptable by using CAT. In contrast, only the higher-order IRT model with 160 items for group factors was acceptable. A comparison of Figures revealed that a longer test yielded a smaller SEs.



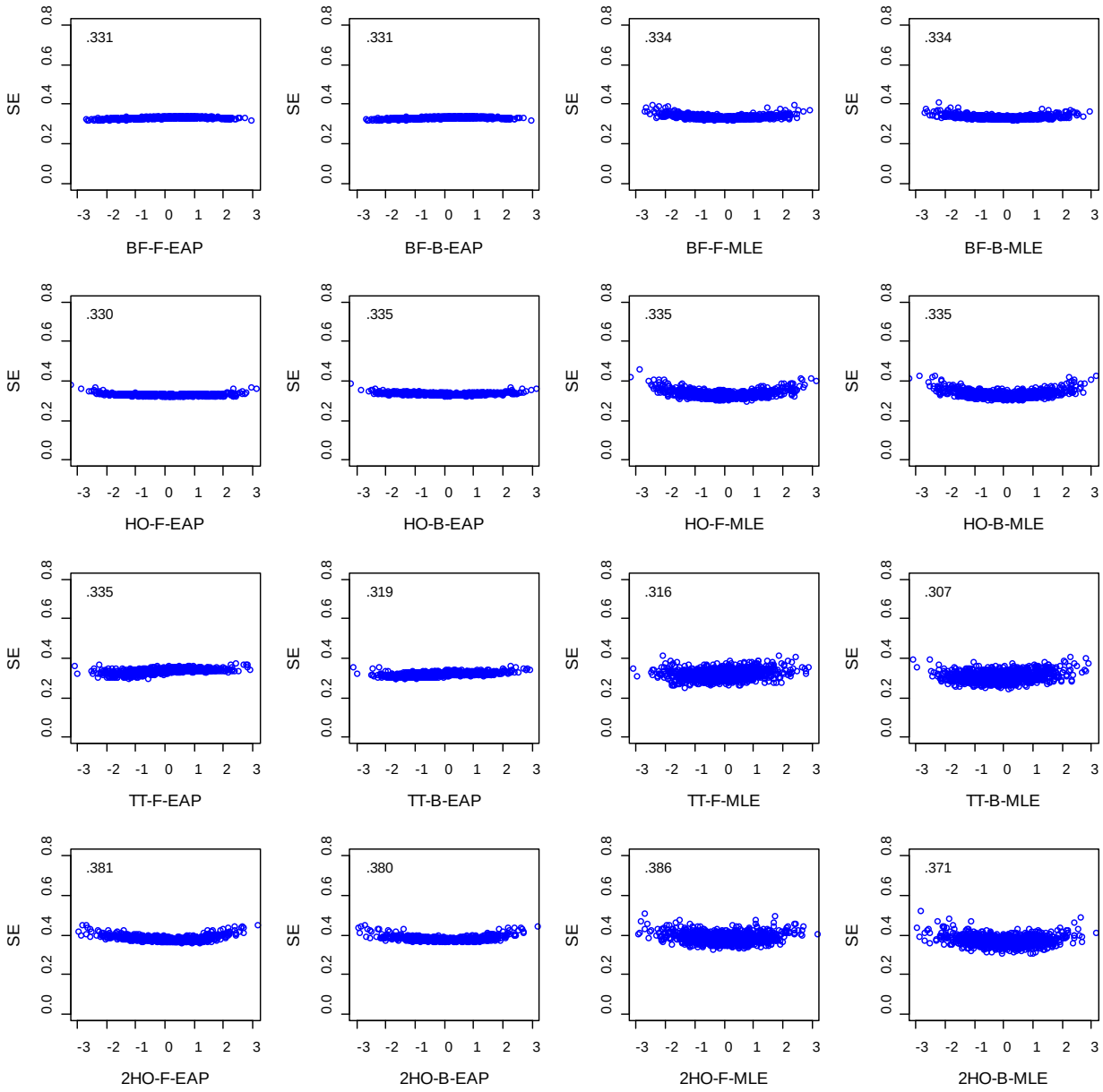
4.10: Average SE (First primary factor with two group factors (40 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



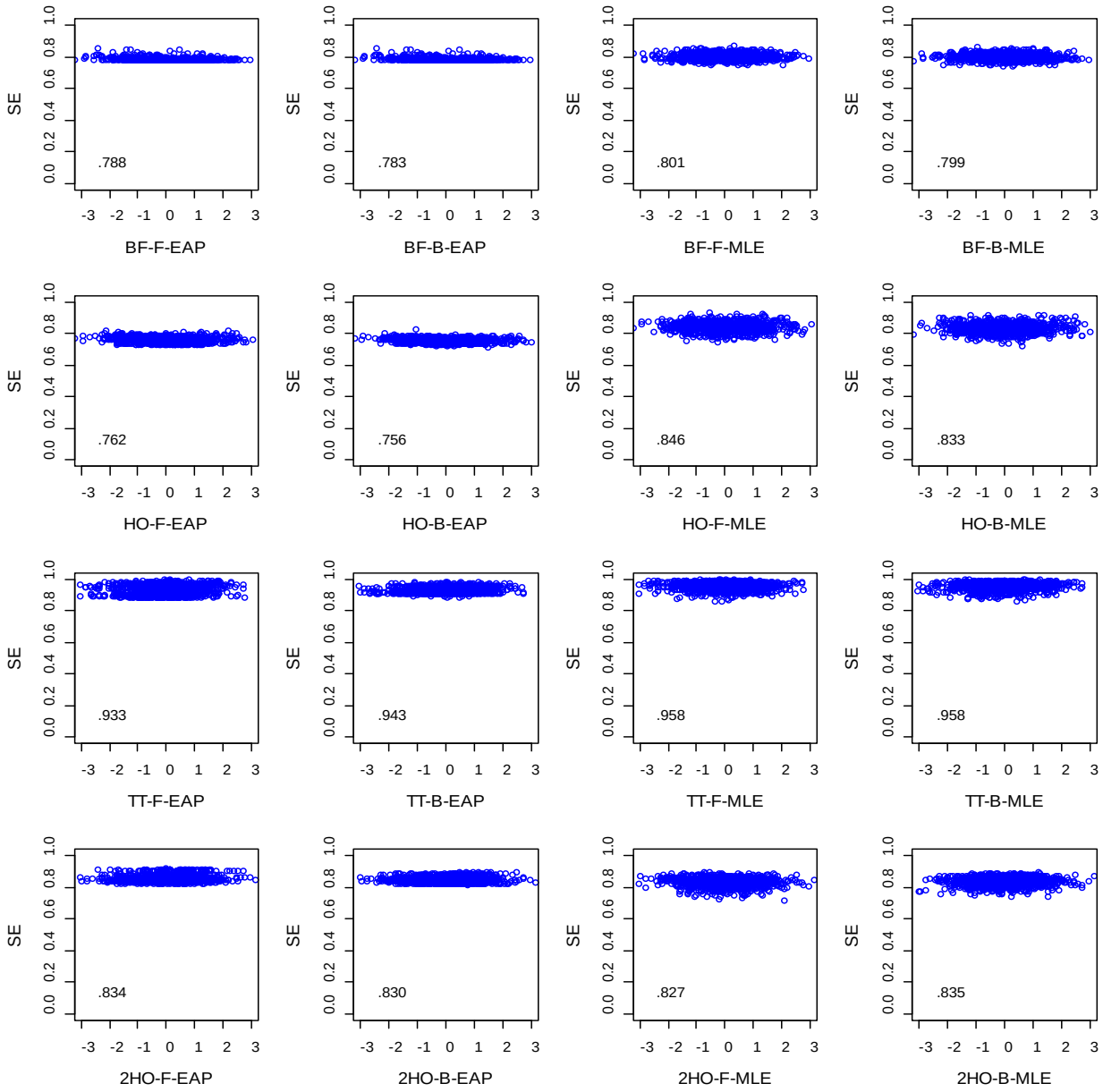
4.11: Average SE (First primary factor with two group factors (80 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



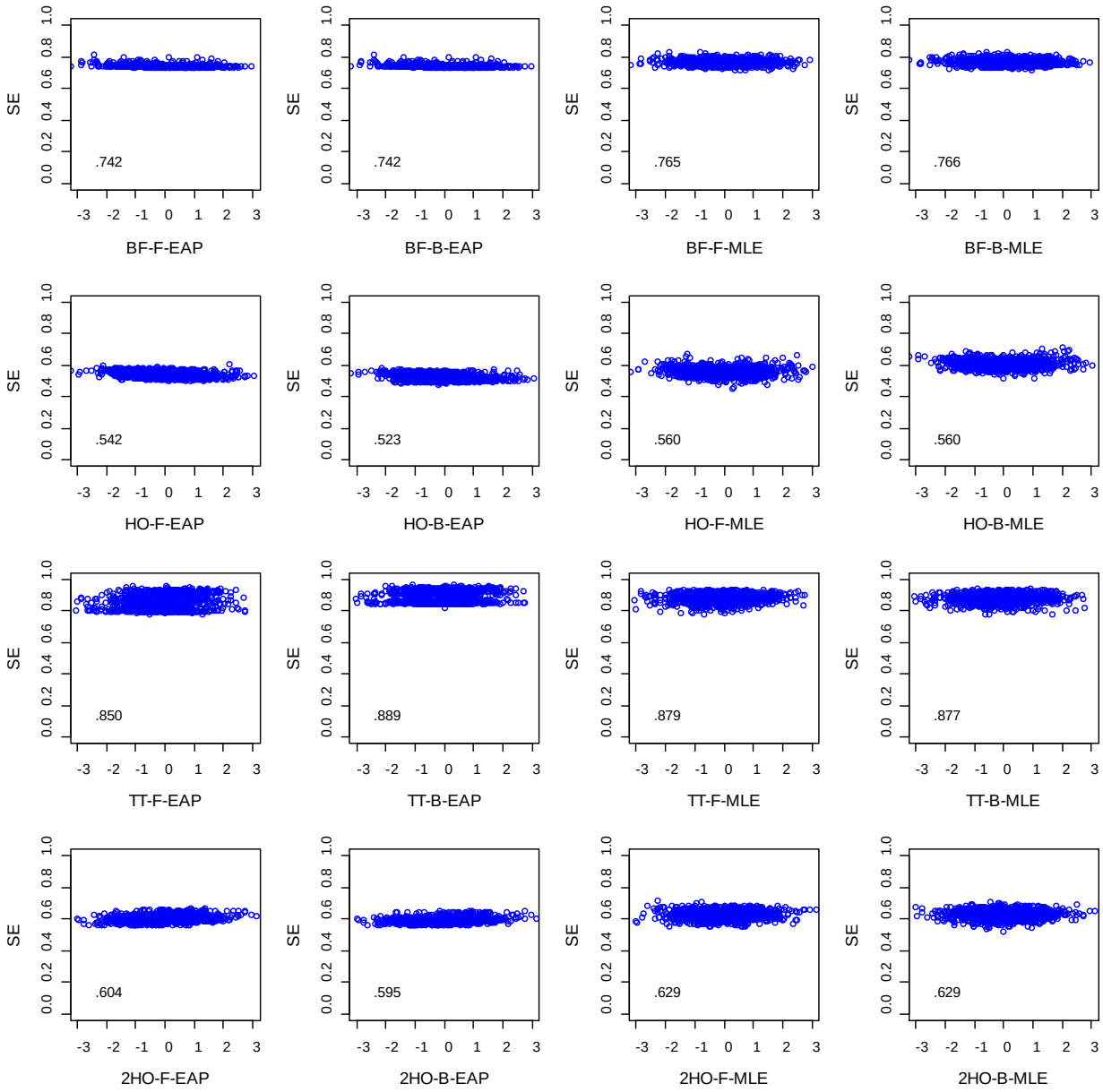
4.12: Average SE (First primary factor with two group factors (160 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



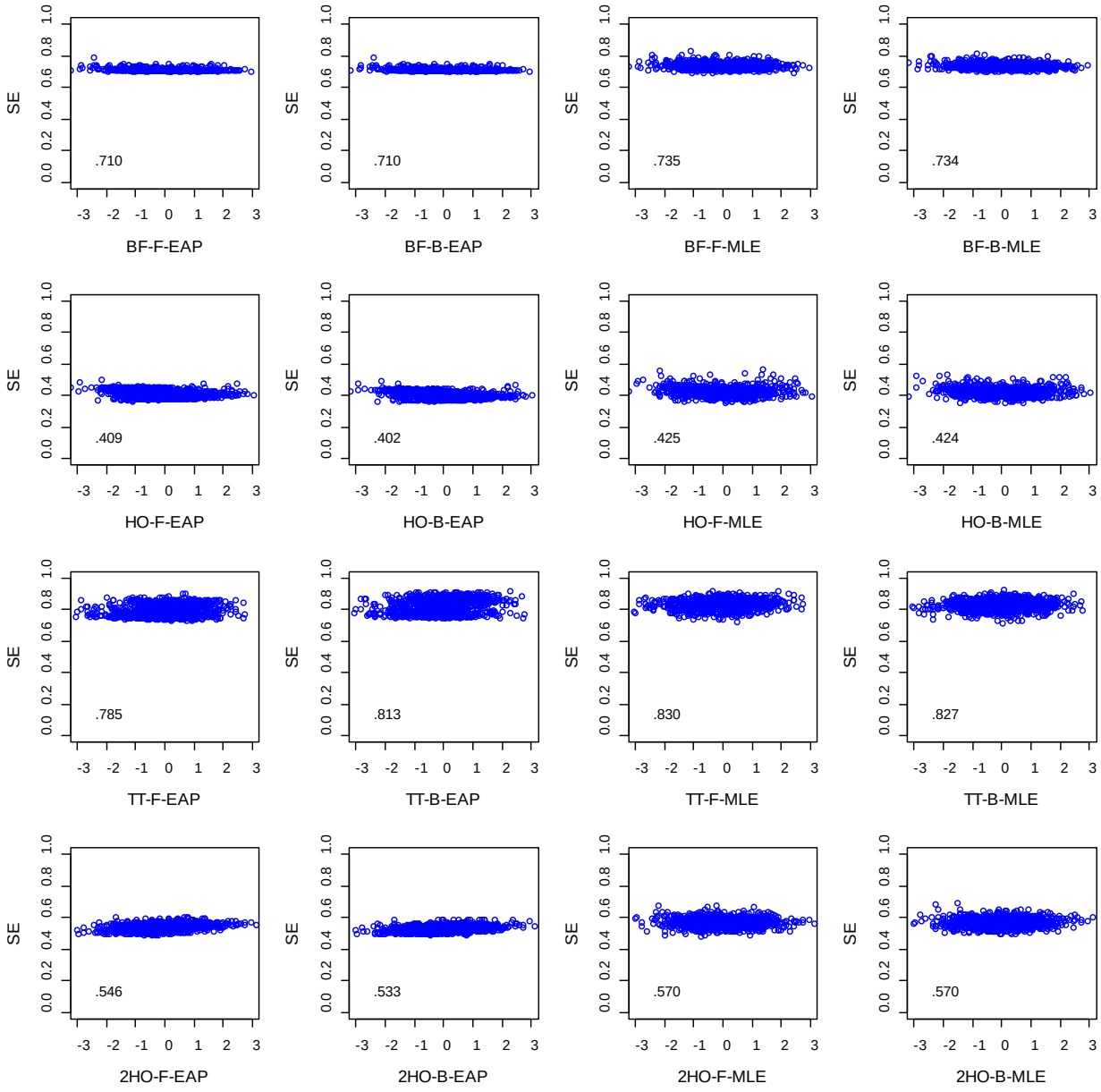
4.13: Average SE (First group factor with two group factors (40 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



4.14: Average SE (First group factor with two group factors (80 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)



4.15: Average SE (First group factor with two group factors (160 items))

Note. BF-F-EAP: Bifactor IRT model – Fisher item selection method – EAP scoring method
HO-B-MLE: Higher-order IRT model – Bayesian item selection method – MLE scoring method
(TT: Two-tier IRT model, 2HO: Higher-order IRT model with two general factors)

CHAPTER 5

Discussions

5.1 Summary

In recent years, the importance of formative assessments has been emphasized in the education field. This type of assessment often includes multiple correlated sub-domains and a hierarchical structure among the proficiencies. In this dissertation, several multidimensional CAT procedures were investigated to improve the measurement aspects of diagnostic testing and to better match the psychometric models to the test structure. The main purpose of this study was to explore various multidimensional CAT models on the measurement precision and accuracy and the impact of correlations among the dimensions under different conditions. To achieve the purpose of the study, two item selection methods (MFI and Bayesian) and two proficiency estimation methods (MLE and EAP) were utilized while four CAT designs (1) Bifactor IRT model, (2) Higher-order IRT model, (3) Two-tier IRT model, and (4) Higher-order IRT model with two primary factors were considered.

Five factors were manipulated to determine the effectiveness of the multidimensional CAT design: (1) the different correlation conditions between two primary factors (low, medium, and high), (2) the number of group factors per primary factor (two and four), (3) the number of items (40, 80 and 160), (4) the item selection method (MFI and Bayesian), and (5) the proficiency score estimation method (MLE and EAP). Three outcome measures, including correlations between true and estimated proficiency scores, Root Mean Square Error (RMSE) of

estimated proficiency scores, and Standard Errors (SE) were computed totaling 192 different conditions. A total of 1000 examinees were simulated for each test condition.

This study undertook a comprehensive comparison of item selection methods and proficiency scores estimation in several multidimensional IRT models in conjunction with a CAT. As expected, the correlation between true and estimated proficiency scores increased when the test length increased under different correlations among the factors, different item selection methods and different scoring methods. The higher-order IRT model algorithms provided higher correlations than the hierarchical IRT models for the primary factor, but the bifactor and the two-tier IRT model showed higher correlations than the higher-order IRT models for the group factors.

Average RMSEs decreased when the test length increased under different correlations among the factors, different item selection methods and different scoring methods. The higher-order IRT model algorithm provided lower RMSEs than the other three multidimensional IRT models for primary factor. However, there were no large differences in RMSEs for the group factors among the models. Smaller RMSEs for the primary factor were observed for the bifactor model and the higher-order model with four group factors than those for the bifactor model and the higher-order model with two group factors. In contrast, the RMSEs for the group factors from the bifactor model and the higher-order model with two group factors were slightly smaller than those from the bifactor model and the higher-order model with four group factors. For the generating conditions with two primary factors, the RMSEs in both of primary and group factors from the higher-order IRT model with two primary factors and the two-tier IRT model with two group factors were slightly smaller than those from the two models with four group factors.

Average SEs decreased as the number of items and correlation between two primary factors increased. The bifactor IRT model and the higher-order IRT model provided smaller SEs for both of primary and group factors than the two-tier and the higher-order IRT models with two primary factors. Smaller SEs for the primary factor were observed for the bifactor model and the higher-order model with four group factors than those for the bifactor model and the higher-order model with two group factors. In contrast, the SEs for the group factor from the bifactor model and the higher-order model with two group factors were slightly smaller than those from the bifactor model and the higher-order model with four group factors. For the generating conditions with two primary factors, the SEs from the two-tier IRT model and the higher-order IRT model with two primary factors and two group factors were slightly smaller than those from two models with four group factors.

In conclusion, the higher-order IRT model CAT has an advantage over the hierarchical IRT model CAT when we need scores for the primary factors. On the other hand, if test designers are interested in more specific group factors, hierarchical IRT models outperformed the higher-order IRT models. The bifactor model and the two-tier model have several advantages over the higher-order models. First, the hierarchical models fit the data significantly better than the corresponding higher-order models. Second, when group factors are used to predict an external variable, it is easier to interpret the results from the hierarchical models (Chen, West, & Sousa, 2006).

In this study, two different item selection methods, Maximum Fisher Information (MFI) and Bayesian, were applied to evaluate the measurement accuracy of item selection methods in CAT. The item selection method was negligible across the four multidimensional IRT CAT

algorithms. However, the Bayesian item selection method had smaller RMSEs and SEs than the MFI method in specific cases.

The effect of proficiency score estimation methods were negligible across the four multidimensional IRT CAT algorithms. However, the EAP proficiency score estimation method outperformed the MLE method, especially for short test length in this study. In general, Bayesian θ estimation such as EAP was recommended as the ability estimation methods. However, with the Bayesian method, the test developers or designers need to select priors, which might not be as objective as the maximum likelihood method. Thus, all factors need to be taken into considerations when choosing the θ estimation method. In general, if the test length is long enough, the two scoring methods should be comparable.

5.2 Limitations and Directions for Future Study

Several issues were not investigated in this study and should be explored in the future. Above all, there is a need to make the multidimensional IRT CAT model more widely applicable in various research settings in the educational and psychological field.

First, the simulation design was limited to two orders of factors: the first-order group factors (either two or four), and the second-order primary factors (either one or two). Despite the fact that third-order latent traits have not been studied, in practice they may be involved more than just first-order latent traits and second-order latent traits. These higher-order and hierarchical IRT CAT algorithms performed fairly well in these restricted simulation conditions. Similar performance can be expected under expanded conditions.

Second, more research needs to be done on how to utilize the collateral information for priors to better assist estimation with the Bayesian method. Instead of the population prior, as was used in this study, an individual prior may be used to increase the accuracy of the estimation. Thus, more studies need to be done to investigate individual prior cases like variances are quite different and correlations vary to assess the impact of item selection methods and proficiency score estimation methods under various conditions.

Third, a fixed-length rule (40, 80 and 160 items) was used to terminate the CAT algorithm in this study. The primary advantage of the fixed length stopping rule is its simplicity. However, one downside of the fixed length stopping rule is that examinees will be measured with different degrees of precision, with larger measurement errors typically occurring at extreme trait levels. In contrast to the fixed length stopping rule, one powerful advantage of the SE stopping rule is that, when the item pool information function is relatively flat, it typically yields near equivalent measurement precision across the examinee trait continuum (Choi, Grady, & Dodd, 2010). The implementation of fixed precision termination rules needs further evaluation in multidimensional IRT model CATs.

Fourth, only dichotomous items were examined in this study. However, in recent years, polytomous items have been widely used in educational and psychological tests. Fortunately, there are various IRT models such as the Graded Response Model (Samejima, 1969, 1972), the Nominal Model (Bock, 1972), and the Generalized Partial Credit Model (Muraki, 1992) for polytomous responses. Thus, the development of higher-order and hierarchical IRT CAT algorithms for polytomous items would be of interest.

Fifth, although there are many methods for item selection, ability estimation, item exposure control, and content balancing, this study investigates only a few of them. However, controlling item exposure is one of the important issues that must be addressed, since a high rate of item exposure leads to a large test security risk in high stake assessments (Davey & Parshall, 1995; Lee, Ip, & Fuh, 2008; Stocking & Lewis, 1995a; 1995b). With such high stakes, future research on multidimensional IRT model CAT should consider implementing item exposure control procedures.

Finally, in addition to CAT, computerized classification testing and multistage tests have received much attention in recent years (Davis & Dodd, 2003; Thompson, 2009; Wang & Huang, 2011). The development of computerized classification testing algorithms and multistage tests under the higher-order and hierarchical IRT models would be a valuable contribution to this field of study in the behavioral sciences.

5.3 Implications for Educational and Psychological Measurement

CAT has become a very important testing mode since it was introduced into the educational and psychological fields in the early 1970's. It has clear advantages over the traditional paper and pencil testing in many aspects including shorter tests and more efficient score reporting. As the movement towards CAT continues to move forward in large-scale educational assessments, understanding the properties of various adaptive testing designs becomes more important.

The purpose of this dissertation has been to comprehensively explore how diagnostic test designs could capitalize on the dimensional and hierarchical structure among the proficiencies

being measured and CAT to improve measurement precision and efficiency. To address the purpose of the study, simulation studies were conducted based on real data. Along with the simulation studies, a complete illustration of how adaptive testing procedures improved the potential benefits of diagnostic assessment. Also, this dissertation addressed the importance of the correlations of estimated proficiencies with different numbers of primary and group factors and various test lengths. By fully utilizing information from each test administration, (i.e., "borrowing strength" from the correlation structure of latent traits), the precision and reliability of the targeted subscale estimates was improved. In addition, approaches to CAT utilizing higher-order and hierarchical models in CAT approaches could enhance the validity and usefulness of a given test by providing diagnostic subscale estimates in addition to an overall scale estimate.

The test features investigated in this dissertation, such as test length, the number of dimensions, item selection methods, proficiency estimation methods, and the use of higher-order or hierarchical models are all issues and decision points that practitioners in the field face regularly. Thus, the findings from this dissertation contribute to the expanding knowledge base in the fields of educational and psychological research and provide practical guidelines to programs that are considering CAT as a test design.

APPENDIX A

Correlation between True and Estimated Proficiency Scores

Table A.1: Correlation between True and Estimated Proficiency Scores (Bifactor IRT Model)

Test Length	Item Selection	Scoring	Number of group factors							
			Two group factors			Four group factors				
			G_1	s_1	s_2	G_1	s_1	s_2	s_3	s_4
40										
	MFI	MLE	0.848	0.841	0.843	0.844	0.852	0.849	0.857	0.855
		EAP	0.858	0.842	0.853	0.861	0.858	0.856	0.858	0.857
	Bayes	MLE	0.847	0.839	0.836	0.852	0.854	0.847	0.859	0.850
		EAP	0.859	0.847	0.850	0.863	0.858	0.850	0.858	0.852
80										
	MFI	MLE	0.871	0.833	0.836	0.858	0.858	0.860	0.851	0.860
		EAP	0.884	0.852	0.862	0.867	0.864	0.850	0.844	0.866
	Bayes	MLE	0.877	0.842	0.831	0.857	0.863	0.854	0.851	0.862
		EAP	0.884	0.857	0.867	0.871	0.864	0.865	0.844	0.864
160										
	MFI	MLE	0.896	0.864	0.867	0.886	0.859	0.858	0.850	0.851
		EAP	0.904	0.868	0.875	0.906	0.859	0.861	0.858	0.864
	Bayes	MLE	0.899	0.859	0.855	0.907	0.858	0.861	0.854	0.853
		EAP	0.907	0.871	0.877	0.916	0.859	0.865	0.862	0.864

Table A.2: Correlation between True and Estimated Proficiency Scores (Higher-order IRT Model)

Test Length	Item Selection	Scoring	Number of group factors							
			Two group factors			Four group factors				
			G_1	s_1	s_2	G_1	s_1	s_2	s_3	s_4
40										
	MFI	MLE	0.861	0.833	0.835	0.875	0.835	0.842	0.846	0.834
		EAP	0.871	0.836	0.842	0.878	0.840	0.852	0.851	0.841
	Bayes	MLE	0.866	0.832	0.846	0.877	0.845	0.849	0.862	0.838
		EAP	0.878	0.839	0.846	0.870	0.858	0.850	0.853	0.845
80										
	MFI	MLE	0.888	0.865	0.863	0.876	0.845	0.841	0.856	0.842
		EAP	0.886	0.865	0.866	0.888	0.840	0.852	0.851	0.841
	Bayes	MLE	0.882	0.871	0.866	0.879	0.844	0.831	0.854	0.837
		EAP	0.891	0.869	0.867	0.880	0.858	0.850	0.853	0.845
160										
	MFI	MLE	0.915	0.874	0.868	0.920	0.840	0.846	0.850	0.855
		EAP	0.911	0.874	0.877	0.927	0.873	0.865	0.868	0.856
	Bayes	MLE	0.916	0.881	0.870	0.926	0.850	0.854	0.863	0.859
		EAP	0.926	0.877	0.870	0.923	0.879	0.872	0.870	0.859

Table A.3: Correlation between True and Estimated Proficiency Scores (Two-tier IRT Model with two group factors)

Test Length	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
	Low correlation b/t two general factors							
	MFI	MLE	0.787	0.800	0.874	0.869	0.897	0.901
		EAP	0.803	0.822	0.842	0.836	0.853	0.896
	Bayes	MLE	0.794	0.792	0.872	0.868	0.910	0.905
		EAP	0.813	0.820	0.843	0.836	0.857	0.895
40	*Medium correlation b/t two general factors*							
	MFI	MLE	0.803	0.804	0.865	0.878	0.900	0.899
		EAP	0.803	0.814	0.842	0.855	0.870	0.894
	Bayes	MLE	0.792	0.803	0.857	0.872	0.910	0.901
		EAP	0.805	0.812	0.853	0.849	0.881	0.897
	High correlation b/t two general factors							
	MFI	MLE	0.802	0.807	0.875	0.879	0.896	0.898
		EAP	0.816	0.818	0.849	0.847	0.857	0.887
	Bayes	MLE	0.808	0.815	0.878	0.882	0.897	0.901
		EAP	0.834	0.824	0.859	0.853	0.879	0.891
	Low correlation b/t two general factors							
	MFI	MLE	0.840	0.848	0.854	0.856	0.860	0.856
		EAP	0.849	0.856	0.844	0.832	0.843	0.851
	Bayes	MLE	0.848	0.853	0.857	0.841	0.858	0.854
		EAP	0.848	0.857	0.844	0.829	0.848	0.853
80	*Medium correlation b/t two general factors*							
	MFI	MLE	0.857	0.872	0.848	0.856	0.862	0.861
		EAP	0.855	0.866	0.829	0.845	0.854	0.858
	Bayes	MLE	0.852	0.868	0.842	0.854	0.866	0.857
		EAP	0.859	0.872	0.834	0.848	0.856	0.866
	High correlation b/t two general factors							
	MFI	MLE	0.868	0.868	0.839	0.850	0.867	0.848
		EAP	0.869	0.866	0.841	0.843	0.852	0.846
	Bayes	MLE	0.876	0.860	0.854	0.859	0.865	0.866
		EAP	0.879	0.876	0.836	0.840	0.850	0.858
	Low correlation b/t two general factors							
	MFI	MLE	0.881	0.894	0.838	0.833	0.844	0.847
		EAP	0.898	0.901	0.846	0.847	0.857	0.853
	Bayes	MLE	0.885	0.894	0.839	0.837	0.839	0.857
		EAP	0.900	0.905	0.847	0.847	0.856	0.853
160	*Medium correlation b/t two general factors*							
	MFI	MLE	0.894	0.899	0.839	0.839	0.859	0.858
		EAP	0.904	0.902	0.846	0.844	0.854	0.866
	Bayes	MLE	0.904	0.905	0.846	0.847	0.859	0.868
		EAP	0.903	0.909	0.845	0.848	0.855	0.864
	High correlation b/t two general factors							
	MFI	MLE	0.893	0.887	0.828	0.855	0.824	0.848
		EAP	0.904	0.905	0.856	0.852	0.856	0.847
	Bayes	MLE	0.903	0.911	0.840	0.849	0.846	0.847
		EAP	0.911	0.910	0.837	0.850	0.851	0.842

Table A.4: Correlation between True and Estimated Proficiency Scores (Two-tier IRT Model with four group factors)

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
40												
Low correlation b/t two general factors												
	MFI	MLE	0.732	0.743	0.874	0.861	0.901	0.907	0.871	0.903	0.889	0.901
		EAP	0.746	0.740	0.889	0.884	0.912	0.861	0.847	0.894	0.913	0.845
	Bayes	MLE	0.739	0.744	0.876	0.873	0.909	0.908	0.865	0.901	0.919	0.904
		EAP	0.750	0.743	0.875	0.878	0.886	0.857	0.857	0.898	0.912	0.850
Medium correlation b/t two general factors												
	MFI	MLE	0.740	0.731	0.882	0.912	0.904	0.910	0.888	0.894	0.918	0.906
		EAP	0.754	0.754	0.879	0.878	0.910	0.876	0.863	0.883	0.919	0.834
	Bayes	MLE	0.746	0.749	0.888	0.910	0.899	0.914	0.882	0.901	0.915	0.907
		EAP	0.759	0.753	0.879	0.875	0.888	0.868	0.884	0.898	0.930	0.842
High correlation b/t two general factors												
	MFI	MLE	0.791	0.780	0.904	0.900	0.909	0.902	0.898	0.908	0.918	0.905
		EAP	0.799	0.803	0.910	0.876	0.915	0.846	0.842	0.890	0.927	0.844
	Bayes	MLE	0.799	0.809	0.904	0.906	0.910	0.896	0.895	0.909	0.915	0.905
		EAP	0.804	0.811	0.899	0.905	0.907	0.864	0.895	0.905	0.921	0.848
80												
Low correlation b/t two general factors												
	MFI	MLE	0.809	0.815	0.890	0.888	0.931	0.924	0.893	0.924	0.941	0.919
		EAP	0.811	0.809	0.888	0.873	0.929	0.880	0.867	0.914	0.953	0.865
	Bayes	MLE	0.812	0.807	0.888	0.891	0.927	0.928	0.904	0.925	0.942	0.921
		EAP	0.819	0.813	0.897	0.873	0.903	0.878	0.877	0.919	0.952	0.870
Medium correlation b/t two general factors												
	MFI	MLE	0.814	0.810	0.886	0.890	0.925	0.932	0.899	0.929	0.938	0.917
		EAP	0.808	0.811	0.885	0.865	0.926	0.895	0.873	0.903	0.949	0.853
	Bayes	MLE	0.819	0.814	0.887	0.888	0.920	0.932	0.900	0.923	0.939	0.923
		EAP	0.812	0.819	0.901	0.867	0.906	0.888	0.903	0.918	0.950	0.862
High correlation b/t two general factors												
	MFI	MLE	0.819	0.831	0.892	0.882	0.937	0.924	0.906	0.921	0.941	0.919
		EAP	0.835	0.829	0.904	0.858	0.934	0.860	0.881	0.910	0.947	0.864
	Bayes	MLE	0.831	0.837	0.902	0.895	0.933	0.923	0.908	0.926	0.937	0.918
		EAP	0.844	0.843	0.901	0.905	0.916	0.882	0.905	0.925	0.930	0.868

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
160												
Low correlation b/t two general factors												
	MFI	MLE	0.873	0.840	0.900	0.881	0.891	0.904	0.880	0.902	0.940	0.905
		EAP	0.883	0.843	0.886	0.878	0.864	0.851	0.854	0.883	0.934	0.848
	Bayes	MLE	0.877	0.846	0.878	0.901	0.887	0.886	0.898	0.903	0.934	0.907
		EAP	0.882	0.839	0.886	0.879	0.861	0.850	0.874	0.898	0.935	0.852
Medium correlation b/t two general factors												
	MFI	MLE	0.883	0.848	0.888	0.889	0.890	0.896	0.900	0.902	0.926	0.914
		EAP	0.882	0.853	0.878	0.884	0.867	0.877	0.874	0.888	0.924	0.860
	Bayes	MLE	0.876	0.853	0.902	0.903	0.887	0.900	0.901	0.905	0.924	0.909
		EAP	0.885	0.851	0.888	0.890	0.873	0.875	0.880	0.897	0.929	0.861
High correlation b/t two general factors												
	MFI	MLE	0.885	0.891	0.904	0.897	0.898	0.875	0.909	0.913	0.929	0.907
		EAP	0.896	0.881	0.896	0.903	0.905	0.879	0.885	0.894	0.933	0.853
	Bayes	MLE	0.884	0.876	0.904	0.891	0.990	0.885	0.905	0.911	0.937	0.909
		EAP	0.896	0.880	0.901	0.904	0.911	0.901	0.917	0.912	0.937	0.881

Table A.5: Correlation between True and Estimated Proficiency Scores (Higher-order IRT Model (2 primary factors) with two group factors)

Test Length	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
40	*Low correlation b/t two general factors*							
	MFI	MLE	0.795	0.799	0.825	0.881	0.820	0.890
		EAP	0.811	0.787	0.832	0.866	0.814	0.873
	Bayes	MLE	0.798	0.793	0.844	0.881	0.825	0.904
		EAP	0.797	0.781	0.843	0.874	0.815	0.867
	Medium correlation b/t two general factors							
	MFI	MLE	0.824	0.816	0.844	0.891	0.819	0.898
		EAP	0.836	0.813	0.838	0.858	0.827	0.871
	Bayes	MLE	0.837	0.810	0.847	0.888	0.823	0.899
		EAP	0.837	0.802	0.845	0.867	0.821	0.872
	High correlation b/t two general factors							
	MFI	MLE	0.850	0.822	0.827	0.884	0.836	0.897
		EAP	0.853	0.838	0.829	0.859	0.832	0.880
	Bayes	MLE	0.845	0.829	0.834	0.889	0.835	0.904
		EAP	0.856	0.830	0.829	0.861	0.828	0.877
80	*Low correlation b/t two general factors*							
	MFI	MLE	0.855	0.889	0.836	0.870	0.815	0.868
		EAP	0.859	0.874	0.825	0.859	0.827	0.843
	Bayes	MLE	0.858	0.890	0.823	0.864	0.790	0.859
		EAP	0.854	0.878	0.843	0.859	0.832	0.843
	Medium correlation b/t two general factors							
	MFI	MLE	0.878	0.889	0.837	0.853	0.822	0.865
		EAP	0.881	0.896	0.839	0.846	0.833	0.852
	Bayes	MLE	0.886	0.893	0.821	0.866	0.819	0.866
		EAP	0.885	0.892	0.843	0.853	0.836	0.852
	High correlation b/t two general factors							
	MFI	MLE	0.877	0.909	0.815	0.870	0.815	0.860
		EAP	0.895	0.908	0.831	0.865	0.837	0.861
	Bayes	MLE	0.890	0.907	0.819	0.875	0.813	0.874
		EAP	0.899	0.908	0.835	0.858	0.841	0.864
160	*Low correlation b/t two general factors*							
	MFI	MLE	0.900	0.908	0.839	0.861	0.823	0.853
		EAP	0.911	0.918	0.845	0.868	0.841	0.840
	Bayes	MLE	0.907	0.916	0.826	0.860	0.821	0.844
		EAP	0.904	0.912	0.859	0.859	0.857	0.845
	Medium correlation b/t two general factors							
	MFI	MLE	0.915	0.926	0.841	0.853	0.833	0.861
		EAP	0.919	0.930	0.853	0.849	0.849	0.842
	Bayes	MLE	0.909	0.924	0.834	0.842	0.835	0.841
		EAP	0.917	0.925	0.865	0.853	0.855	0.839
	High correlation b/t two general factors							
	MFI	MLE	0.923	0.940	0.855	0.853	0.848	0.889
		EAP	0.928	0.942	0.857	0.884	0.859	0.878
	Bayes	MLE	0.923	0.942	0.850	0.863	0.857	0.887
		EAP	0.932	0.939	0.869	0.872	0.864	0.886

Table A.6: Correlation between True and Estimated Proficiency Scores (Higher-order IRT Model (2 primary factors) with four group factors)

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
40												
Low correlation b/t two general factors												
	MFI	MLE	0.743	0.754	0.822	0.819	0.829	0.815	0.819	0.821	0.837	0.829
		EAP	0.757	0.751	0.837	0.832	0.830	0.799	0.795	0.842	0.831	0.799
	Bayes	MLE	0.750	0.755	0.828	0.825	0.841	0.810	0.813	0.849	0.847	0.822
		EAP	0.761	0.754	0.823	0.826	0.834	0.805	0.805	0.846	0.845	0.798
Medium correlation b/t two general factors												
	MFI	MLE	0.751	0.742	0.819	0.860	0.852	0.818	0.836	0.842	0.866	0.854
		EAP	0.765	0.765	0.827	0.826	0.851	0.817	0.824	0.844	0.867	0.782
	Bayes	MLE	0.757	0.760	0.820	0.858	0.857	0.822	0.830	0.849	0.863	0.855
		EAP	0.770	0.764	0.827	0.823	0.856	0.816	0.832	0.846	0.878	0.790
High correlation b/t two general factors												
	MFI	MLE	0.802	0.791	0.852	0.848	0.857	0.830	0.846	0.856	0.866	0.853
		EAP	0.801	0.814	0.858	0.824	0.853	0.814	0.790	0.838	0.875	0.802
	Bayes	MLE	0.810	0.820	0.854	0.856	0.850	0.826	0.845	0.859	0.863	0.813
		EAP	0.815	0.821	0.847	0.853	0.855	0.842	0.843	0.853	0.869	0.796
80												
Low correlation b/t two general factors												
	MFI	MLE	0.820	0.826	0.838	0.836	0.879	0.852	0.841	0.872	0.889	0.867
		EAP	0.832	0.820	0.836	0.821	0.877	0.838	0.845	0.862	0.901	0.813
	Bayes	MLE	0.823	0.812	0.836	0.839	0.875	0.846	0.852	0.873	0.890	0.869
		EAP	0.830	0.824	0.845	0.821	0.851	0.846	0.825	0.867	0.900	0.818
Medium correlation b/t two general factors												
	MFI	MLE	0.825	0.821	0.834	0.838	0.873	0.860	0.847	0.877	0.886	0.865
		EAP	0.819	0.822	0.840	0.813	0.874	0.853	0.821	0.851	0.897	0.801
	Bayes	MLE	0.830	0.825	0.835	0.836	0.868	0.861	0.848	0.871	0.887	0.871
		EAP	0.823	0.830	0.849	0.815	0.854	0.846	0.851	0.866	0.898	0.810
High correlation b/t two general factors												
	MFI	MLE	0.828	0.842	0.840	0.830	0.875	0.862	0.854	0.869	0.889	0.867
		EAP	0.846	0.840	0.852	0.806	0.879	0.840	0.831	0.860	0.897	0.812
	Bayes	MLE	0.842	0.838	0.850	0.843	0.881	0.851	0.856	0.874	0.885	0.866
		EAP	0.855	0.854	0.849	0.853	0.864	0.830	0.862	0.882	0.887	0.825

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
160												
Low correlation b/t two general factors												
	MFI	MLE	0.884	0.851	0.848	0.829	0.839	0.852	0.828	0.850	0.888	0.853
		EAP	0.894	0.854	0.834	0.826	0.812	0.799	0.802	0.831	0.882	0.796
	Bayes	MLE	0.888	0.827	0.830	0.853	0.839	0.838	0.846	0.851	0.882	0.855
		EAP	0.893	0.850	0.834	0.827	0.809	0.798	0.822	0.846	0.883	0.800
Medium correlation b/t two general factors												
	MFI	MLE	0.894	0.859	0.836	0.837	0.838	0.844	0.848	0.850	0.874	0.862
		EAP	0.893	0.864	0.826	0.832	0.828	0.838	0.835	0.849	0.872	0.808
	Bayes	MLE	0.887	0.864	0.850	0.851	0.835	0.848	0.849	0.853	0.872	0.857
		EAP	0.896	0.862	0.836	0.838	0.821	0.823	0.828	0.845	0.877	0.809
High correlation b/t two general factors												
	MFI	MLE	0.896	0.902	0.852	0.845	0.866	0.823	0.857	0.861	0.877	0.855
		EAP	0.907	0.892	0.844	0.851	0.873	0.827	0.833	0.842	0.881	0.801
	Bayes	MLE	0.895	0.887	0.854	0.841	0.910	0.835	0.855	0.861	0.885	0.857
		EAP	0.907	0.891	0.849	0.852	0.889	0.849	0.865	0.860	0.885	0.829

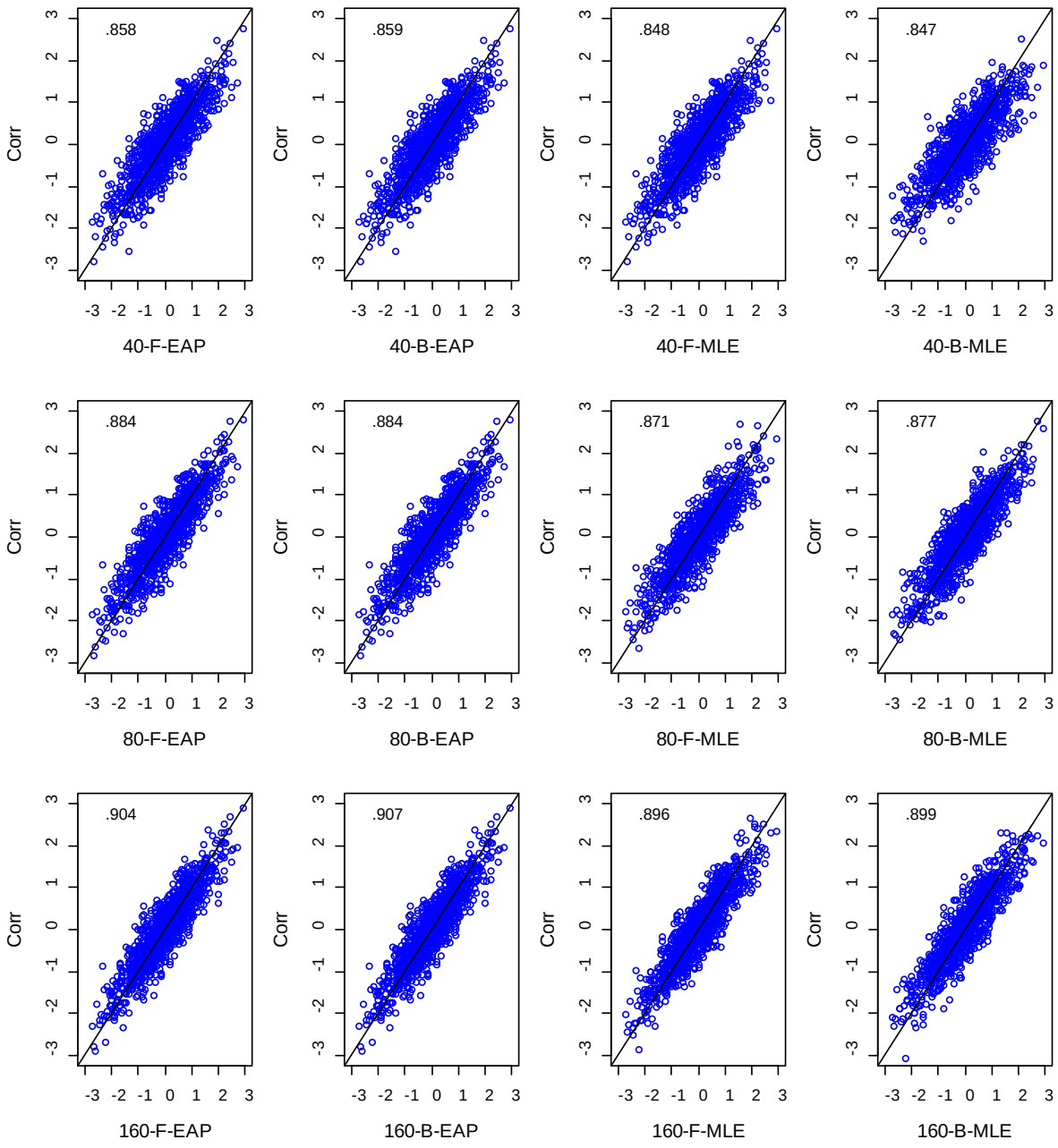


Figure A.1: Correlation between True and Estimated Proficiency Scores (Primary factor for Bifactor IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

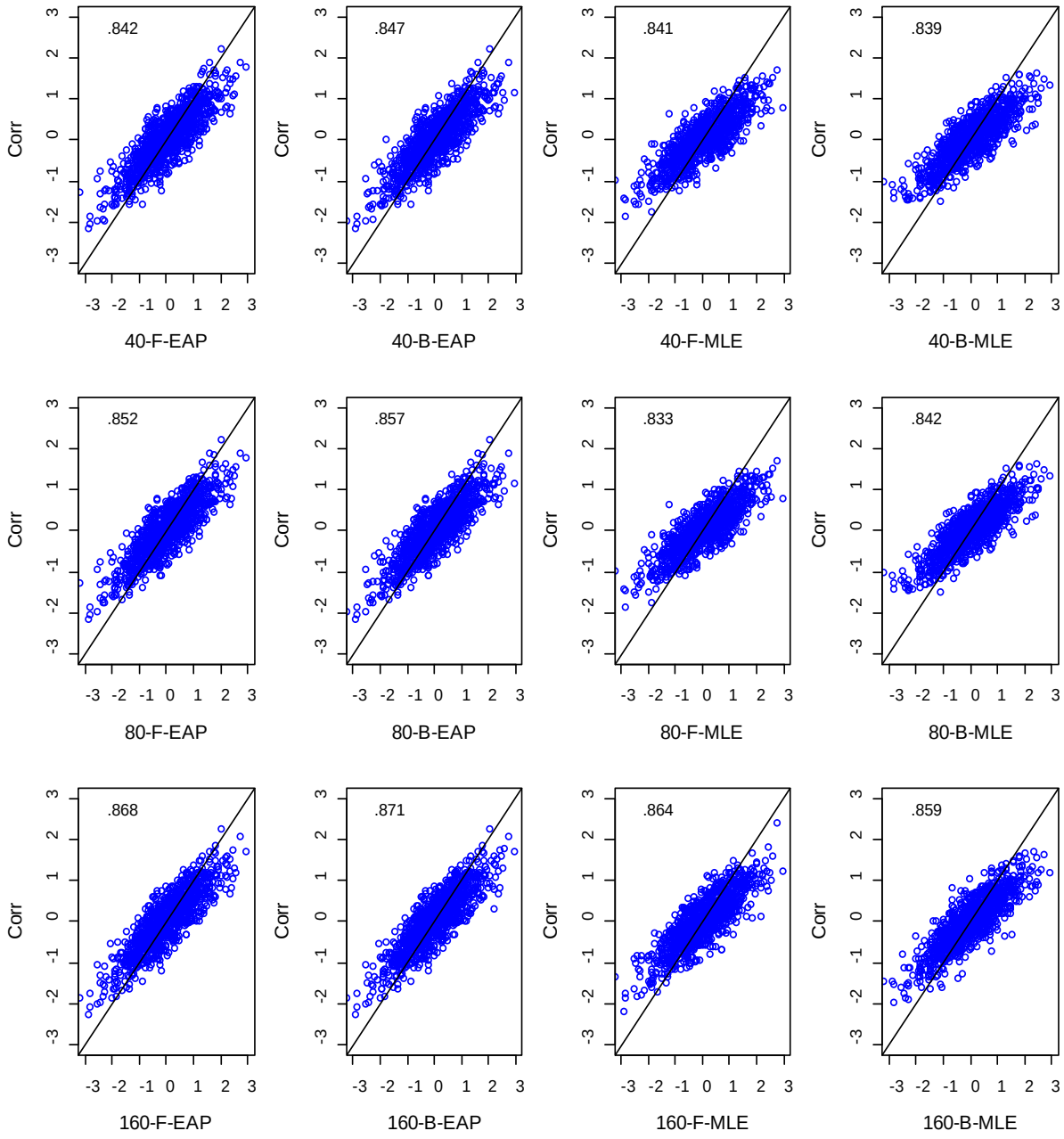


Figure A.2: Correlation between True and Estimated Proficiency Scores (First group factor for Bifactor IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

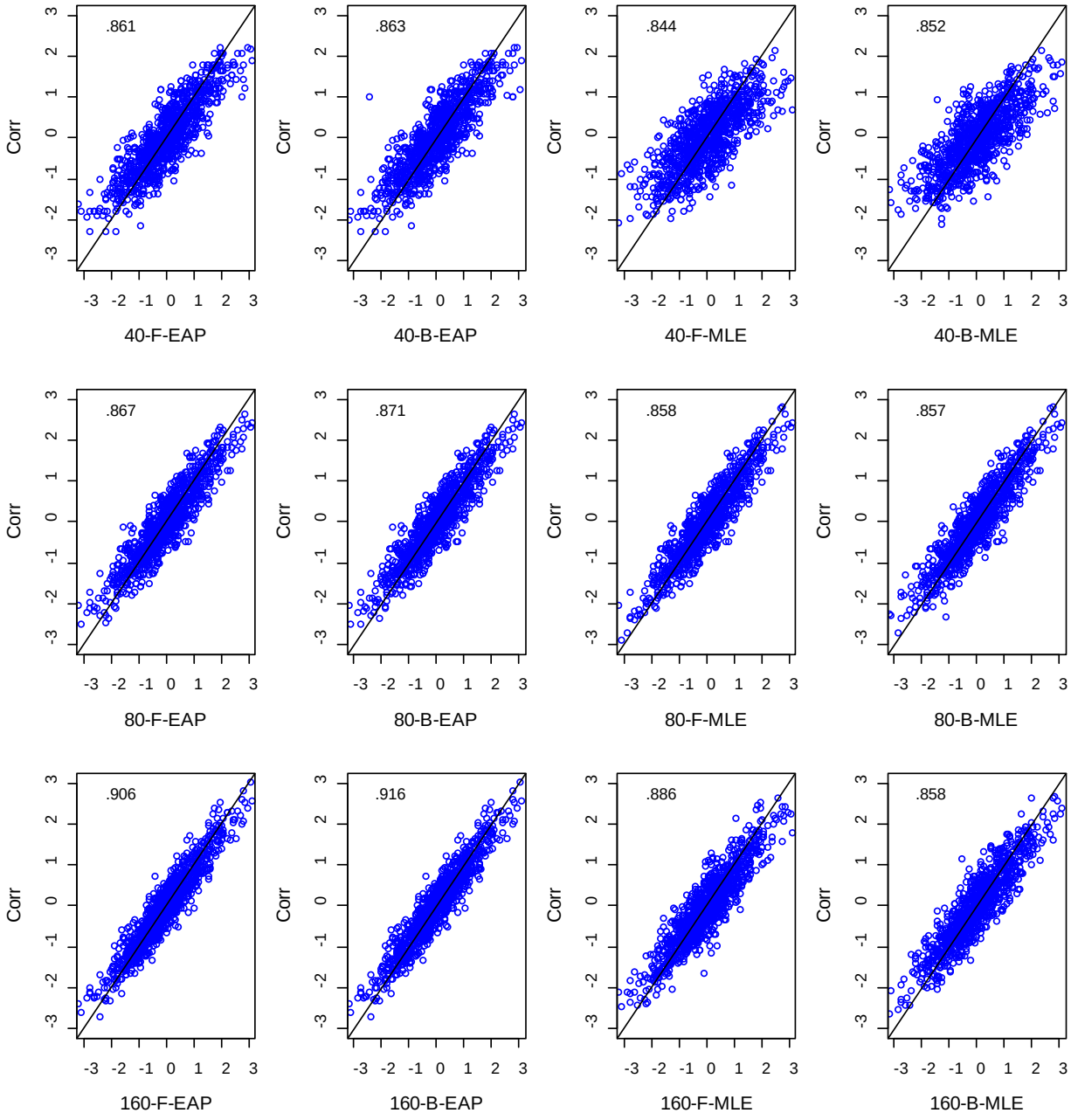


Figure A.3: Correlation between True and Estimated Proficiency Scores (Primary factor for Bifactor IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

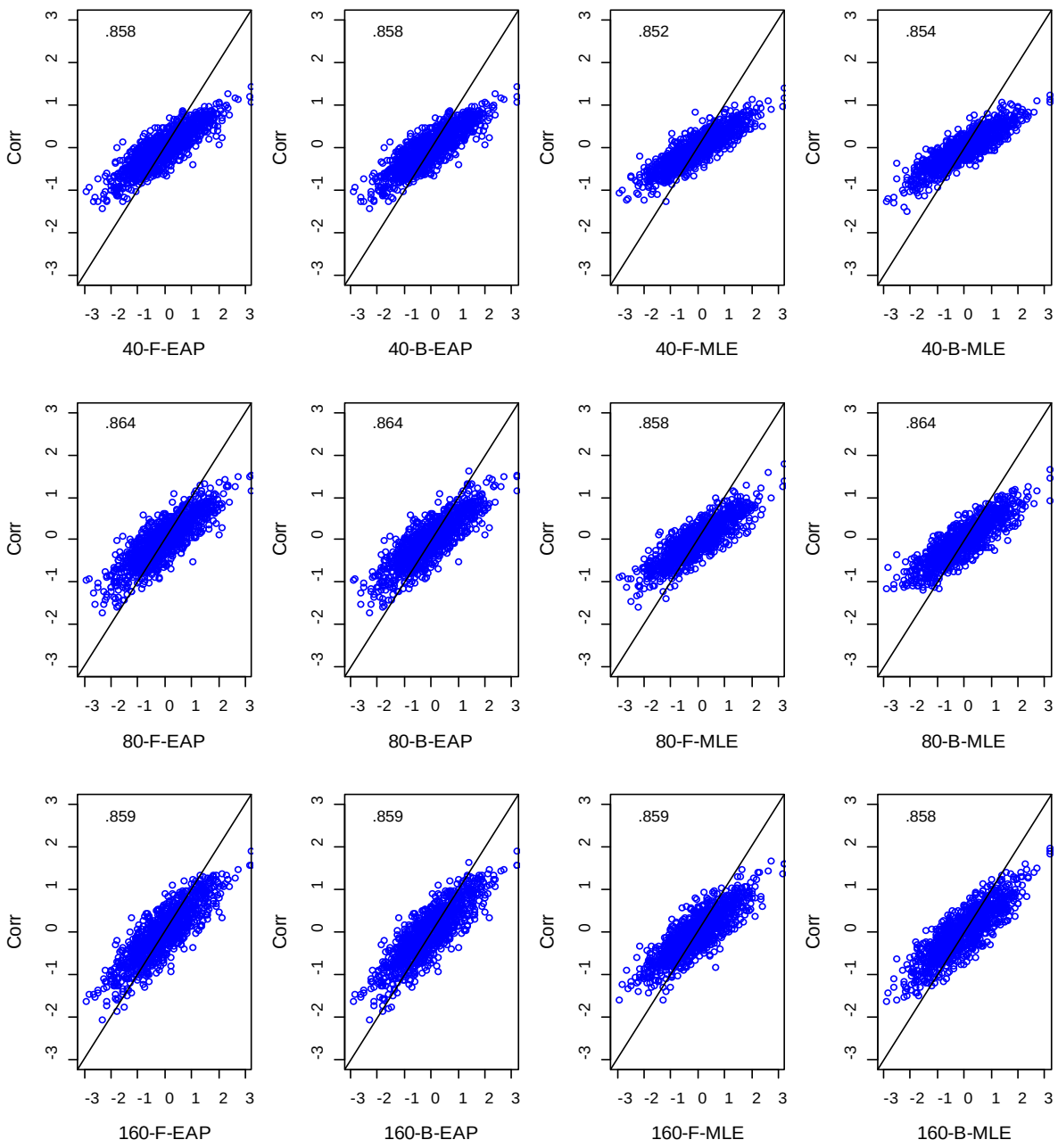


Figure A.4: Correlation between True and Estimated Proficiency Scores (First group factor for Bifactor IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

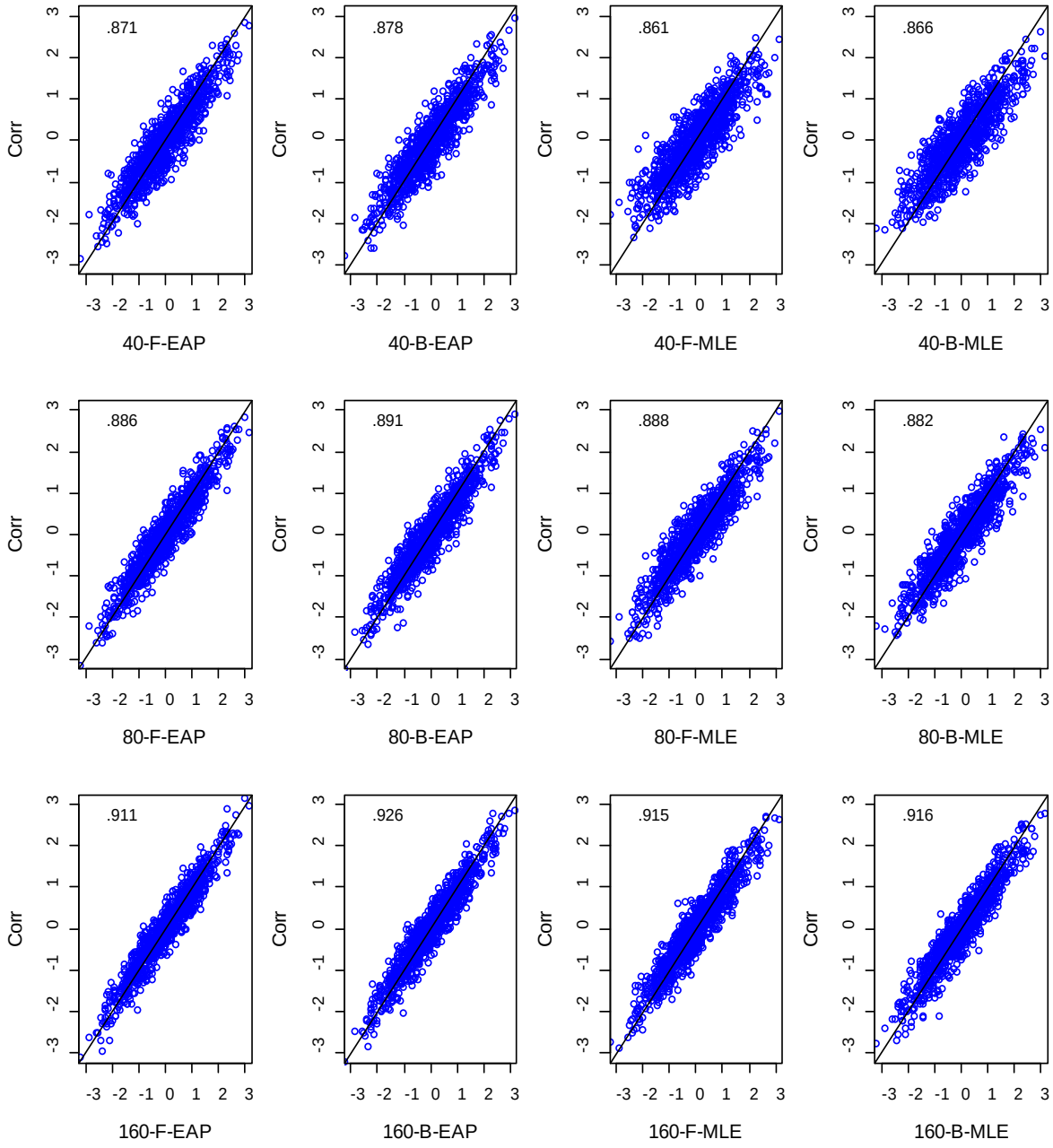


Figure A.5: Correlation between True and Estimated Proficiency Scores (Primary factor for Higher-order IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

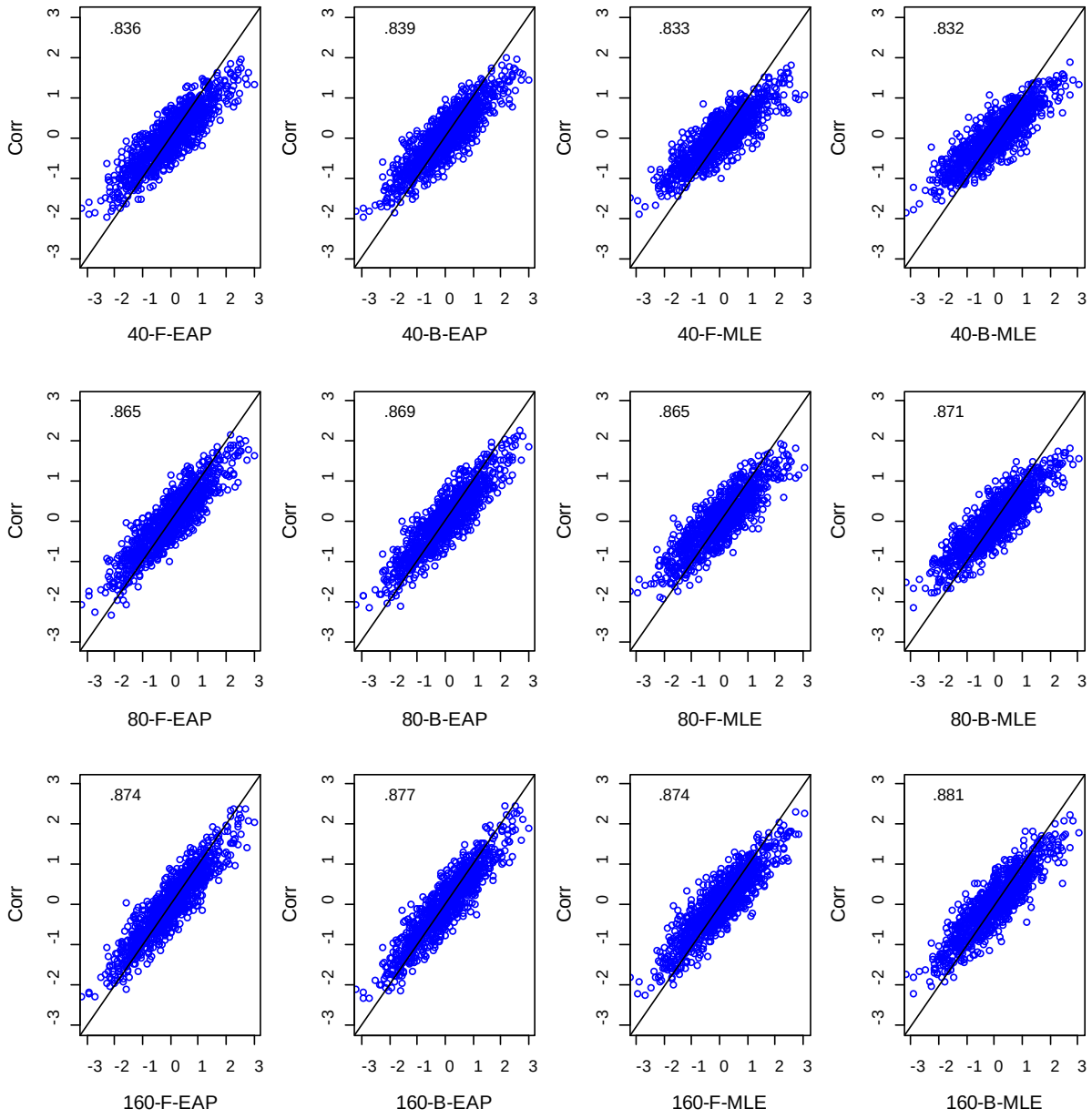


Figure A.6: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

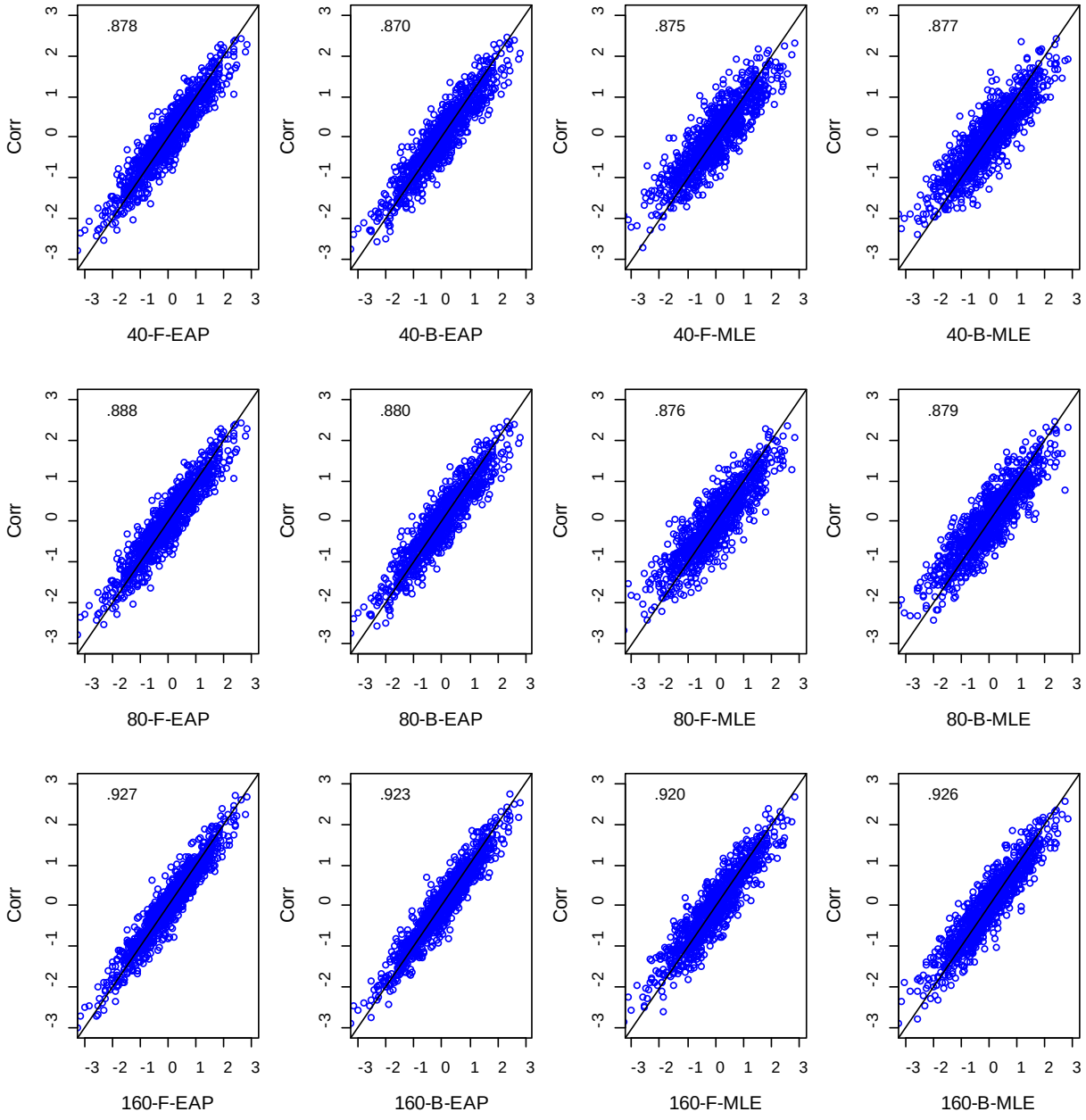


Figure A.7: Correlation between True and Estimated Proficiency Scores (Primary factor for Higher-order IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

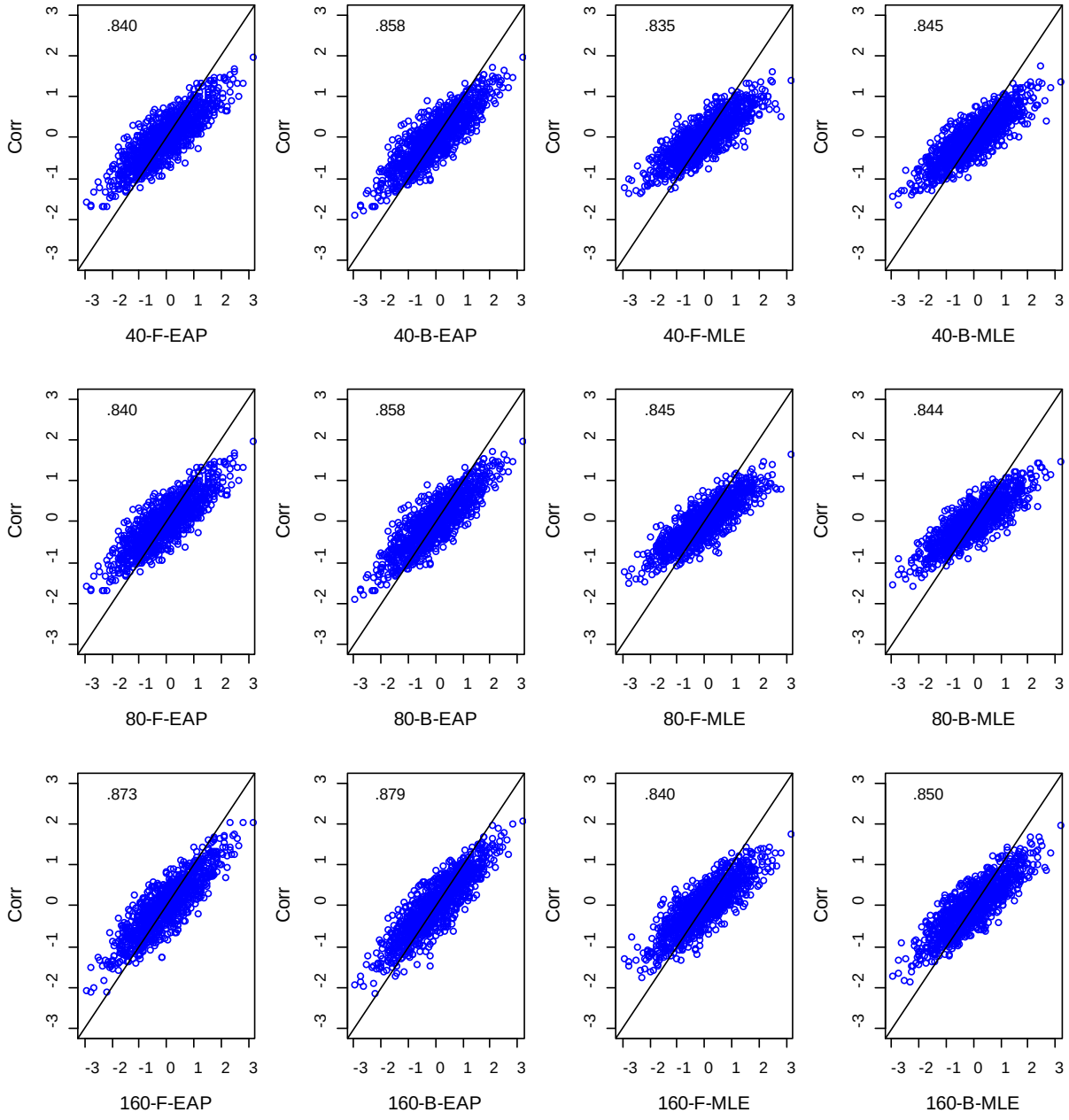


Figure A.8: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

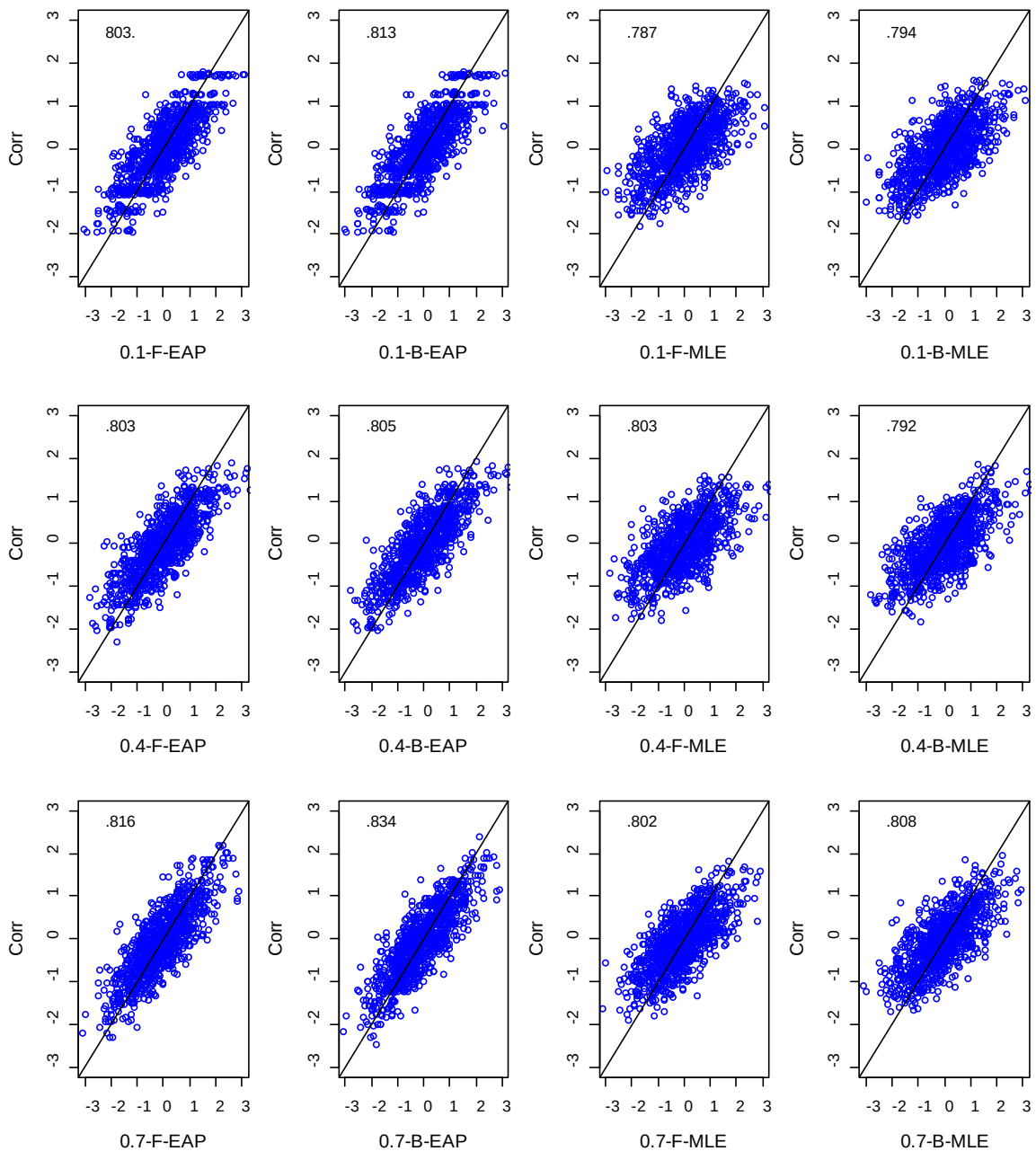


Figure A.9: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

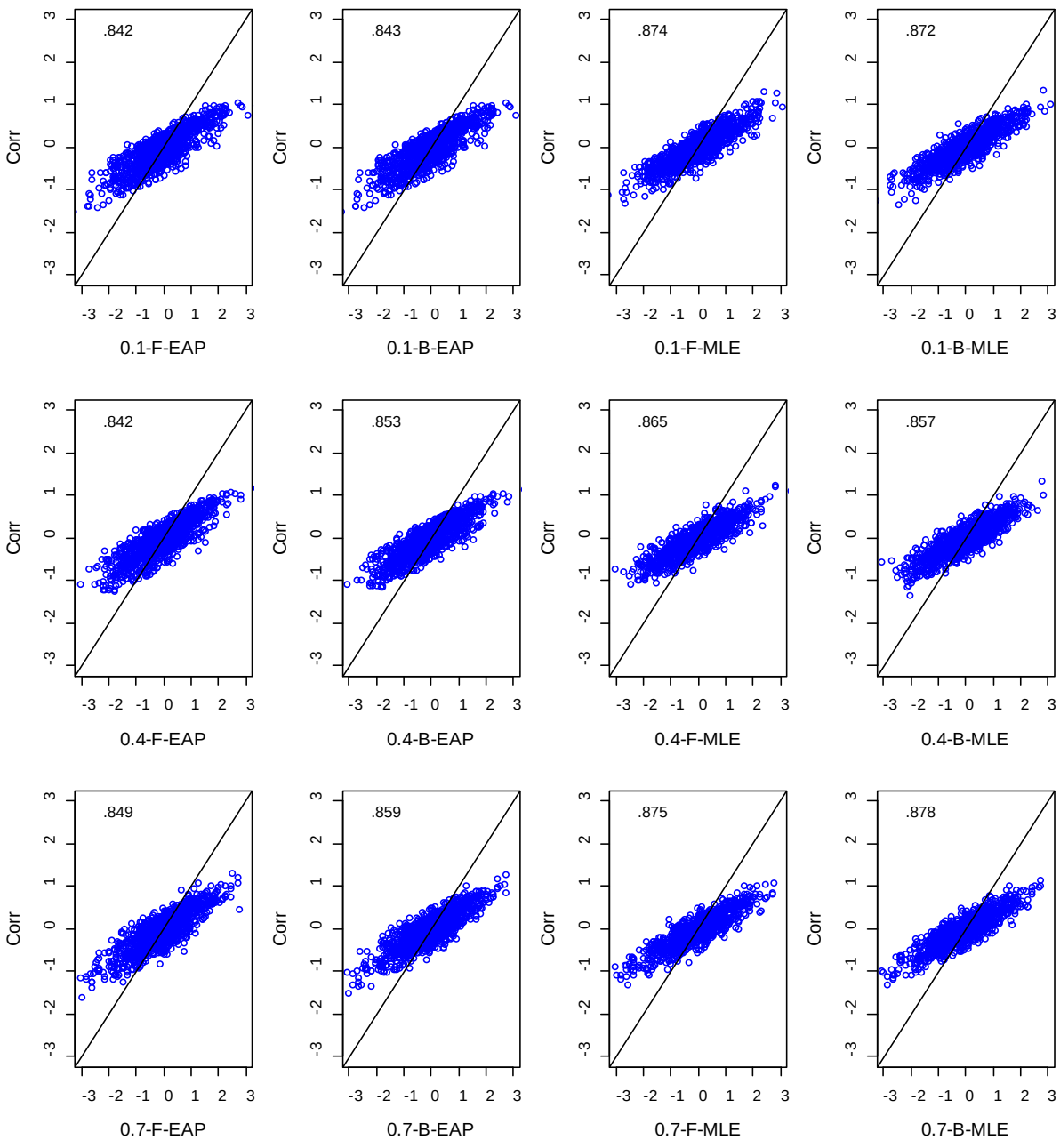


Figure A.10: Correlation between True and Estimated Proficiency Scores (First group factor for Two-tier IRT model with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

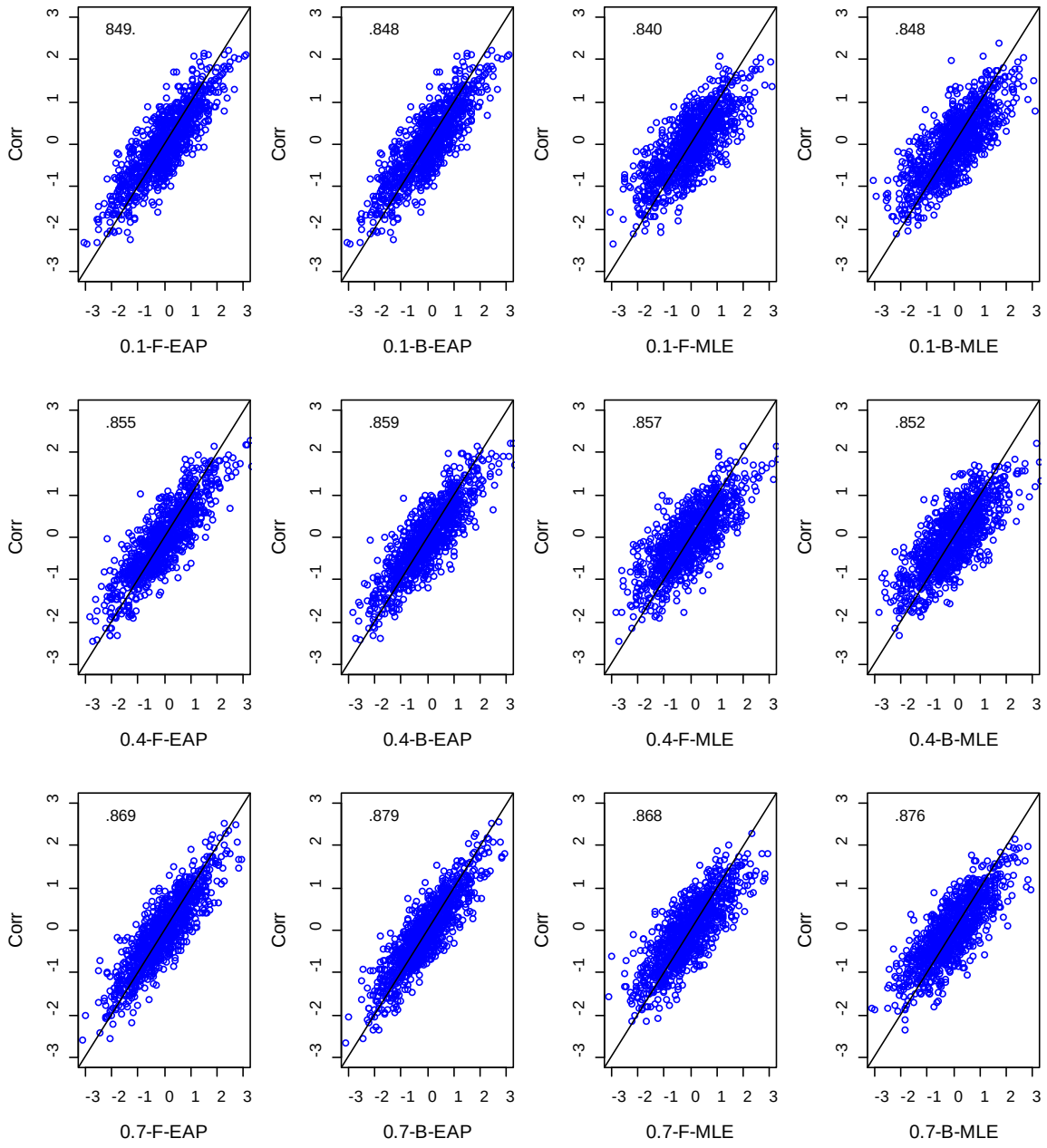


Figure A.11: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

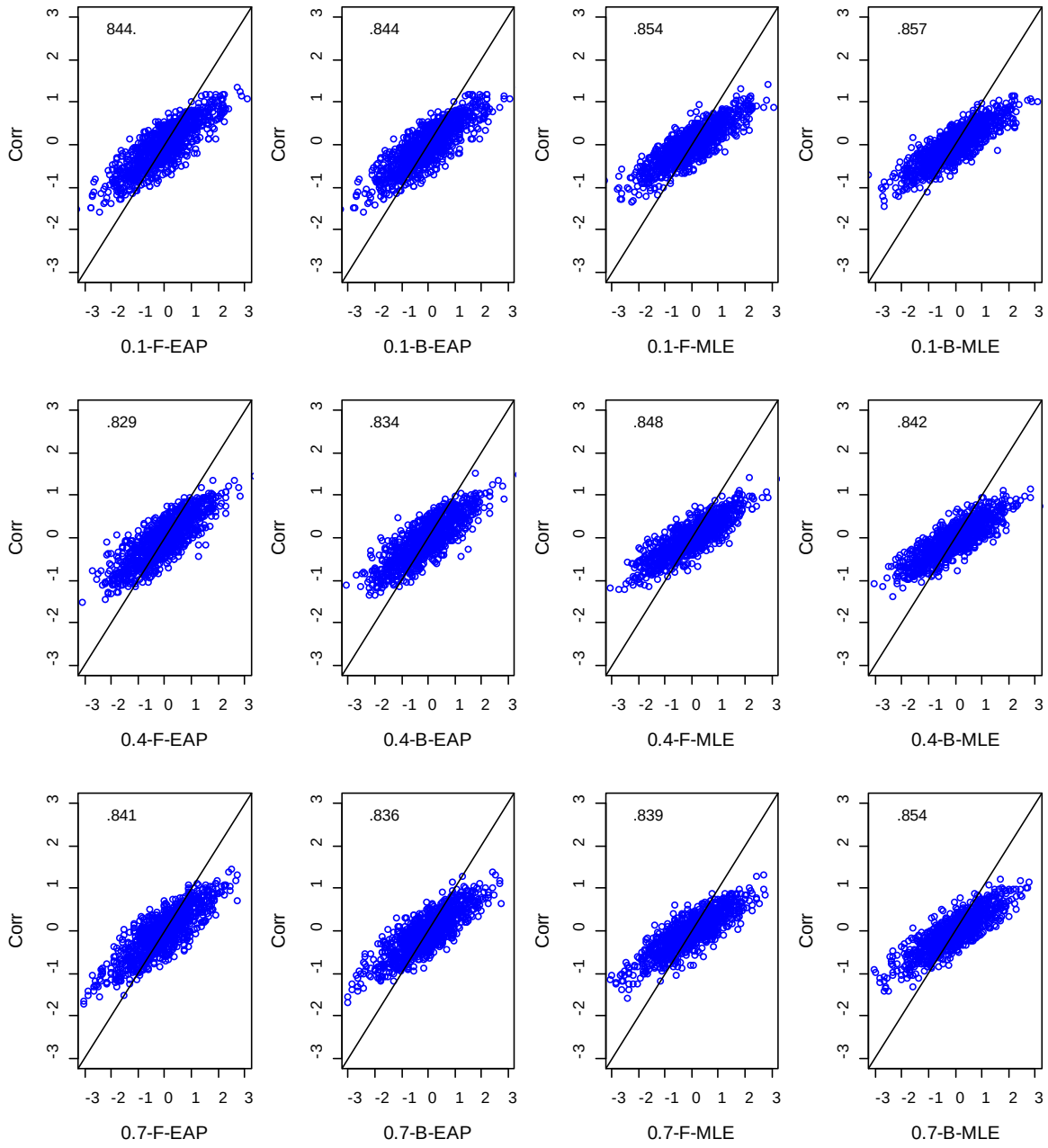


Figure A.12: Correlation between True and Estimated Proficiency Scores (First group factor for Two-tier IRT model with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

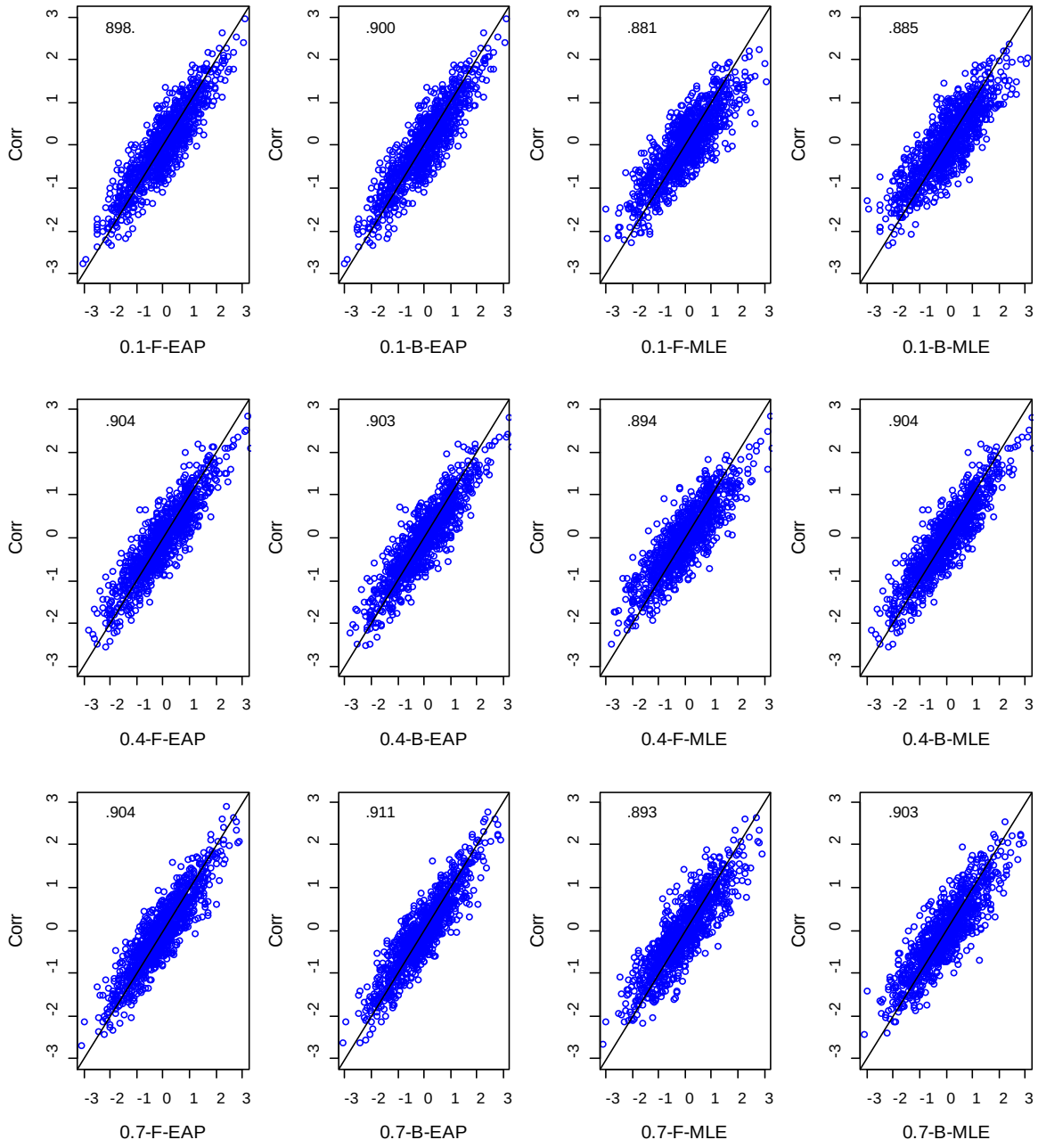


Figure A.13: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

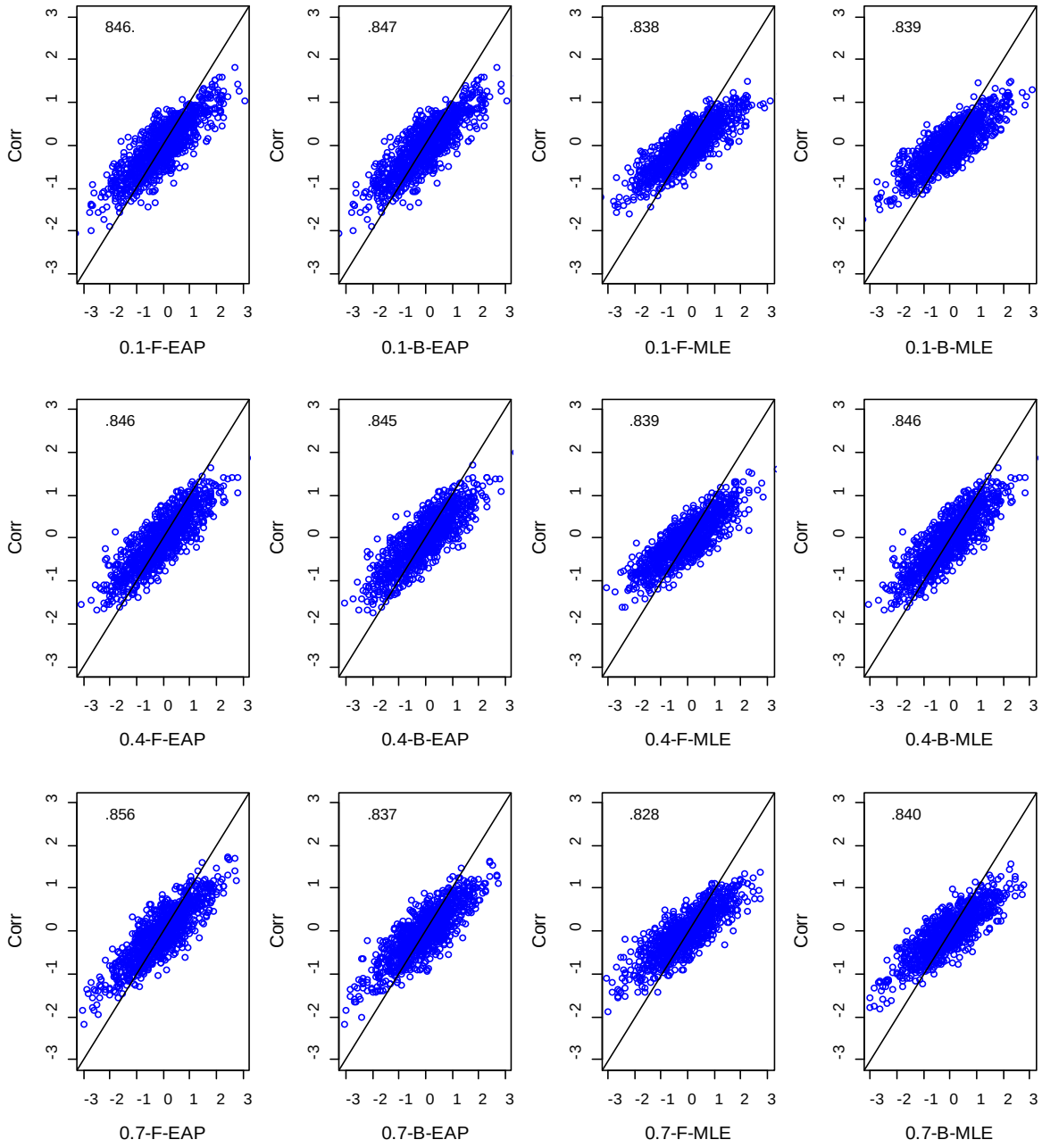


Figure A.14: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

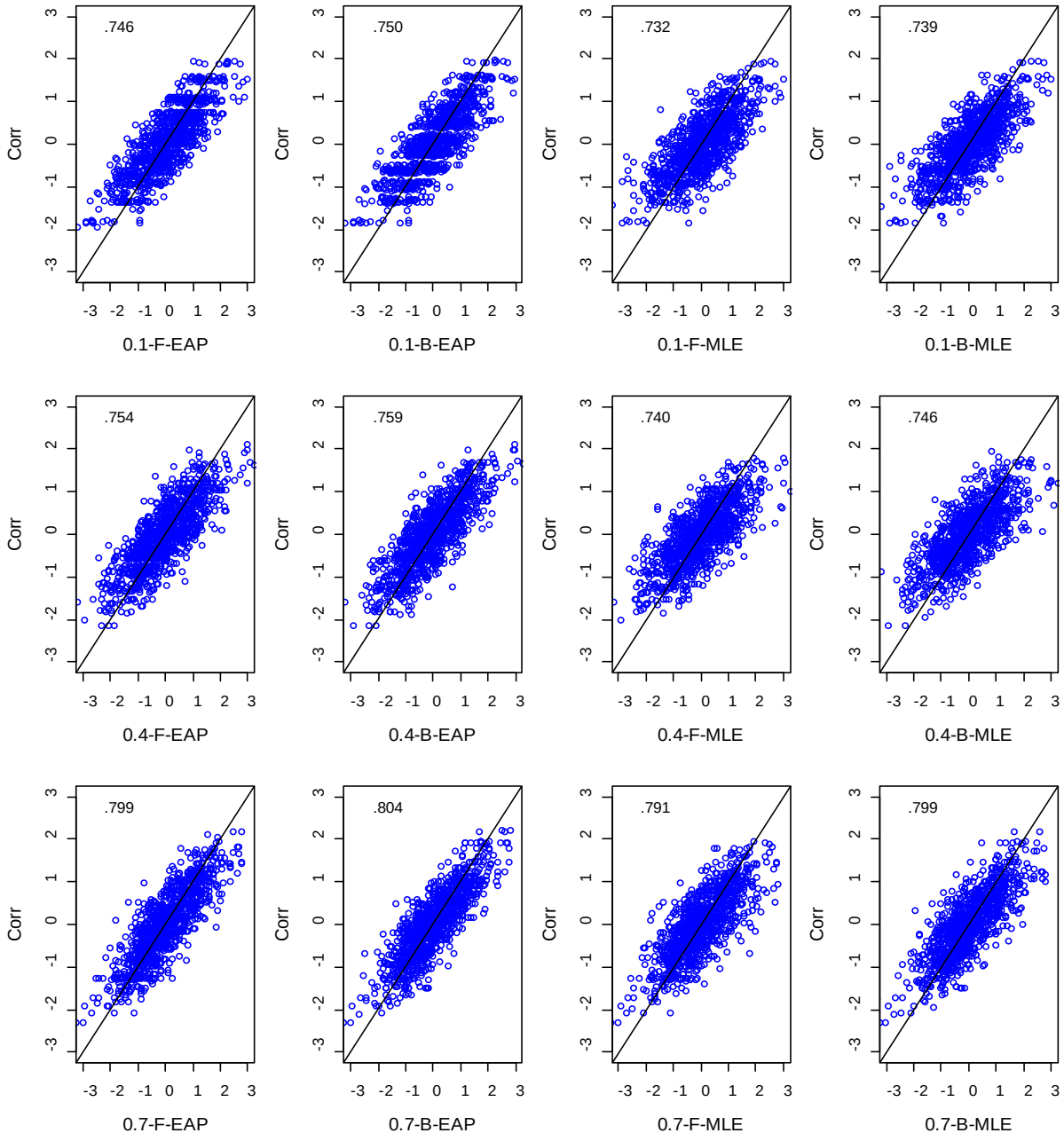


Figure A.15: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

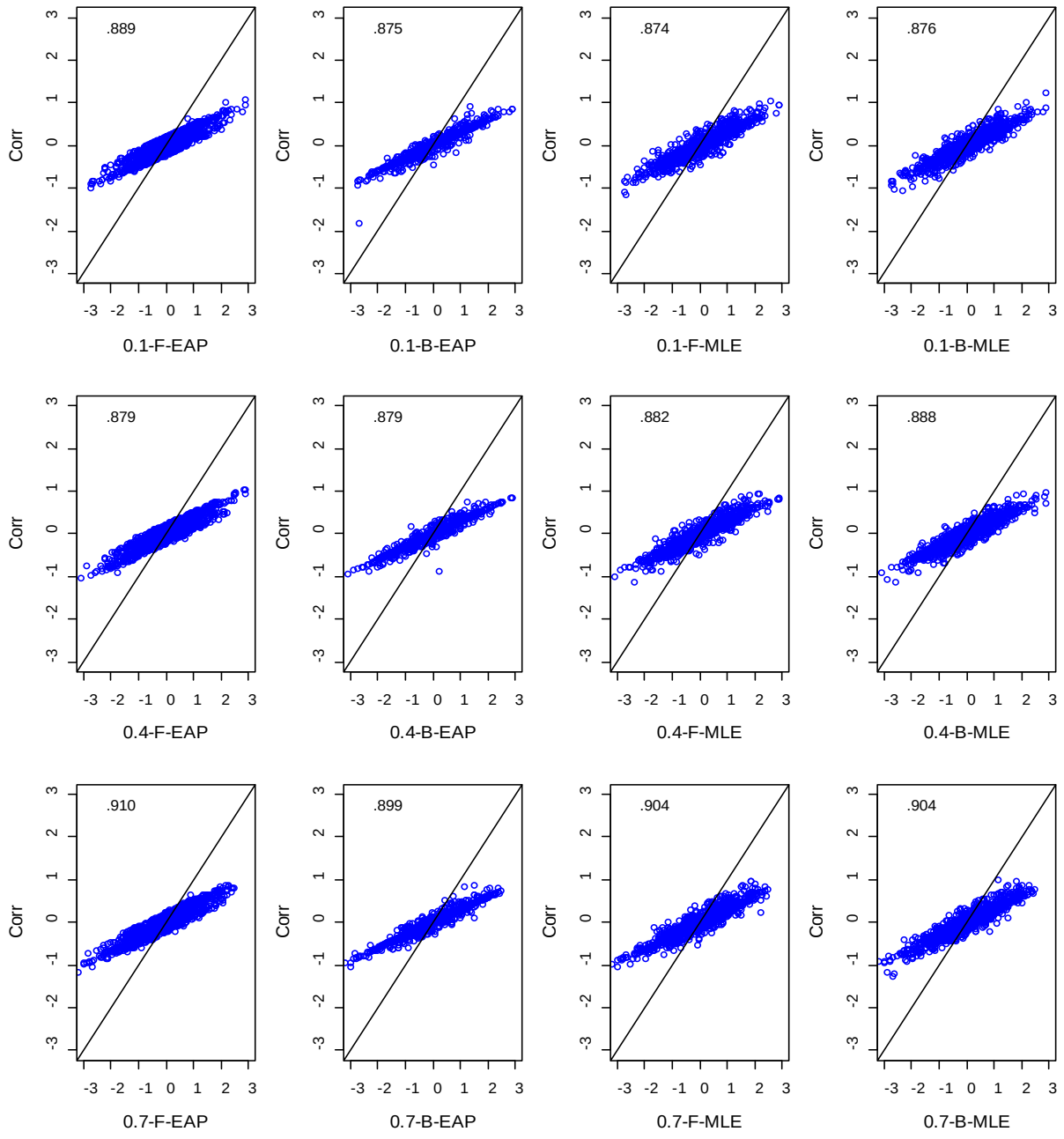


Figure A.16: Correlation between True and Estimated Proficiency Scores (First group factor for Two-tier IRT model with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

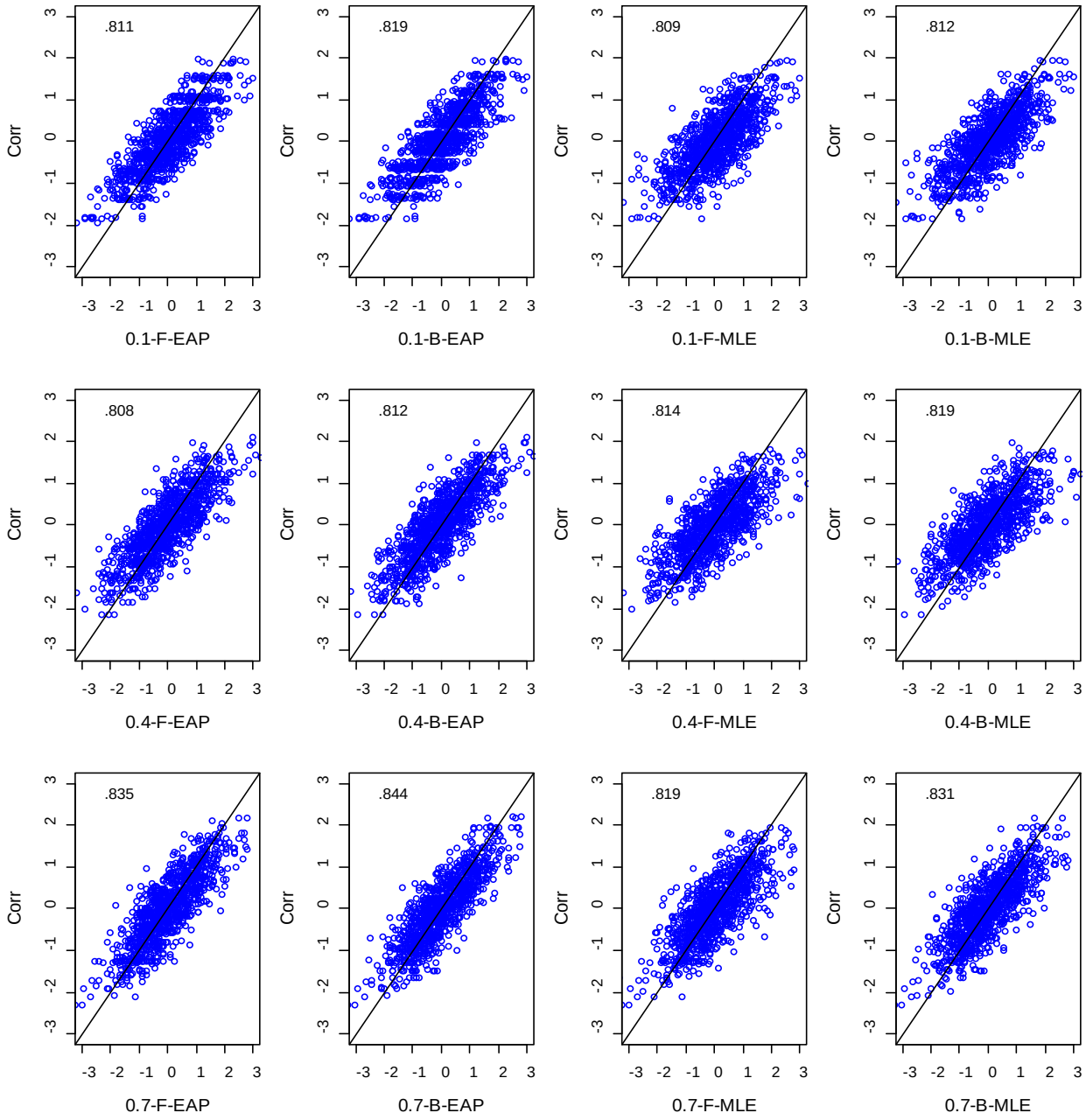


Figure A.17: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

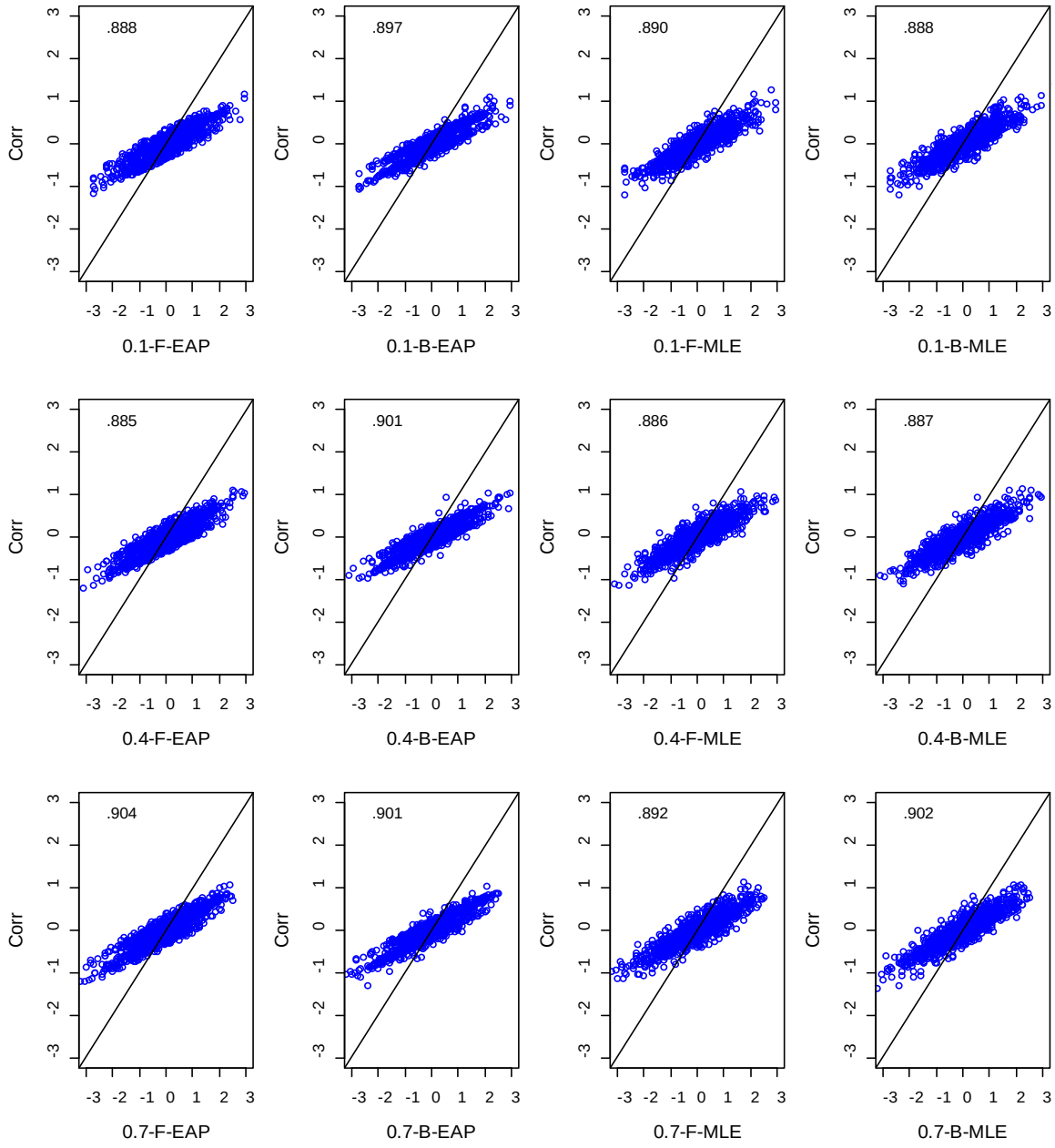


Figure A.18: Correlation between True and Estimated Proficiency Scores (First group factor for Two-tier IRT model with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

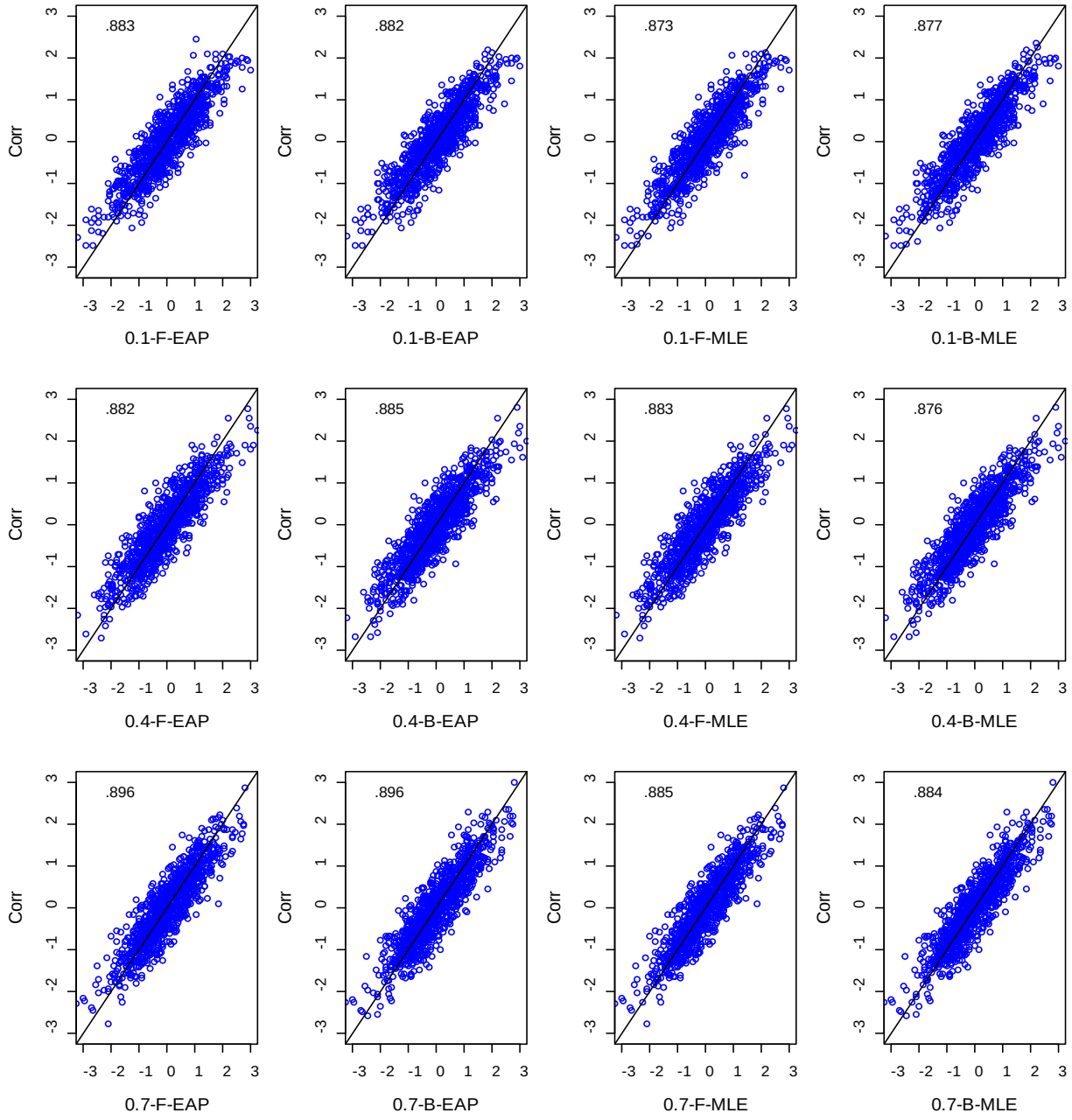


Figure A.19: Correlation between True and Estimated Proficiency Scores (First primary factor for Two-tier IRT model with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

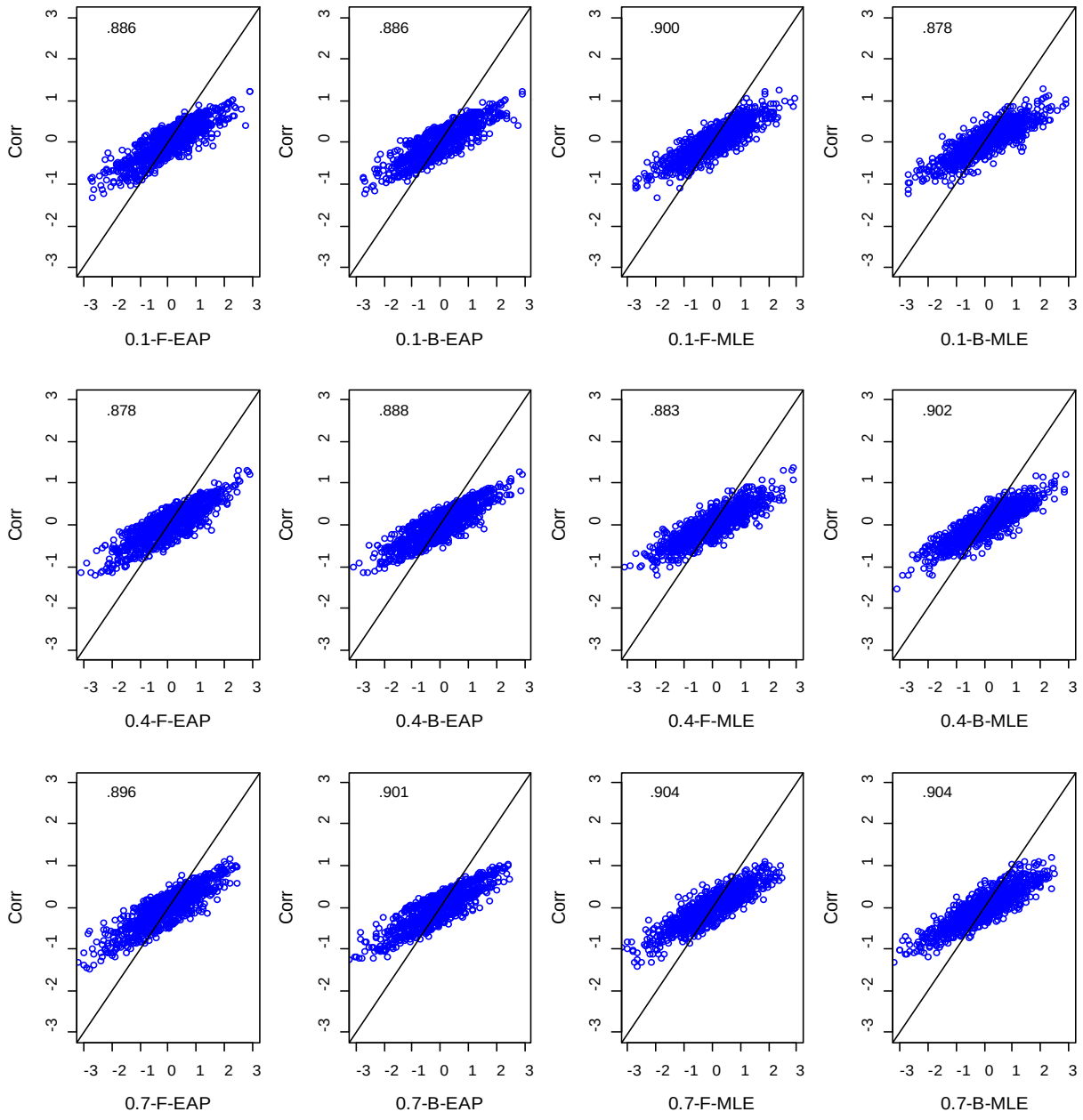


Figure A.20: Correlation between True and Estimated Proficiency Scores (First group factor for Two-tier IRT model with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring method

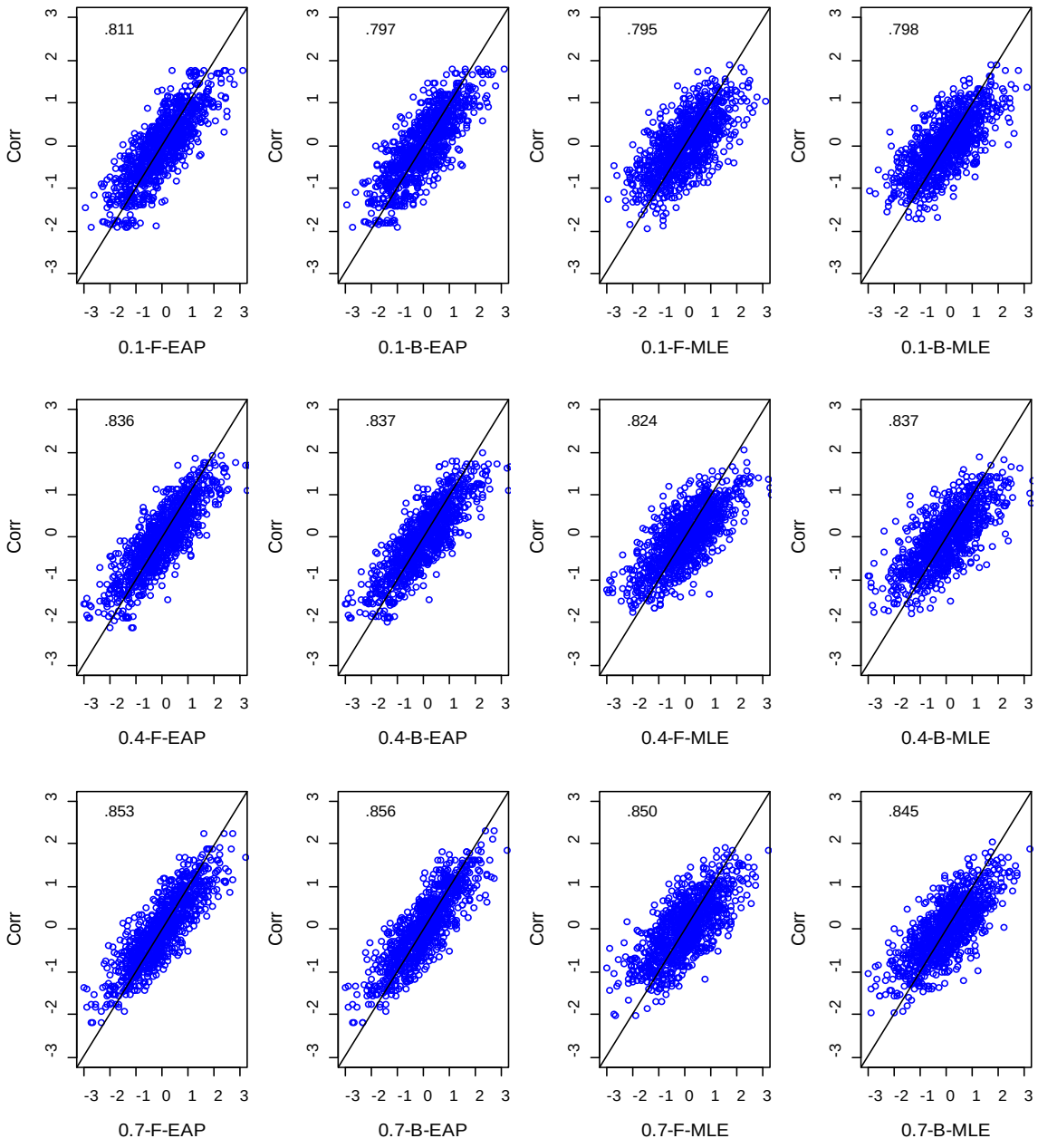


Figure A.21: Correlation between True and Estimated Proficiency Scores (First primary factor for Higher-order IRT model (2 primary factors) with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

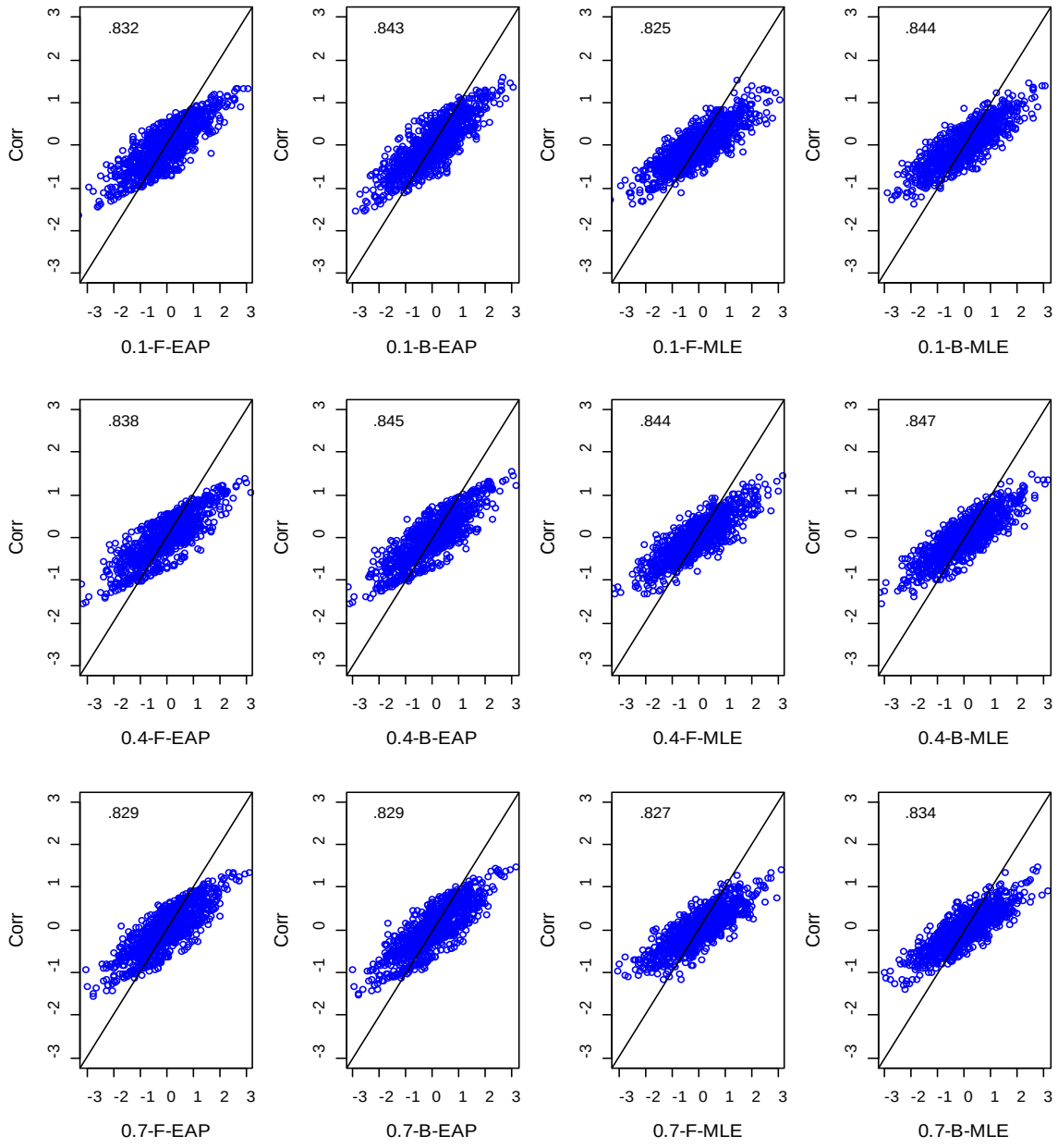


Figure A.22: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model (2 primary factors) with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

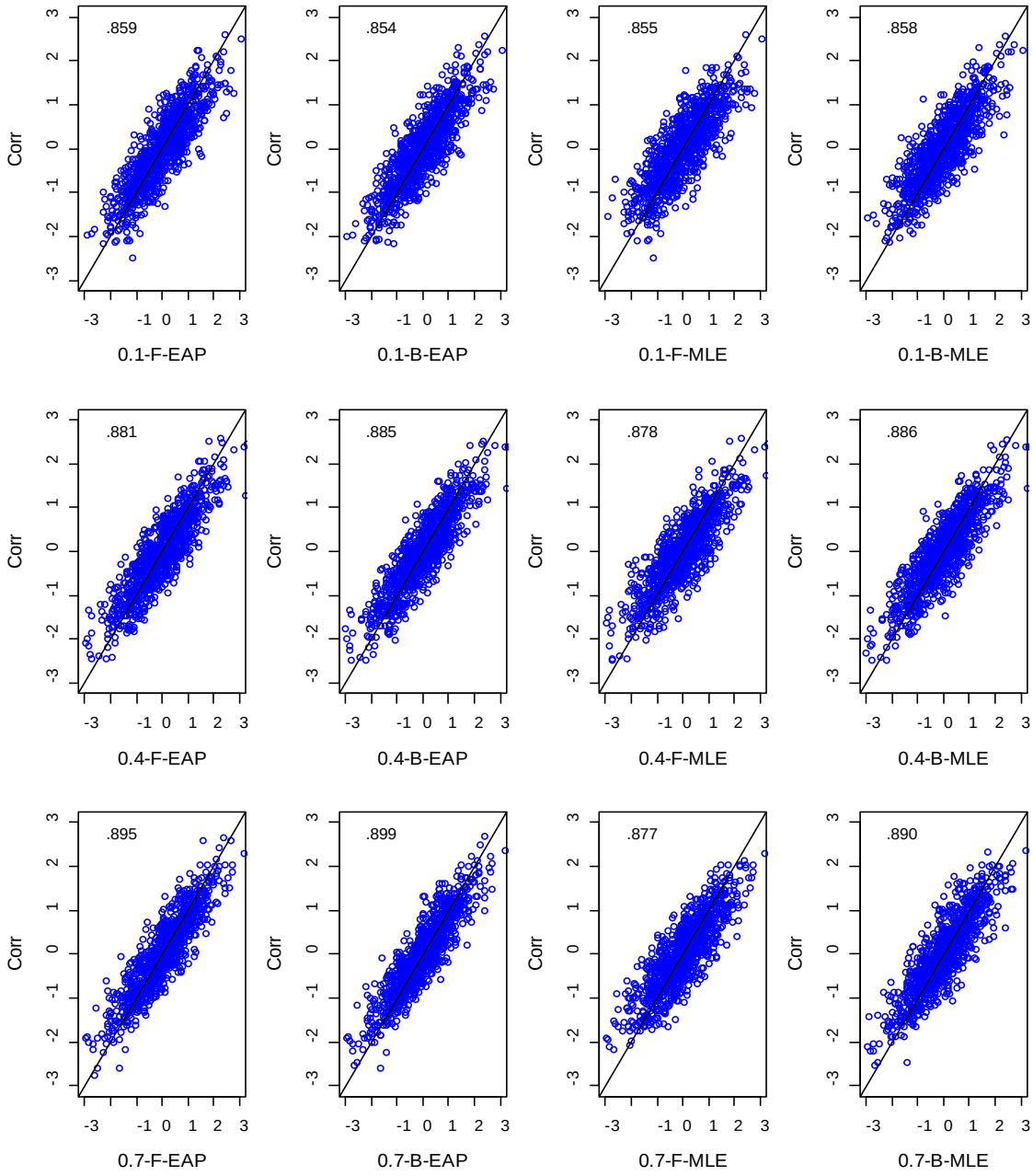


Figure A.23: Correlation between True and Estimated Proficiency Scores (First primary factor for Higher-order IRT model (2 primary factors) with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

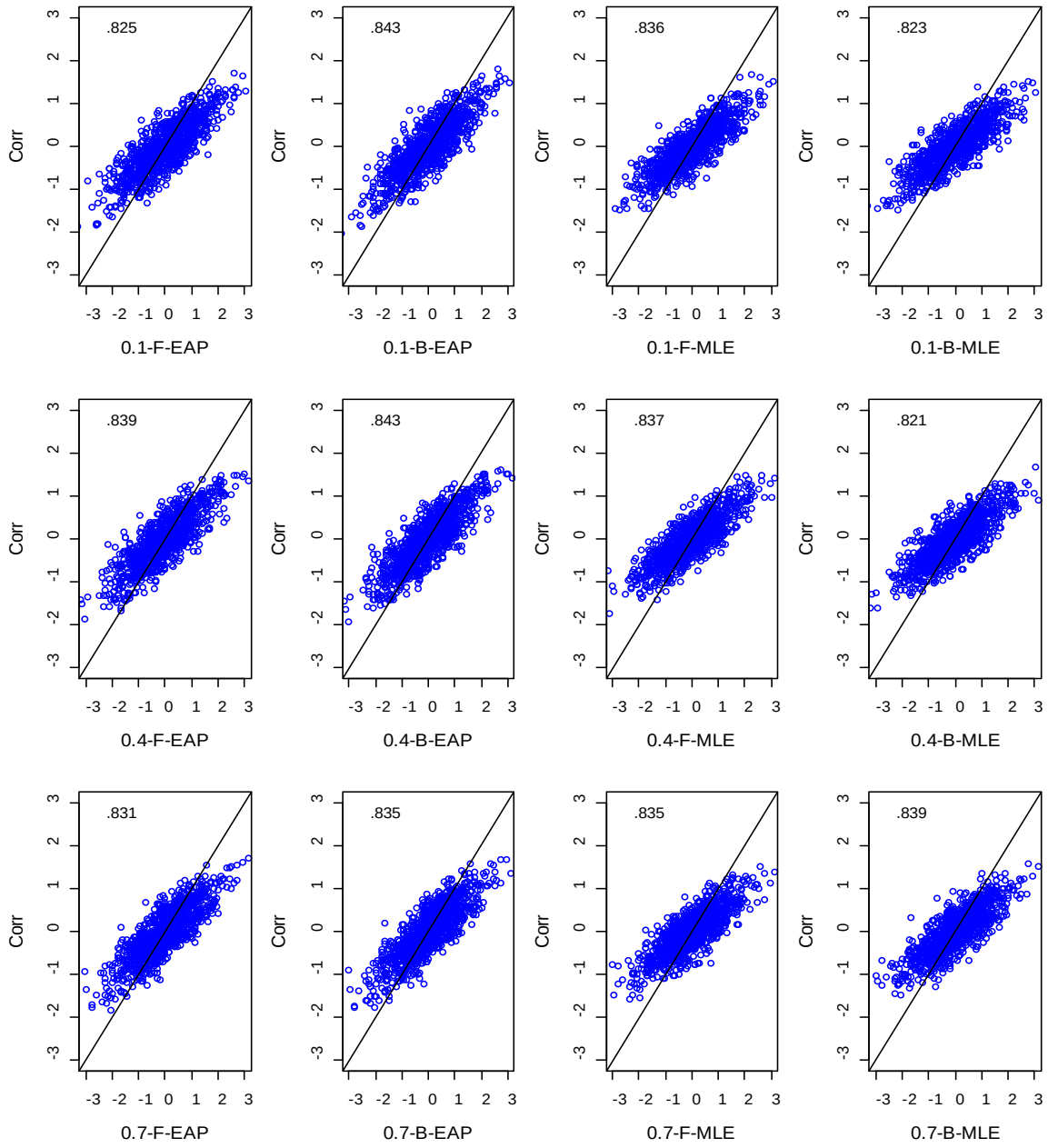


Figure A.24: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model (2 primary factors) with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

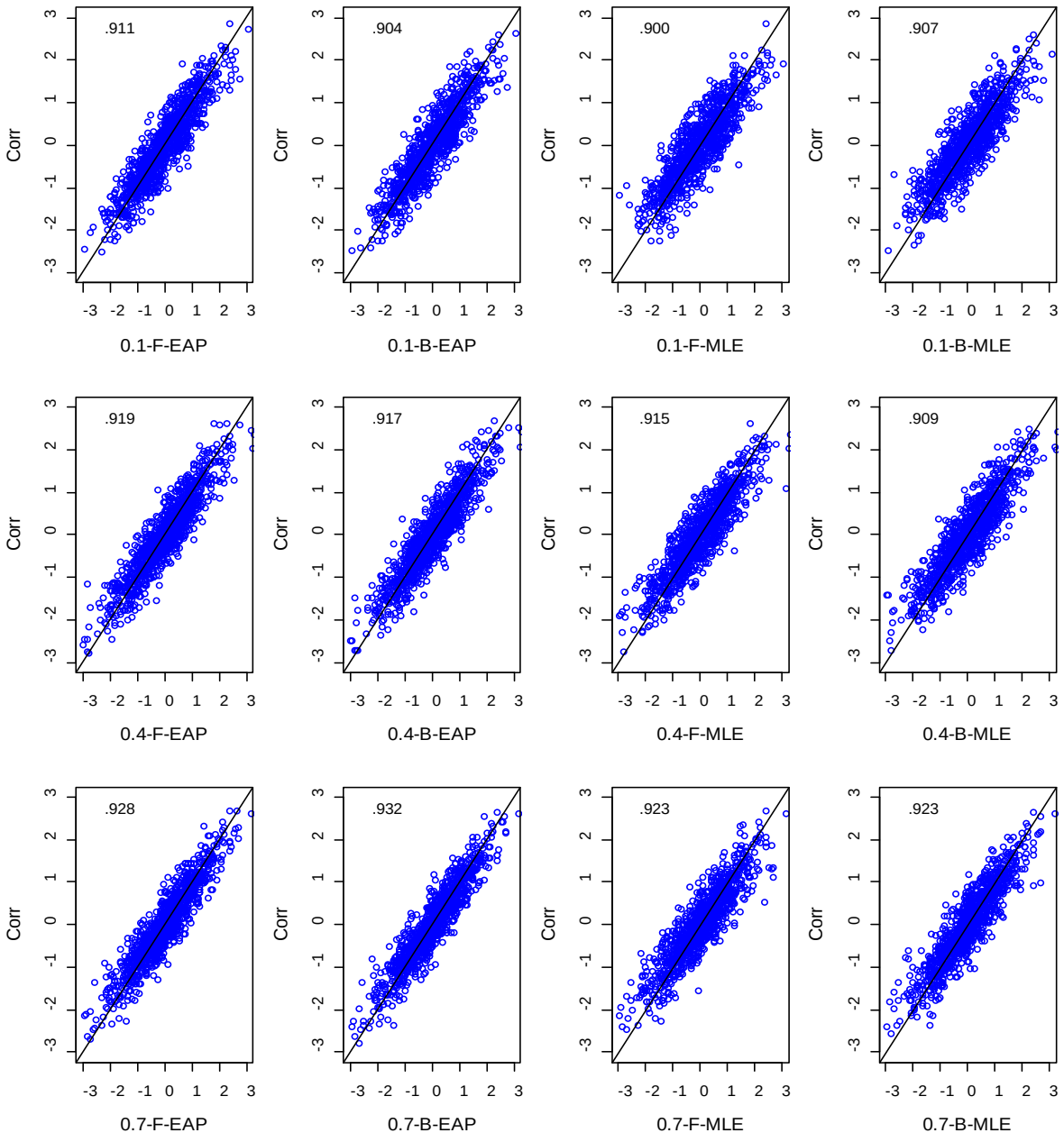


Figure A.25: Correlation between True and Estimated Proficiency Scores (First primary factor for Higher-order IRT model (2 primary factors) with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

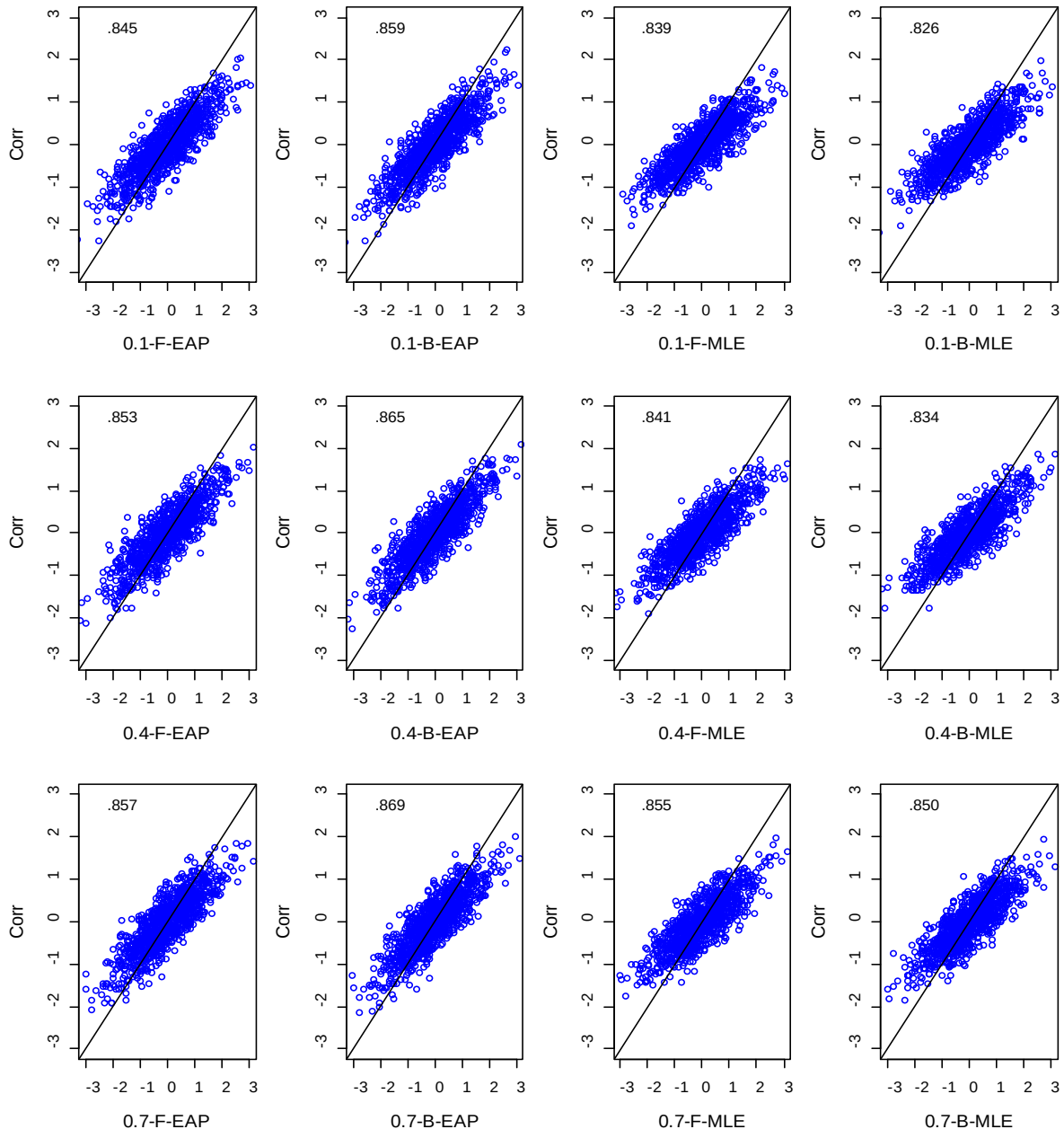


Figure A.26: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model (2 primary factors) with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

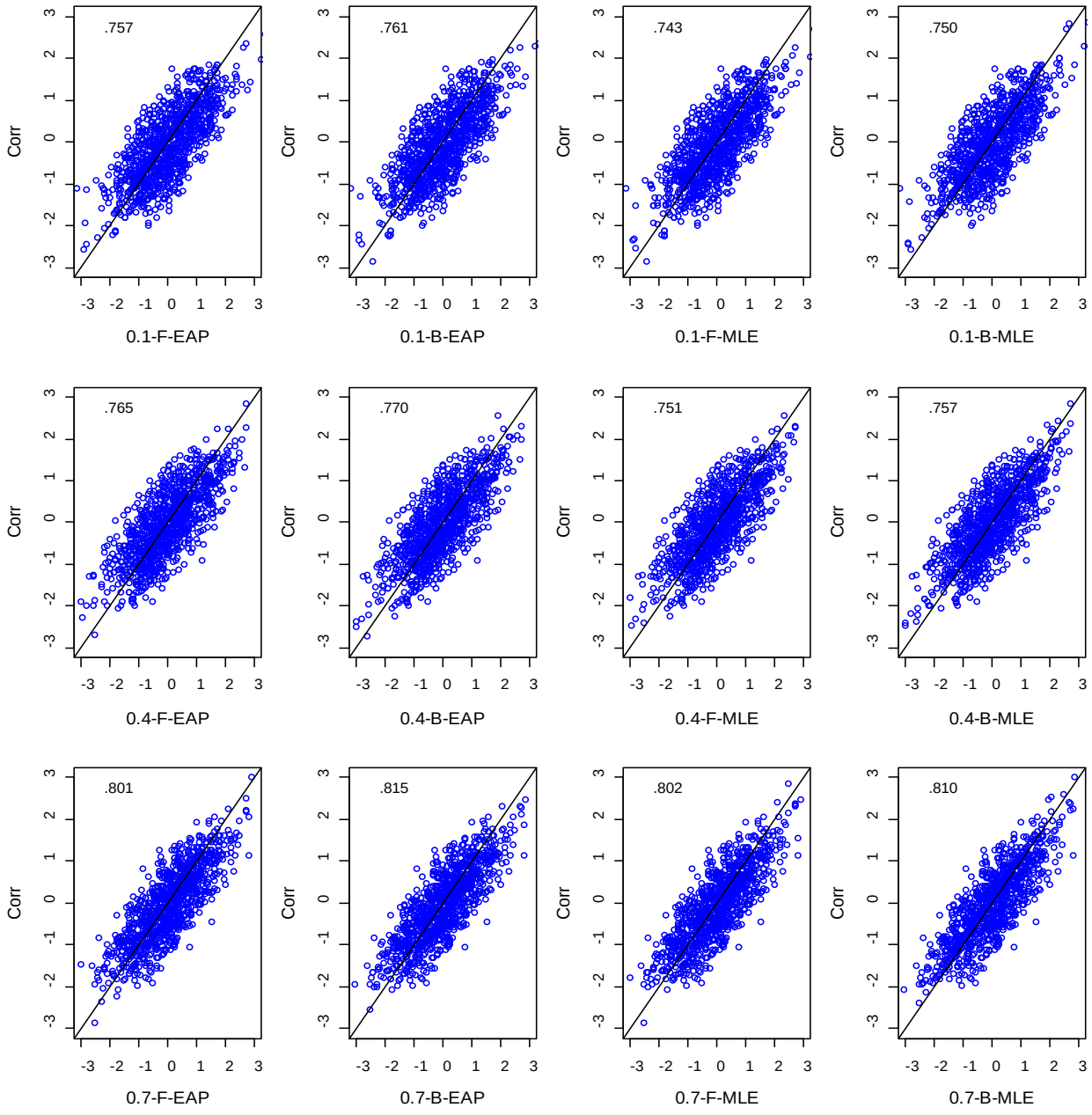


Figure A.27: Correlation between True and Estimated Proficiency Scores (First primary factor for Higher-order IRT model (2 primary factors) with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

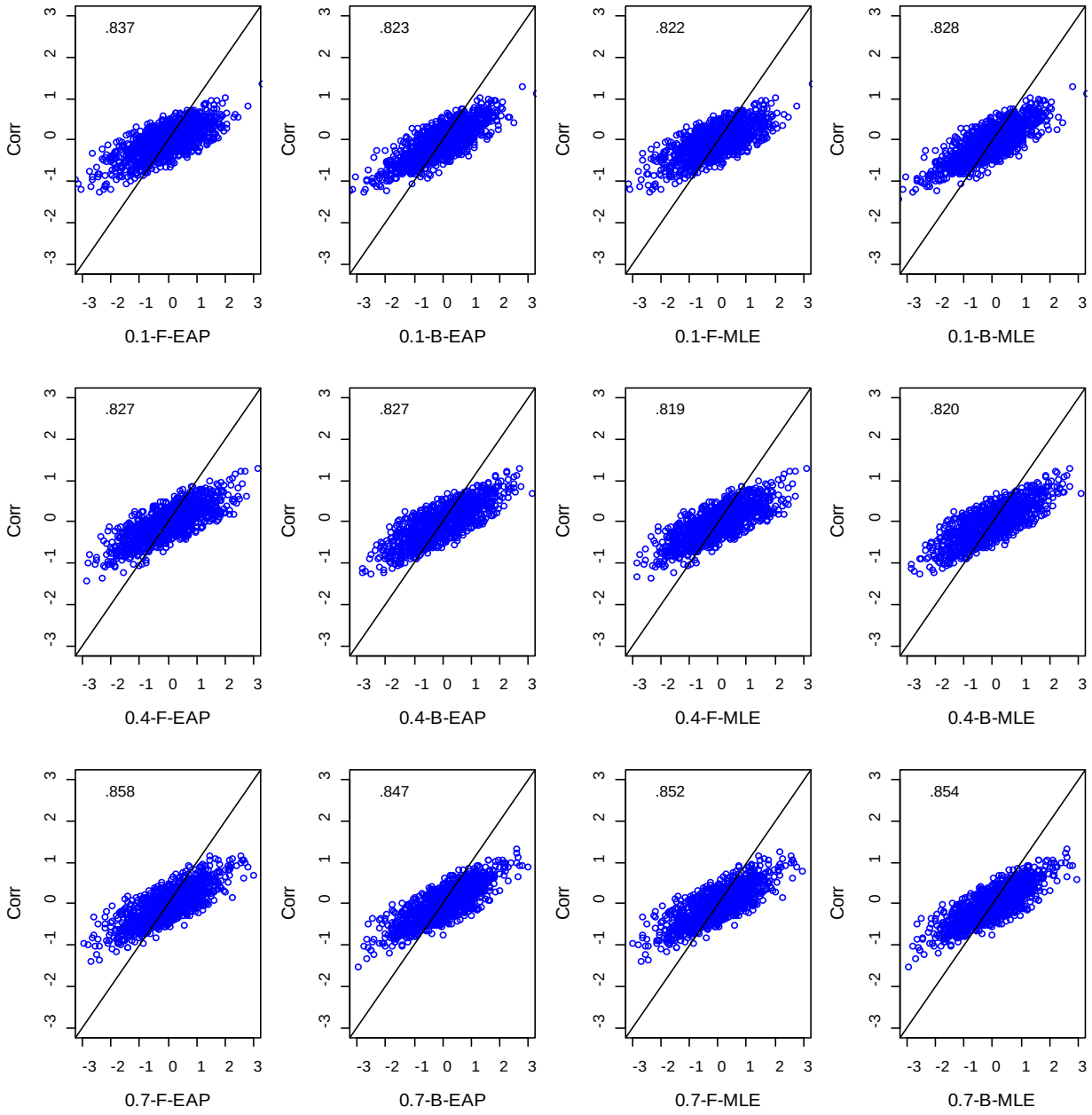


Figure A.28: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model (2 primary factors) with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

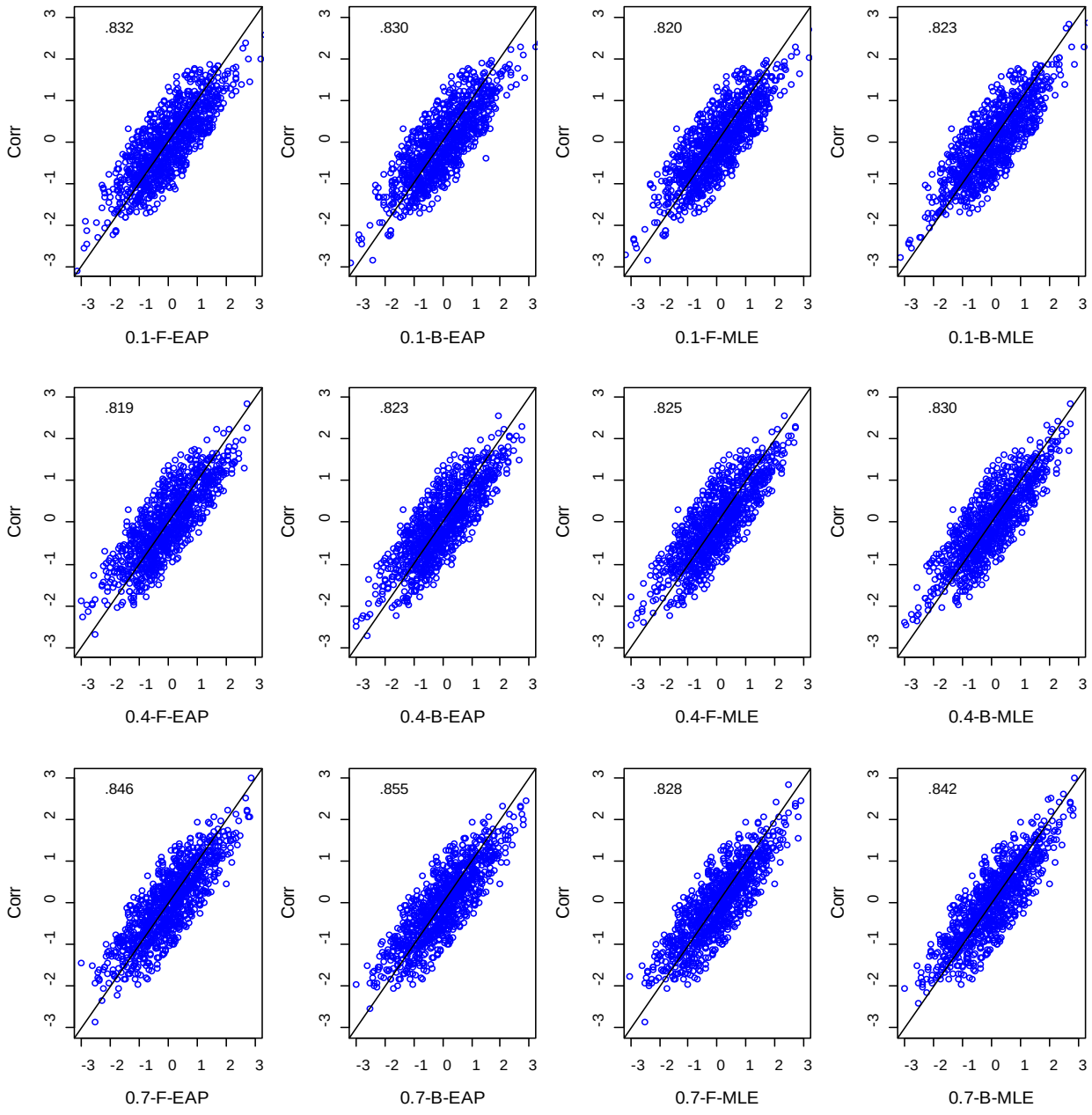


Figure A.29: Correlation between True and Estimated Proficiency Scores (First primary factor for Higher-order IRT model (2 primary factors) with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

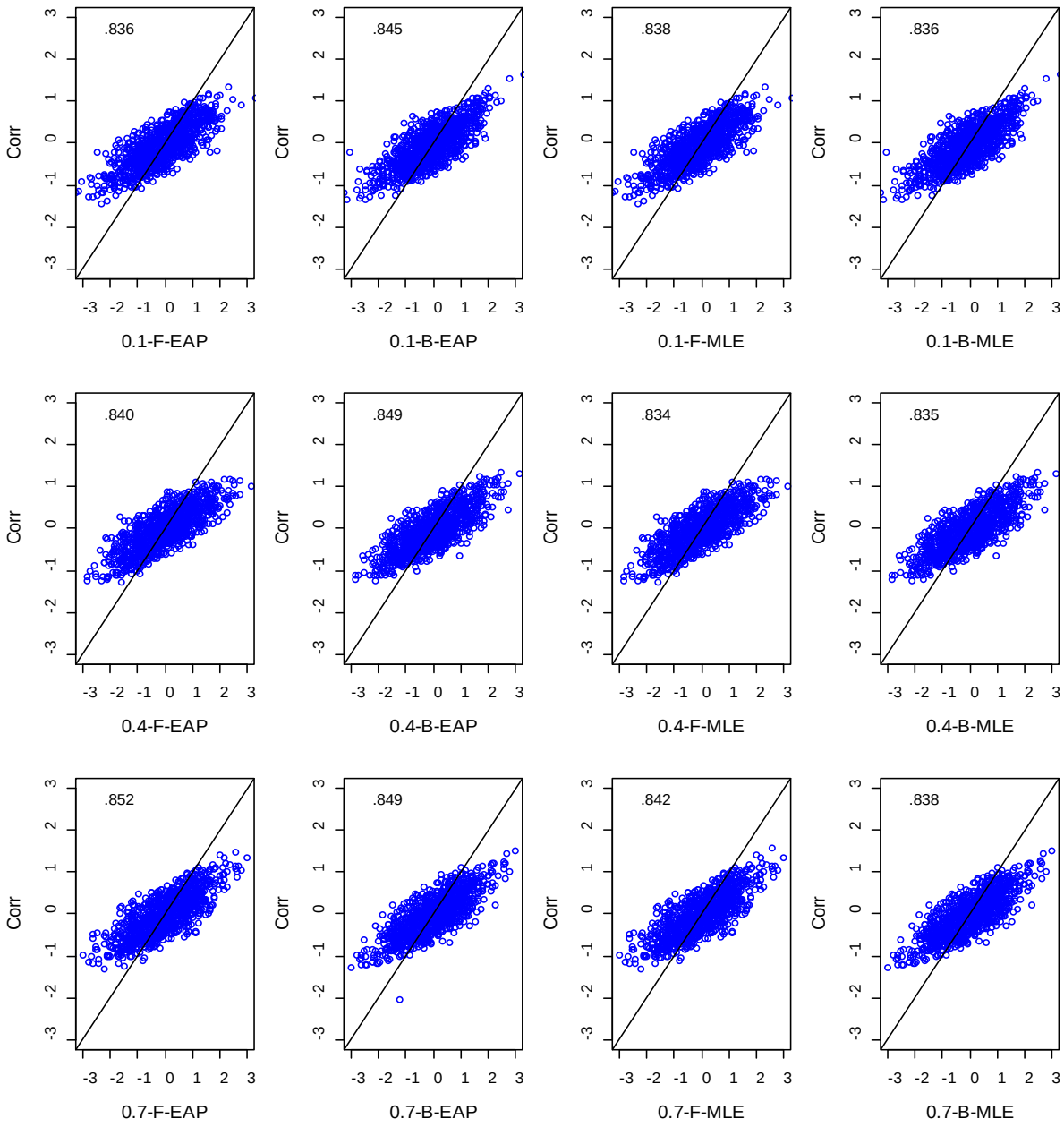


Figure A.30: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model (2 primary factors) with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

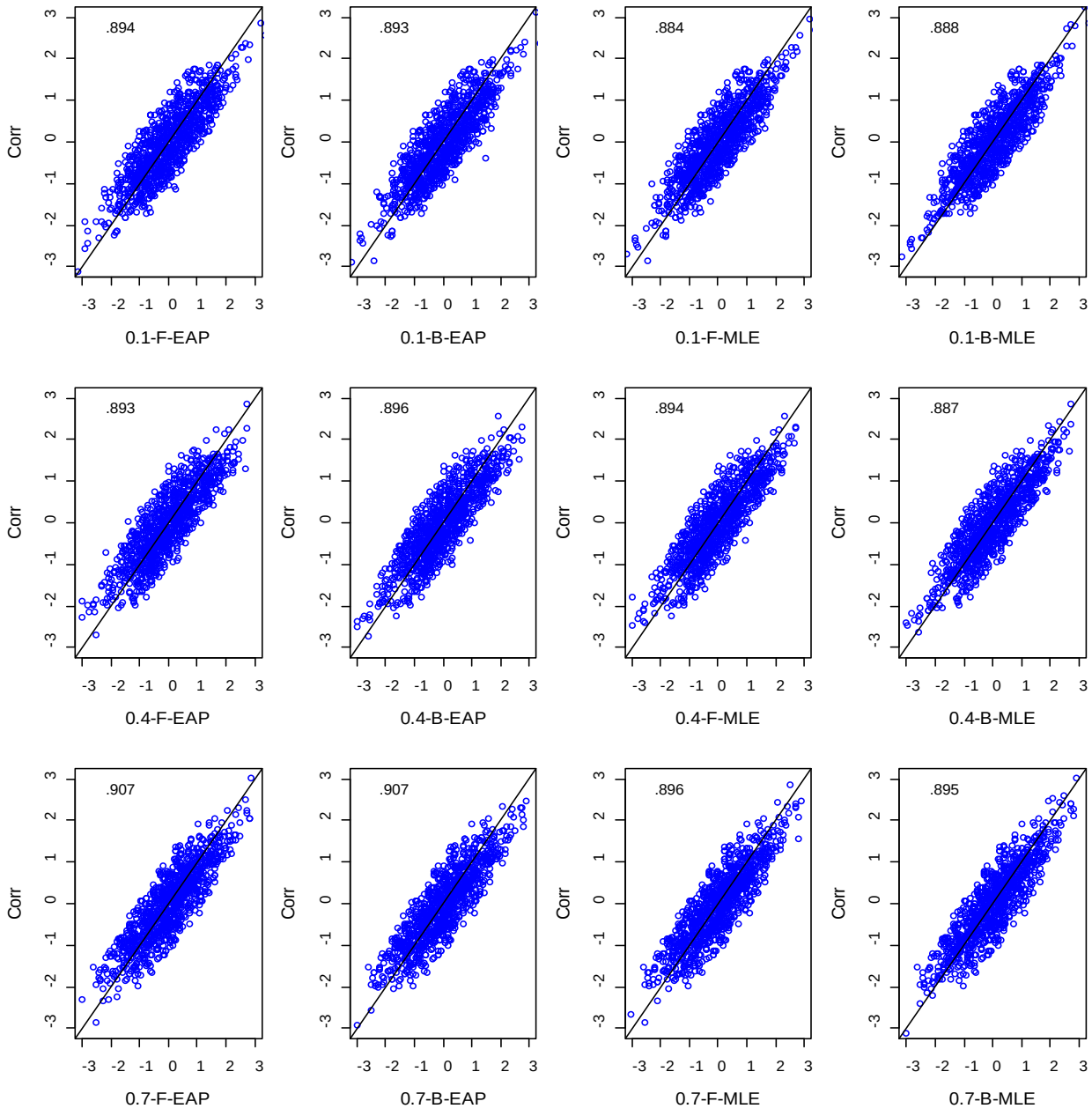


Figure A.31: Correlation between True and Estimated Proficiency Scores (First primary factor for Higher-order IRT model (2 primary factors) with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

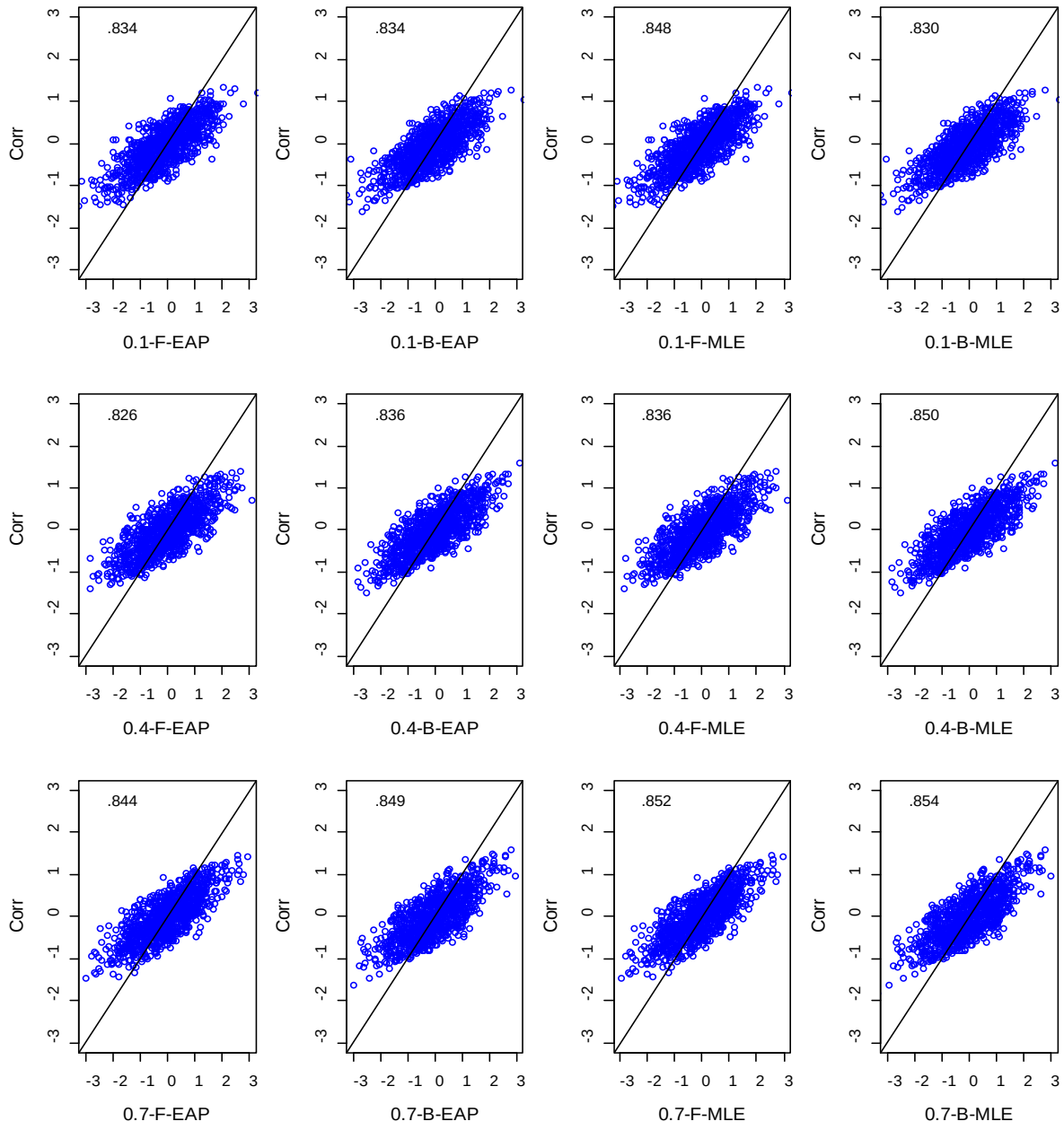


Figure A.32: Correlation between True and Estimated Proficiency Scores (First group factor for Higher-order IRT model (2 primary factors) with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two primary factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two primary factors – Bayesian item selection method – MLE scoring

APPENDIX B

Average Root Mean Squared Error (RMSE)

Table B.1: Average RMSE (Bifactor IRT Model)

Test Length	Item Selection	Scoring	Number of group factors							
			Two group factors			Four group factors				
			G_1	s_1	s_2	G_1	s_1	s_2	s_3	s_4
40										
	MFI	MLE	0.415	0.444	0.472	0.416	0.526	0.525	0.528	0.541
		EAP	0.405	0.452	0.468	0.411	0.509	0.500	0.498	0.510
	Bayes	MLE	0.415	0.466	0.512	0.415	0.523	0.521	0.528	0.547
		EAP	0.409	0.444	0.470	0.411	0.509	0.502	0.497	0.508
80										
	MFI	MLE	0.385	0.422	0.455	0.343	0.507	0.516	0.512	0.523
		EAP	0.366	0.431	0.441	0.330	0.486	0.483	0.465	0.485
	Bayes	MLE	0.374	0.419	0.444	0.353	0.497	0.523	0.512	0.517
		EAP	0.372	0.416	0.445	0.338	0.486	0.489	0.475	0.475
160										
	MFI	MLE	0.334	0.417	0.448	0.334	0.487	0.490	0.484	0.492
		EAP	0.334	0.424	0.432	0.327	0.462	0.463	0.470	0.484
	Bayes	MLE	0.343	0.416	0.426	0.340	0.486	0.490	0.487	0.474
		EAP	0.337	0.414	0.432	0.332	0.482	0.483	0.471	0.469

Table B.2: Average RMSE (Higher-order IRT Model)

Test Length	Item Selection	Scoring	Number of group factors							
			Two group factors			Four group factors				
			G_1	s_1	s_2	G_1	s_1	s_2	s_3	s_4
40										
	MFI	MLE	0.414	0.458	0.468	0.399	0.495	0.487	0.523	0.503
		EAP	0.418	0.450	0.449	0.397	0.471	0.454	0.474	0.465
	Bayes	MLE	0.413	0.446	0.463	0.388	0.479	0.477	0.511	0.490
		EAP	0.415	0.469	0.482	0.399	0.469	0.449	0.463	0.449
80										
	MFI	MLE	0.371	0.445	0.449	0.328	0.470	0.489	0.514	0.487
		EAP	0.370	0.472	0.475	0.327	0.466	0.454	0.466	0.460
	Bayes	MLE	0.365	0.457	0.467	0.333	0.477	0.491	0.502	0.491
		EAP	0.367	0.465	0.473	0.319	0.464	0.447	0.461	0.439
160										
	MFI	MLE	0.324	0.437	0.441	0.319	0.469	0.460	0.478	0.450
		EAP	0.324	0.431	0.434	0.314	0.456	0.449	0.455	0.438
	Bayes	MLE	0.323	0.428	0.441	0.312	0.459	0.446	0.480	0.460
		EAP	0.317	0.439	0.432	0.323	0.434	0.441	0.458	0.429

Table B.3: Average RMSE (Two-tier IRT Model with two group factors)

Test Length	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
	Low correlation b/t two general factors							
	MFI	MLE	0.472	0.480	0.534	0.562	0.550	0.530
		EAP	0.481	0.499	0.528	0.533	0.515	0.528
	Bayes	MLE	0.486	0.487	0.541	0.570	0.532	0.534
		EAP	0.480	0.508	0.528	0.532	0.516	0.529
40	*Medium correlation b/t two general factors*							
	MFI	MLE	0.464	0.472	0.527	0.551	0.558	0.534
		EAP	0.470	0.469	0.506	0.513	0.523	0.536
	Bayes	MLE	0.477	0.460	0.536	0.556	0.557	0.530
		EAP	0.459	0.471	0.509	0.518	0.533	0.537
	High correlation b/t two general factors							
	MFI	MLE	0.460	0.465	0.529	0.558	0.530	0.535
		EAP	0.464	0.458	0.504	0.531	0.496	0.529
	Bayes	MLE	0.473	0.452	0.527	0.558	0.537	0.534
		EAP	0.453	0.458	0.517	0.536	0.513	0.529
	Low correlation b/t two general factors							
	MFI	MLE	0.434	0.433	0.526	0.546	0.523	0.521
		EAP	0.423	0.426	0.490	0.515	0.490	0.494
	Bayes	MLE	0.429	0.437	0.527	0.557	0.519	0.518
		EAP	0.425	0.427	0.490	0.516	0.488	0.494
80	*Medium correlation b/t two general factors*							
	MFI	MLE	0.430	0.436	0.515	0.550	0.535	0.507
		EAP	0.421	0.432	0.481	0.499	0.500	0.491
	Bayes	MLE	0.418	0.428	0.515	0.538	0.539	0.516
		EAP	0.414	0.417	0.485	0.498	0.505	0.483
	High correlation b/t two general factors							
	MFI	MLE	0.411	0.424	0.524	0.561	0.506	0.526
		EAP	0.413	0.424	0.476	0.514	0.466	0.504
	Bayes	MLE	0.417	0.416	0.507	0.536	0.501	0.521
		EAP	0.409	0.400	0.499	0.517	0.490	0.491
	Low correlation b/t two general factors							
	MFI	MLE	0.375	0.393	0.508	0.520	0.487	0.487
		EAP	0.373	0.383	0.471	0.479	0.457	0.445
	Bayes	MLE	0.384	0.375	0.502	0.522	0.496	0.478
		EAP	0.370	0.382	0.461	0.475	0.458	0.447
160	*Medium correlation b/t two general factors*							
	MFI	MLE	0.369	0.370	0.485	0.505	0.501	0.478
		EAP	0.362	0.374	0.462	0.466	0.482	0.430
	Bayes	MLE	0.364	0.374	0.452	0.478	0.479	0.439
		EAP	0.361	0.362	0.453	0.494	0.481	0.438
	High correlation b/t two general factors							
	MFI	MLE	0.358	0.368	0.486	0.510	0.503	0.490
		EAP	0.357	0.355	0.445	0.477	0.449	0.464
	Bayes	MLE	0.354	0.364	0.486	0.519	0.483	0.490
		EAP	0.362	0.351	0.468	0.486	0.457	0.481

Table B.4: Average RMSE (Two-tier IRT Model with four group factors)

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
40												
Low correlation b/t two general factors												
	MFI	MLE	0.524	0.535	0.527	0.588	0.562	0.563	0.556	0.585	0.523	0.560
		EAP	0.536	0.536	0.528	0.579	0.561	0.559	0.538	0.588	0.522	0.547
	Bayes	MLE	0.528	0.539	0.522	0.591	0.556	0.566	0.560	0.589	0.526	0.557
		EAP	0.528	0.536	0.535	0.579	0.563	0.557	0.538	0.587	0.523	0.546
Medium correlation b/t two general factors												
	MFI	MLE	0.543	0.515	0.536	0.577	0.538	0.560	0.566	0.582	0.511	0.558
		EAP	0.529	0.496	0.542	0.570	0.541	0.547	0.539	0.571	0.500	0.557
	Bayes	MLE	0.523	0.521	0.544	0.570	0.535	0.571	0.569	0.573	0.511	0.550
		EAP	0.532	0.497	0.549	0.571	0.535	0.558	0.562	0.577	0.502	0.558
High correlation b/t two general factors												
	MFI	MLE	0.504	0.500	0.558	0.558	0.558	0.529	0.561	0.603	0.528	0.542
		EAP	0.502	0.497	0.550	0.548	0.563	0.526	0.546	0.594	0.527	0.534
	Bayes	MLE	0.500	0.505	0.555	0.556	0.559	0.526	0.569	0.605	0.530	0.542
		EAP	0.506	0.489	0.567	0.557	0.562	0.529	0.571	0.603	0.523	0.541
80												
Low correlation b/t two general factors												
	MFI	MLE	0.462	0.444	0.486	0.546	0.516	0.535	0.524	0.547	0.485	0.527
		EAP	0.465	0.455	0.482	0.528	0.524	0.522	0.501	0.551	0.485	0.510
	Bayes	MLE	0.453	0.463	0.478	0.544	0.520	0.530	0.524	0.554	0.487	0.526
		EAP	0.455	0.456	0.487	0.527	0.527	0.519	0.501	0.550	0.486	0.509
Medium correlation b/t two general factors												
	MFI	MLE	0.438	0.445	0.502	0.530	0.498	0.525	0.534	0.536	0.463	0.519
		EAP	0.434	0.442	0.504	0.526	0.505	0.509	0.502	0.534	0.463	0.520
	Bayes	MLE	0.438	0.442	0.504	0.529	0.507	0.527	0.529	0.538	0.469	0.519
		EAP	0.425	0.436	0.505	0.528	0.498	0.518	0.525	0.541	0.465	0.521
High correlation b/t two general factors												
	MFI	MLE	0.420	0.440	0.519	0.515	0.529	0.480	0.525	0.571	0.485	0.512
		EAP	0.424	0.434	0.502	0.502	0.525	0.489	0.509	0.556	0.489	0.497
	Bayes	MLE	0.425	0.439	0.509	0.511	0.524	0.490	0.529	0.570	0.493	0.511
		EAP	0.421	0.435	0.504	0.510	0.524	0.493	0.534	0.566	0.486	0.504

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
160												
Low correlation b/t two general factors												
	MFI	MLE	0.422	0.436	0.478	0.514	0.515	0.519	0.528	0.546	0.487	0.527
		EAP	0.418	0.420	0.475	0.505	0.515	0.509	0.506	0.539	0.486	0.505
	Bayes	MLE	0.426	0.436	0.492	0.505	0.523	0.527	0.520	0.535	0.482	0.527
		EAP	0.423	0.426	0.481	0.504	0.516	0.509	0.506	0.545	0.487	0.504
Medium correlation b/t two general factors												
	MFI	MLE	0.405	0.434	0.495	0.514	0.501	0.523	0.539	0.538	0.474	0.520
		EAP	0.415	0.395	0.492	0.502	0.494	0.494	0.502	0.532	0.474	0.514
	Bayes	MLE	0.414	0.416	0.499	0.509	0.488	0.514	0.533	0.535	0.465	0.519
		EAP	0.400	0.399	0.490	0.502	0.485	0.497	0.511	0.532	0.474	0.522
High correlation b/t two general factors												
	MFI	MLE	0.389	0.383	0.492	0.493	0.511	0.495	0.513	0.539	0.489	0.505
		EAP	0.377	0.384	0.499	0.497	0.510	0.480	0.508	0.537	0.478	0.499
	Bayes	MLE	0.386	0.390	0.487	0.504	0.516	0.487	0.514	0.521	0.484	0.507
		EAP	0.385	0.384	0.489	0.491	0.502	0.478	0.502	0.525	0.479	0.502

Table B.5: Average RMSE (Higher-order IRT Model (2 primary factors) with two group factors)

Test Length	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
	Low correlation b/t two general factors							
	MFI	MLE	0.438	0.447	0.523	0.532	0.534	0.559
		EAP	0.430	0.437	0.506	0.477	0.518	0.525
	Bayes	MLE	0.436	0.445	0.512	0.528	0.533	0.545
		EAP	0.430	0.444	0.480	0.487	0.516	0.525
40	*Medium correlation b/t two general factors*							
	MFI	MLE	0.435	0.432	0.518	0.503	0.531	0.539
		EAP	0.426	0.424	0.496	0.470	0.506	0.522
	Bayes	MLE	0.427	0.443	0.514	0.509	0.523	0.542
		EAP	0.422	0.436	0.487	0.479	0.505	0.523
	High correlation b/t two general factors							
	MFI	MLE	0.423	0.433	0.502	0.533	0.517	0.561
		EAP	0.421	0.430	0.473	0.505	0.499	0.523
	Bayes	MLE	0.425	0.425	0.508	0.540	0.511	0.558
		EAP	0.415	0.421	0.466	0.526	0.499	0.530
	Low correlation b/t two general factors							
	MFI	MLE	0.412	0.404	0.501	0.517	0.509	0.514
		EAP	0.408	0.395	0.485	0.452	0.491	0.507
	Bayes	MLE	0.416	0.409	0.512	0.516	0.526	0.525
		EAP	0.395	0.411	0.465	0.468	0.475	0.507
80	*Medium correlation b/t two general factors*							
	MFI	MLE	0.397	0.409	0.503	0.498	0.499	0.517
		EAP	0.394	0.405	0.472	0.460	0.472	0.493
	Bayes	MLE	0.395	0.407	0.506	0.492	0.494	0.515
		EAP	0.393	0.397	0.462	0.460	0.466	0.492
	High correlation b/t two general factors							
	MFI	MLE	0.391	0.404	0.493	0.512	0.498	0.520
		EAP	0.390	0.396	0.455	0.468	0.473	0.498
	Bayes	MLE	0.382	0.404	0.482	0.502	0.495	0.523
		EAP	0.384	0.390	0.447	0.497	0.464	0.502
	Low correlation b/t two general factors							
	MFI	MLE	0.386	0.394	0.475	0.477	0.485	0.499
		EAP	0.384	0.387	0.454	0.420	0.450	0.475
	Bayes	MLE	0.378	0.386	0.484	0.469	0.488	0.503
		EAP	0.371	0.387	0.436	0.430	0.425	0.472
160	*Medium correlation b/t two general factors*							
	MFI	MLE	0.372	0.380	0.477	0.474	0.465	0.477
		EAP	0.371	0.370	0.440	0.432	0.439	0.464
	Bayes	MLE	0.369	0.367	0.484	0.474	0.472	0.500
		EAP	0.367	0.373	0.423	0.434	0.430	0.466
	High correlation b/t two general factors							
	MFI	MLE	0.357	0.369	0.466	0.492	0.469	0.485
		EAP	0.348	0.357	0.426	0.413	0.433	0.464
	Bayes	MLE	0.351	0.361	0.457	0.476	0.464	0.500
		EAP	0.352	0.361	0.422	0.442	0.426	0.478

Table B.6: Average RMSE (Higher-order IRT Model (2 primary factors) with four group factors)

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
40												
Low correlation b/t two general factors												
	MFI	MLE	0.508	0.519	0.604	0.628	0.601	0.604	0.605	0.634	0.530	0.612
		EAP	0.520	0.520	0.575	0.612	0.559	0.560	0.553	0.599	0.518	0.557
	Bayes	MLE	0.512	0.523	0.547	0.610	0.562	0.578	0.576	0.612	0.505	0.564
		EAP	0.512	0.514	0.574	0.606	0.565	0.560	0.552	0.605	0.519	0.553
Medium correlation b/t two general factors												
	MFI	MLE	0.507	0.499	0.600	0.608	0.602	0.613	0.598	0.606	0.556	0.610
		EAP	0.511	0.492	0.586	0.570	0.556	0.560	0.548	0.599	0.546	0.577
	Bayes	MLE	0.507	0.505	0.578	0.587	0.556	0.583	0.564	0.582	0.520	0.559
		EAP	0.516	0.491	0.586	0.569	0.559	0.562	0.541	0.594	0.539	0.563
High correlation b/t two general factors												
	MFI	MLE	0.488	0.486	0.558	0.603	0.569	0.574	0.556	0.577	0.526	0.546
		EAP	0.486	0.481	0.566	0.609	0.561	0.553	0.552	0.592	0.542	0.566
	Bayes	MLE	0.484	0.489	0.557	0.609	0.558	0.576	0.569	0.568	0.525	0.553
		EAP	0.490	0.483	0.564	0.609	0.569	0.549	0.547	0.593	0.538	0.565
80												
Low correlation b/t two general factors												
	MFI	MLE	0.445	0.447	0.538	0.612	0.565	0.588	0.575	0.609	0.509	0.564
		EAP	0.448	0.448	0.571	0.610	0.558	0.560	0.553	0.610	0.521	0.559
	Bayes	MLE	0.442	0.446	0.542	0.607	0.557	0.573	0.568	0.617	0.517	0.566
		EAP	0.448	0.440	0.557	0.606	0.562	0.563	0.552	0.618	0.521	0.555
Medium correlation b/t two general factors												
	MFI	MLE	0.421	0.428	0.556	0.585	0.552	0.595	0.563	0.580	0.528	0.575
		EAP	0.417	0.430	0.593	0.572	0.554	0.562	0.548	0.610	0.548	0.576
	Bayes	MLE	0.421	0.425	0.570	0.592	0.561	0.597	0.558	0.585	0.532	0.569
		EAP	0.408	0.439	0.592	0.567	0.556	0.565	0.541	0.600	0.541	0.562
High correlation b/t two general factors												
	MFI	MLE	0.413	0.423	0.545	0.596	0.566	0.574	0.558	0.577	0.523	0.542
		EAP	0.417	0.417	0.576	0.599	0.556	0.553	0.552	0.606	0.544	0.561
	Bayes	MLE	0.408	0.422	0.548	0.615	0.564	0.586	0.573	0.573	0.527	0.557
		EAP	0.404	0.418	0.564	0.604	0.565	0.552	0.547	0.605	0.541	0.561

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
160												
Low correlation b/t two general factors												
	MFI	MLE	0.404	0.412	0.534	0.615	0.539	0.578	0.575	0.605	0.503	0.571
		EAP	0.399	0.411	0.575	0.624	0.520	0.556	0.556	0.617	0.534	0.573
	Bayes	MLE	0.407	0.417	0.550	0.603	0.556	0.572	0.576	0.601	0.527	0.561
		EAP	0.404	0.407	0.562	0.632	0.536	0.553	0.555	0.630	0.528	0.569
Medium correlation b/t two general factors												
	MFI	MLE	0.386	0.395	0.561	0.590	0.545	0.584	0.552	0.585	0.543	0.567
		EAP	0.396	0.386	0.586	0.592	0.530	0.564	0.537	0.616	0.544	0.575
	Bayes	MLE	0.395	0.397	0.554	0.583	0.543	0.580	0.557	0.593	0.545	0.567
		EAP	0.381	0.384	0.586	0.605	0.528	0.559	0.527	0.610	0.542	0.563
High correlation b/t two general factors												
	MFI	MLE	0.370	0.364	0.537	0.611	0.553	0.561	0.561	0.589	0.526	0.527
		EAP	0.364	0.373	0.568	0.623	0.541	0.551	0.543	0.614	0.544	0.574
	Bayes	MLE	0.367	0.371	0.547	0.620	0.564	0.564	0.552	0.568	0.534	0.542
		EAP	0.366	0.365	0.562	0.622	0.526	0.546	0.534	0.612	0.537	0.574

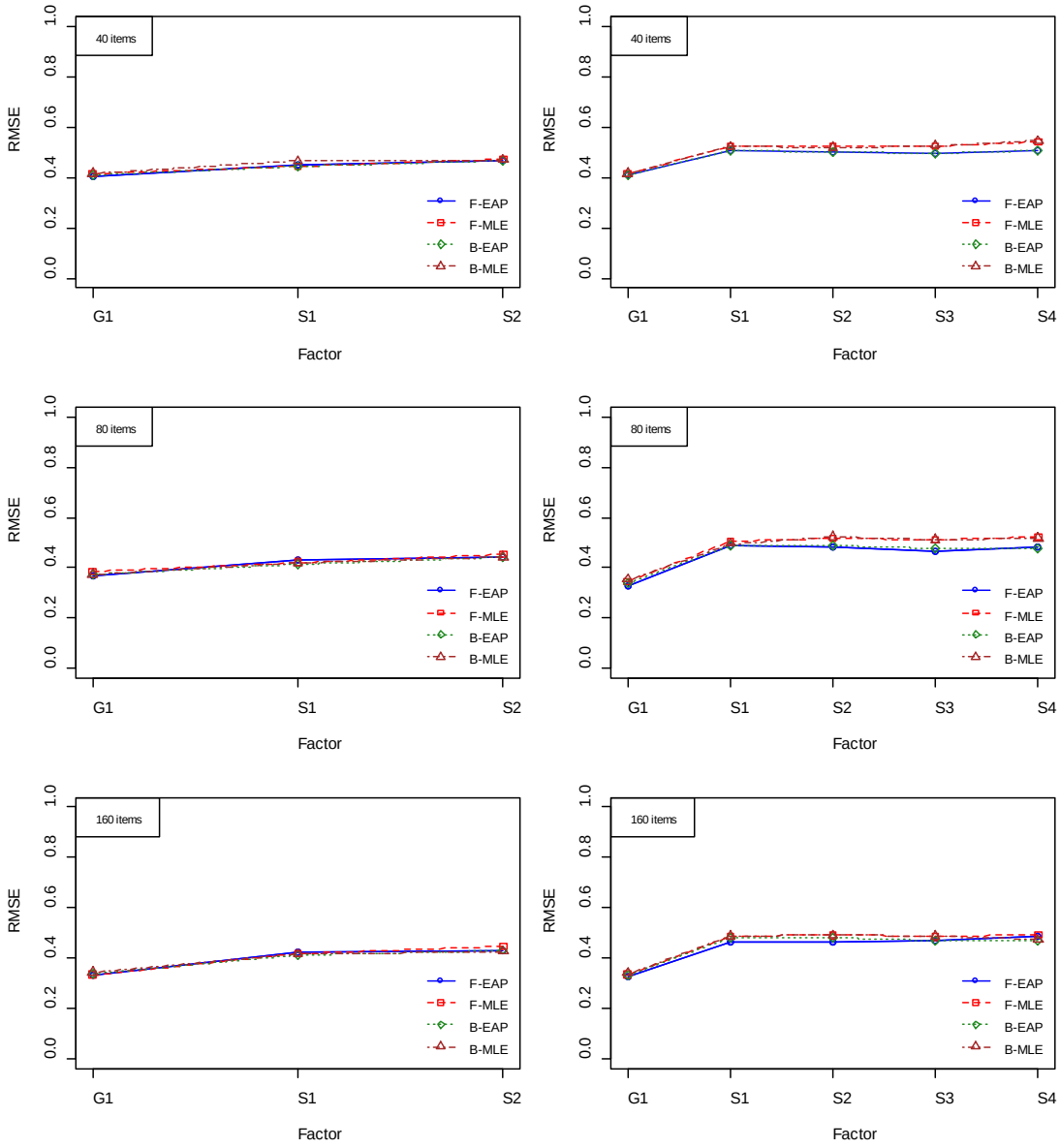


Figure B.1: Average RMSE (Bifactor IRT model)

Note. F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method

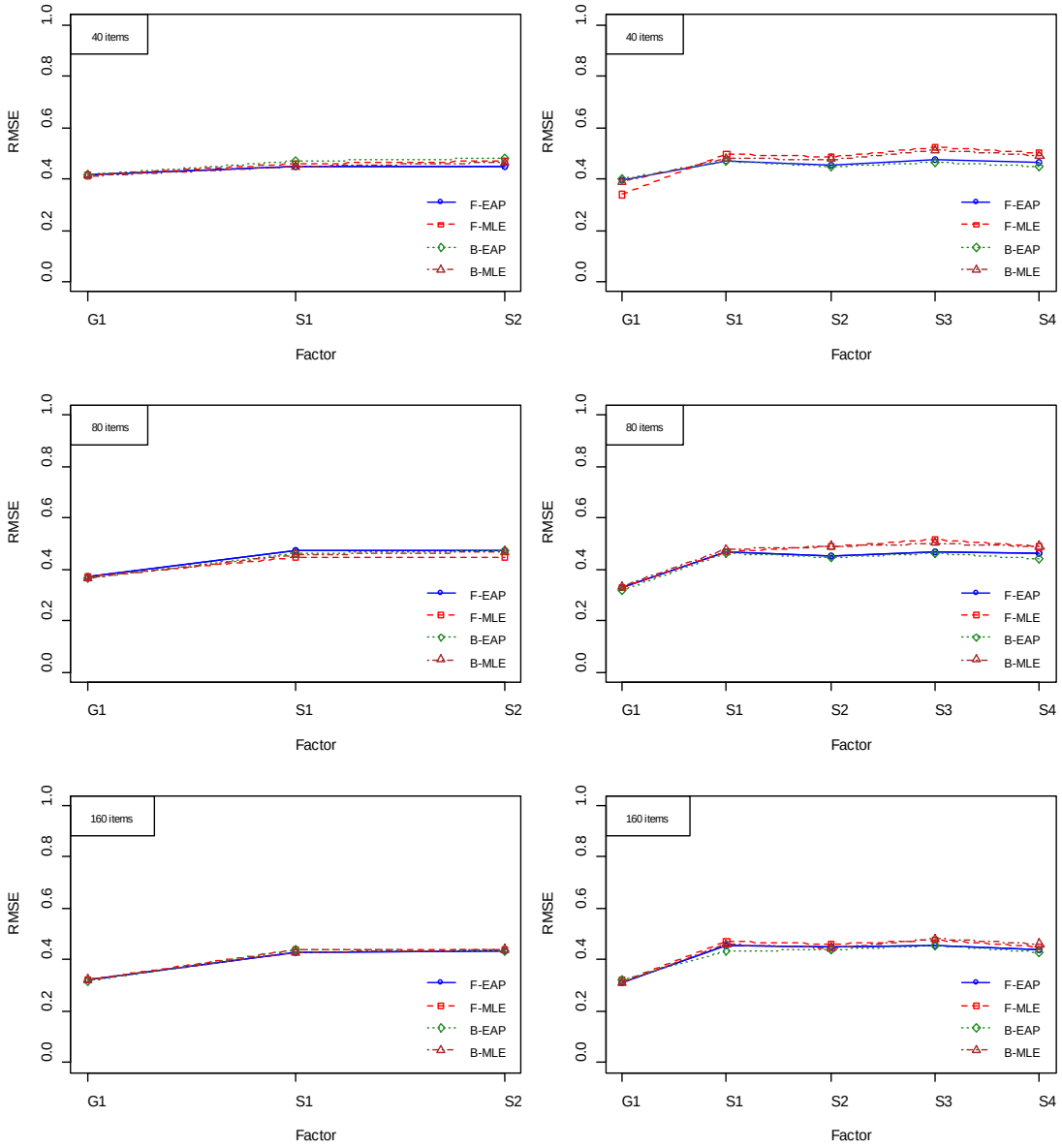


Figure B.2: Average RMSE (Higher-order IRT model)

Note. F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method

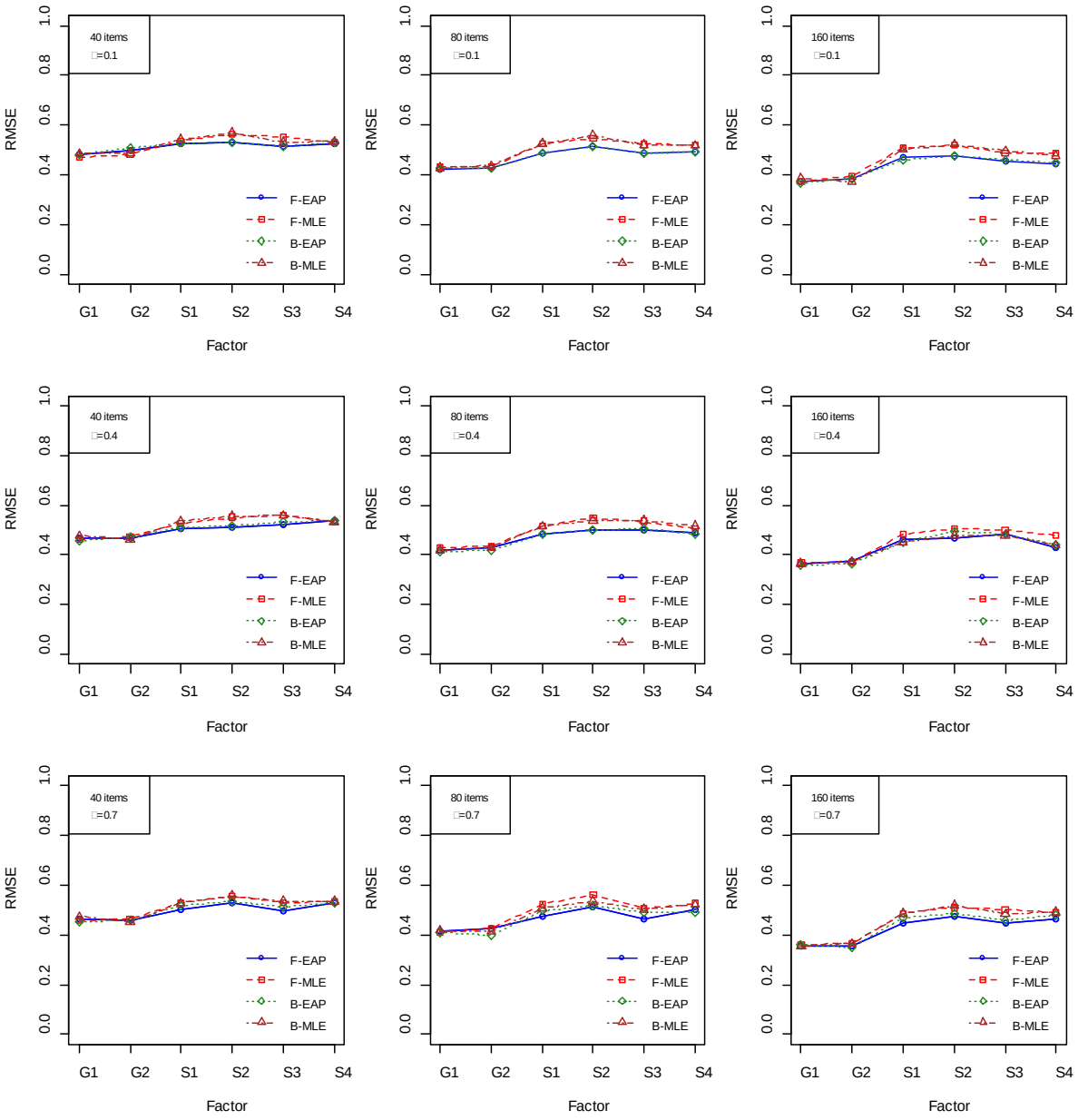


Figure B.3: Average RMSE (Two-tier IRT model with two group factors)

Note. F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method

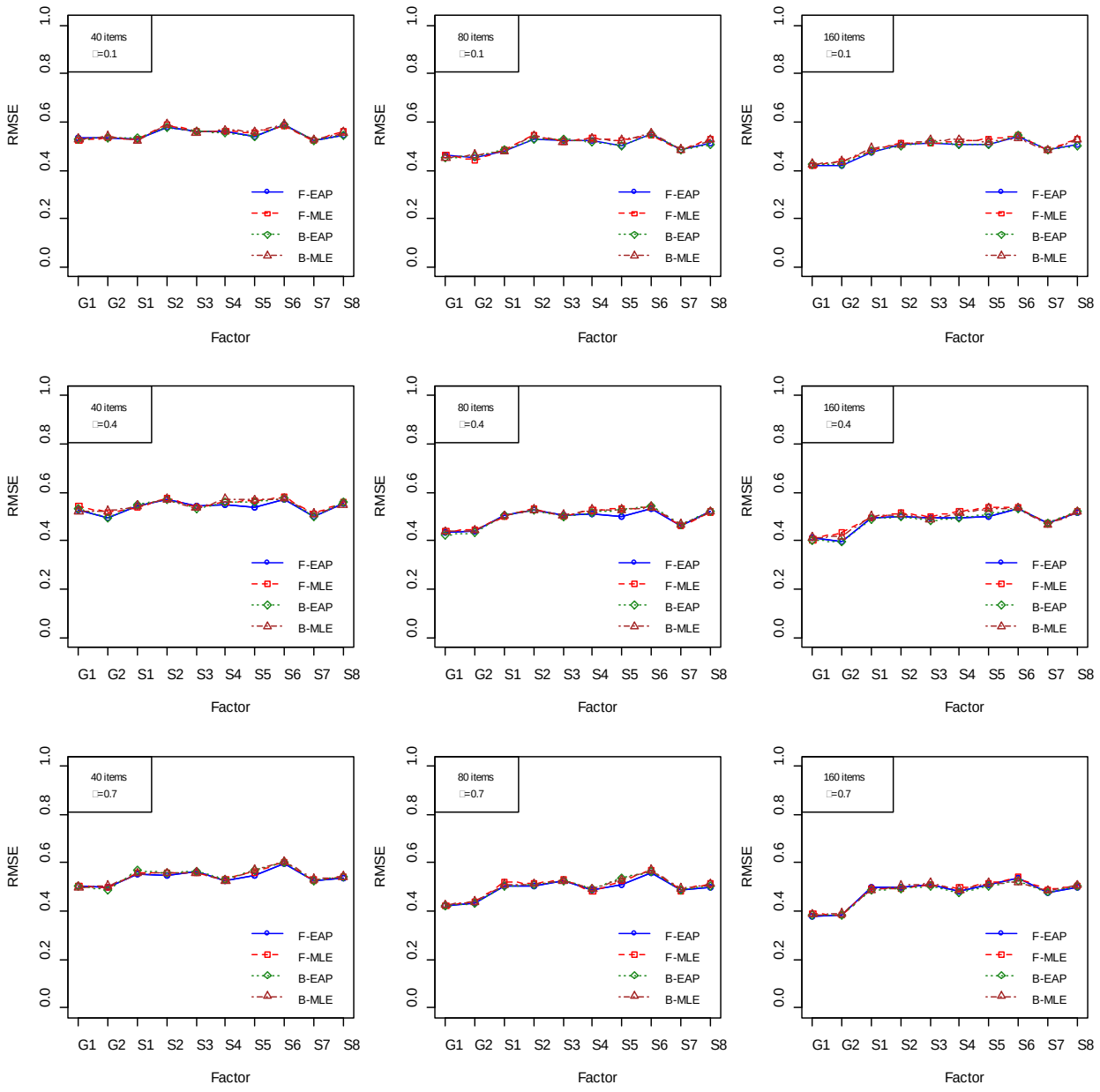


Figure B.4: Average RMSE (Two-tier IRT model with four group factors)

Note. F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method

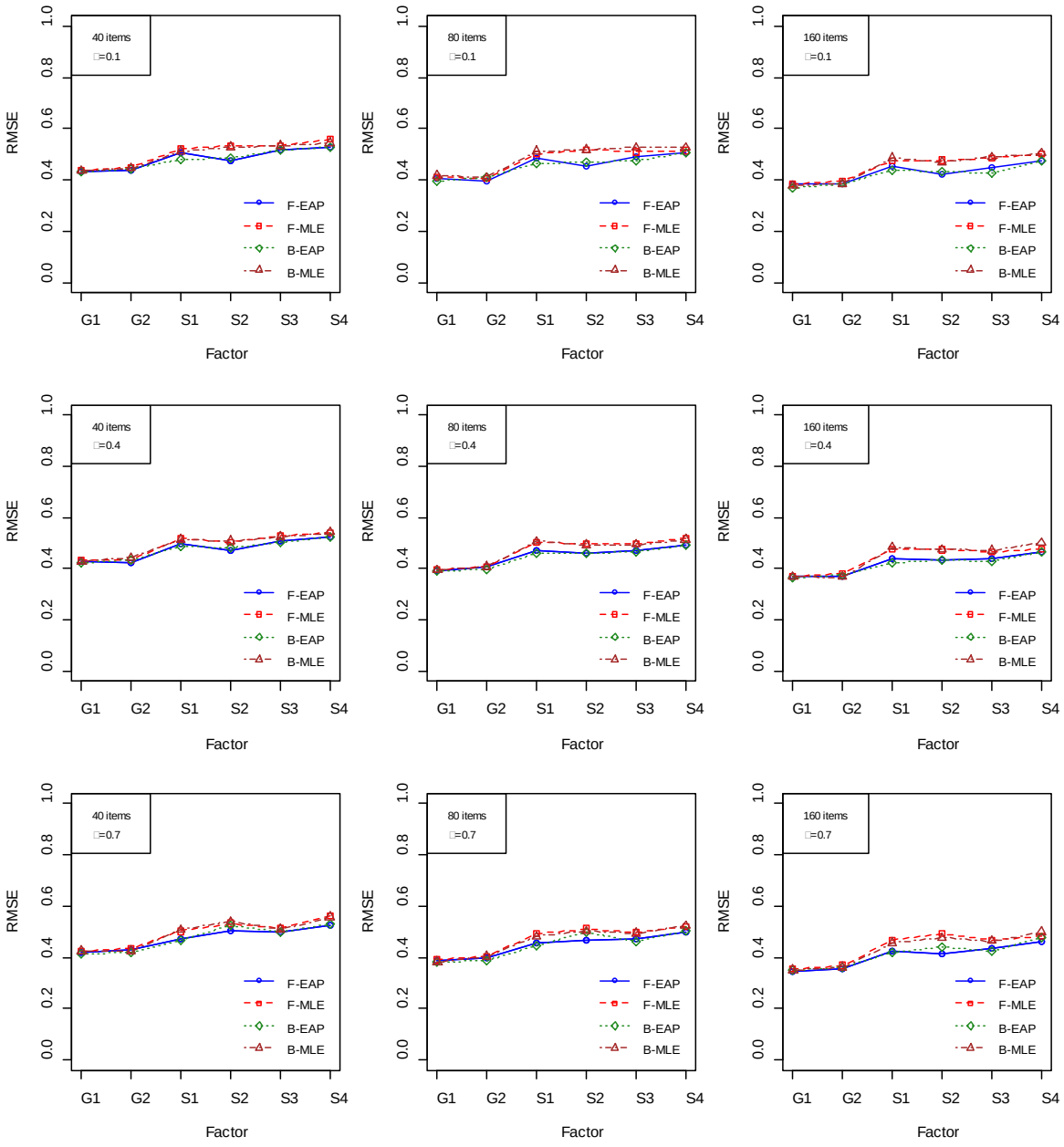


Figure B.5: Average RMSE (Higher-order IRT model (2 primary factors) with two group factors)

Note. F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method

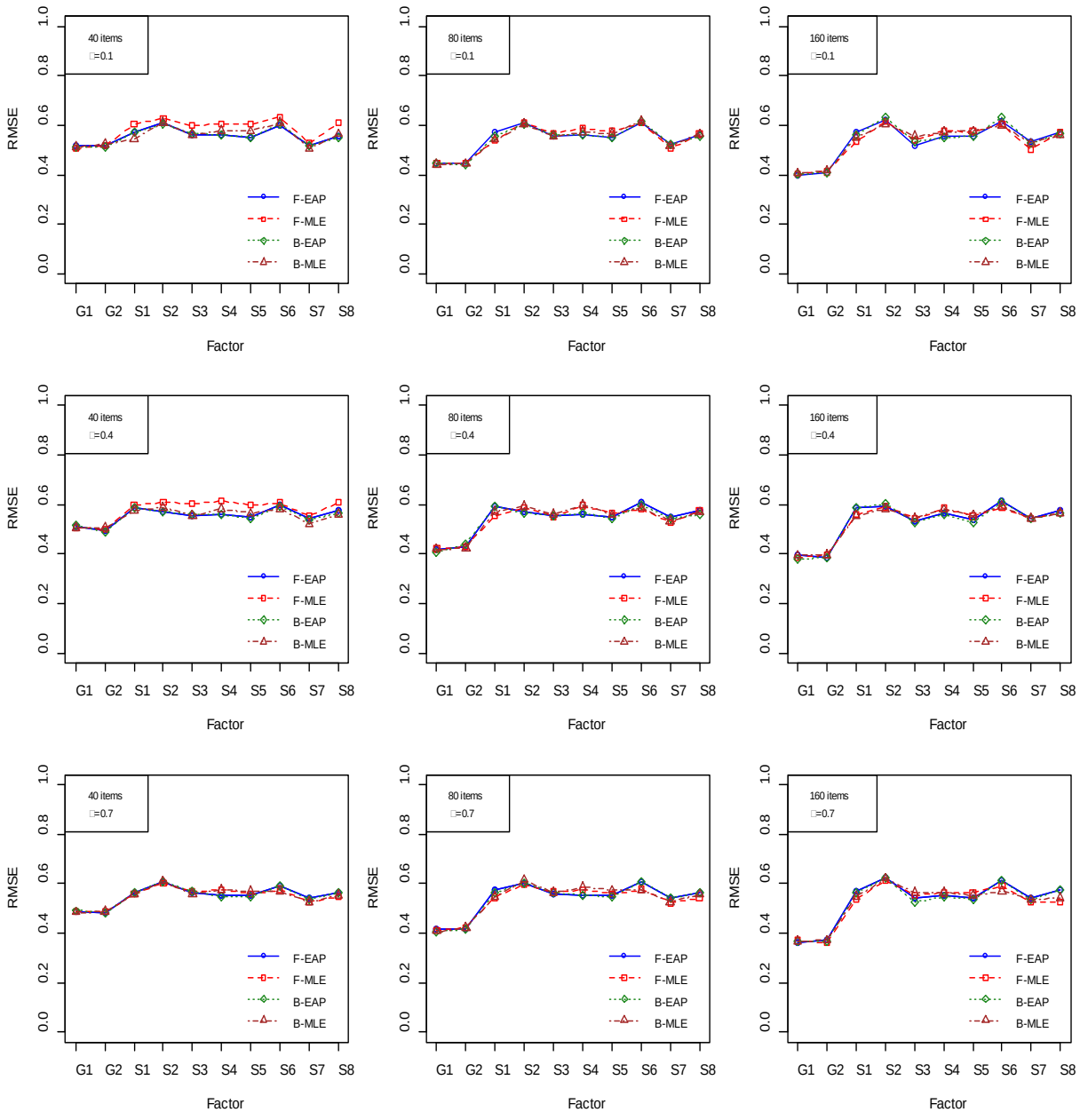


Figure B.6: Average RMSE (Higher-order IRT model (2 primary factors) with four group factors)

Note. F-EAP: Fisher item selection method – EAP scoring method
B-MLE: Bayesian item selection method – MLE scoring method

APPENDIX C

Average Standard Error (SE)

Table C.1: Average SE (Bifactor IRT Model)

Test Length	Item Selection	Scoring	Number of group factors							
			Two group factors			Four group factors				
			G_1	s_1	s_2	G_1	s_1	s_2	s_3	s_4
40										
	MFI	MLE	0.503	0.801	0.800	0.489	0.904	0.916	0.975	0.968
		EAP	0.495	0.788	0.795	0.477	0.897	0.909	0.923	0.970
	Bayes	MLE	0.500	0.799	0.802	0.510	0.917	0.974	0.953	0.967
		EAP	0.496	0.783	0.789	0.477	0.899	0.905	0.943	0.959
80										
	MFI	MLE	0.381	0.765	0.772	0.362	0.872	0.878	0.874	0.879
		EAP	0.370	0.742	0.758	0.322	0.871	0.846	0.830	0.847
	Bayes	MLE	0.382	0.766	0.772	0.351	0.820	0.876	0.875	0.877
		EAP	0.370	0.742	0.758	0.322	0.871	0.846	0.830	0.847
160										
	MFI	MLE	0.334	0.735	0.743	0.245	0.815	0.823	0.817	0.826
		EAP	0.331	0.710	0.725	0.254	0.792	0.769	0.765	0.766
	Bayes	MLE	0.334	0.734	0.743	0.246	0.761	0.824	0.818	0.822
		EAP	0.331	0.710	0.725	0.254	0.792	0.769	0.765	0.766

Table C.2: Average SE (Higher-order IRT Model)

Test Length	Item Selection	Scoring	Number of group factors							
			Two group factors			Four group factors				
			G_1	s_1	s_2	G_1	s_1	s_2	s_3	s_4
40										
	MFI	MLE	0.481	0.846	0.787	0.459	0.861	0.820	0.816	0.865
		EAP	0.483	0.762	0.761	0.460	0.806	0.793	0.826	0.839
	Bayes	MLE	0.508	0.833	0.838	0.490	0.841	0.858	0.847	0.880
		EAP	0.476	0.756	0.708	0.476	0.785	0.794	0.851	0.793
80										
	MFI	MLE	0.423	0.560	0.560	0.392	0.624	0.622	0.668	0.628
		EAP	0.413	0.542	0.505	0.358	0.615	0.612	0.647	0.609
	Bayes	MLE	0.424	0.560	0.560	0.372	0.624	0.620	0.667	0.629
		EAP	0.423	0.523	0.496	0.378	0.588	0.585	0.642	0.584
160										
	MFI	MLE	0.335	0.425	0.423	0.296	0.573	0.567	0.617	0.577
		EAP	0.330	0.409	0.387	0.297	0.534	0.542	0.553	0.552
	Bayes	MLE	0.335	0.424	0.425	0.296	0.572	0.566	0.616	0.578
		EAP	0.335	0.402	0.388	0.315	0.517	0.517	0.552	0.530

Table C.3: Average SE (Two-tier IRT Model with two group factors)

Test Length	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
	Low correlation b/t two general factors							
	MFI	MLE	0.565	0.635	0.910	0.945	0.950	0.958
		EAP	0.559	0.636	0.941	0.926	0.921	0.916
	Bayes	MLE	0.581	0.634	0.974	0.968	0.940	0.947
		EAP	0.559	0.626	0.952	0.926	0.920	0.916
40	*Medium correlation b/t two general factors*							
	MFI	MLE	0.547	0.602	0.954	0.943	0.920	0.974
		EAP	0.544	0.607	0.887	0.873	0.902	0.949
	Bayes	MLE	0.542	0.610	0.952	0.896	0.904	0.975
		EAP	0.539	0.619	0.913	0.914	0.903	0.972
	High correlation b/t two general factors							
	MFI	MLE	0.525	0.573	0.958	0.931	0.966	0.890
		EAP	0.518	0.565	0.933	0.917	0.903	0.924
	Bayes	MLE	0.525	0.564	0.958	0.923	0.909	0.976
		EAP	0.513	0.558	0.943	0.956	0.902	0.940
	Low correlation b/t two general factors							
	MFI	MLE	0.450	0.445	0.881	0.885	0.882	0.875
		EAP	0.442	0.421	0.862	0.869	0.860	0.900
	Bayes	MLE	0.451	0.446	0.883	0.887	0.880	0.874
		EAP	0.442	0.421	0.863	0.870	0.861	0.900
80	*Medium correlation b/t two general factors*							
	MFI	MLE	0.439	0.434	0.880	0.883	0.881	0.874
		EAP	0.432	0.413	0.859	0.868	0.858	0.896
	Bayes	MLE	0.439	0.435	0.879	0.884	0.881	0.874
		EAP	0.421	0.407	0.875	0.872	0.868	0.897
	High correlation b/t two general factors							
	MFI	MLE	0.402	0.401	0.879	0.881	0.877	0.871
		EAP	0.403	0.389	0.850	0.863	0.853	0.891
	Bayes	MLE	0.402	0.398	0.877	0.882	0.878	0.871
		EAP	0.386	0.377	0.889	0.872	0.884	0.893
	Low correlation b/t two general factors							
	MFI	MLE	0.373	0.373	0.835	0.840	0.836	0.827
		EAP	0.361	0.347	0.801	0.813	0.820	0.812
	Bayes	MLE	0.376	0.371	0.837	0.841	0.835	0.827
		EAP	0.361	0.347	0.801	0.813	0.821	0.813
160	*Medium correlation b/t two general factors*							
	MFI	MLE	0.364	0.364	0.833	0.840	0.833	0.826
		EAP	0.355	0.342	0.796	0.811	0.819	0.808
	Bayes	MLE	0.255	0.242	0.746	0.761	0.769	0.758
		EAP	0.347	0.337	0.806	0.814	0.825	0.822
	High correlation b/t two general factors							
	MFI	MLE	0.316	0.320	0.830	0.835	0.831	0.820
		EAP	0.335	0.324	0.785	0.803	0.811	0.800
	Bayes	MLE	0.307	0.335	0.827	0.834	0.830	0.821
		EAP	0.319	0.314	0.813	0.812	0.825	0.826

Table C.4: Average SE (Two-tier IRT Model with four group factors)

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
40												
Low correlation b/t two general factors												
	MFI	MLE	0.647	0.683	0.997	0.996	0.979	0.975	0.995	0.986	0.961	0.997
		EAP	0.646	0.684	0.988	0.966	0.946	0.923	0.956	0.993	0.990	0.954
	Bayes	MLE	0.657	0.687	0.965	0.983	0.980	0.953	0.959	0.970	0.975	0.956
		EAP	0.636	0.679	0.996	0.979	0.991	0.959	0.955	0.993	0.989	0.953
Medium correlation b/t two general factors												
	MFI	MLE	0.625	0.614	0.964	0.973	0.969	0.987	0.987	0.973	0.927	0.991
		EAP	0.622	0.605	0.977	0.962	0.936	0.938	0.922	0.971	0.979	0.953
	Bayes	MLE	0.613	0.602	0.969	0.997	0.992	0.979	0.959	0.960	0.991	0.978
		EAP	0.611	0.605	0.979	0.961	0.935	0.935	0.922	0.971	0.979	0.953
High correlation b/t two general factors												
	MFI	MLE	0.588	0.570	0.940	0.986	0.988	0.969	0.987	0.976	0.987	0.959
		EAP	0.607	0.579	0.974	0.965	0.941	0.946	0.927	0.969	0.980	0.903
	Bayes	MLE	0.582	0.569	0.975	0.990	0.996	0.993	0.937	0.950	0.968	0.984
		EAP	0.571	0.564	0.970	0.988	0.990	0.957	0.995	0.969	0.982	0.975
80												
Low correlation b/t two general factors												
	MFI	MLE	0.502	0.579	0.968	0.971	0.982	0.985	0.981	0.983	0.978	0.984
		EAP	0.485	0.560	0.964	0.951	0.982	0.962	0.933	0.979	0.982	0.957
	Bayes	MLE	0.503	0.579	0.967	0.971	0.983	0.985	0.982	0.984	0.978	0.985
		EAP	0.482	0.559	0.970	0.953	0.975	0.960	0.933	0.980	0.982	0.958
Medium correlation b/t two general factors												
	MFI	MLE	0.491	0.560	0.968	0.971	0.983	0.984	0.982	0.983	0.978	0.983
		EAP	0.477	0.543	0.962	0.953	0.982	0.963	0.939	0.975	0.981	0.956
	Bayes	MLE	0.491	0.561	0.968	0.972	0.983	0.984	0.982	0.984	0.979	0.983
		EAP	0.473	0.528	0.969	0.958	0.974	0.961	0.972	0.981	0.982	0.960
High correlation b/t two general factors												
	MFI	MLE	0.463	0.517	0.967	0.971	0.983	0.985	0.982	0.982	0.979	0.983
		EAP	0.452	0.496	0.961	0.951	0.982	0.962	0.950	0.973	0.980	0.956
	Bayes	MLE	0.463	0.515	0.968	0.971	0.984	0.985	0.981	0.984	0.979	0.984
		EAP	0.441	0.481	0.970	0.965	0.983	0.972	0.990	0.981	0.981	0.960

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
160												
Low correlation b/t two general factors												
	MFI	MLE	0.386	0.470	0.954	0.959	0.956	0.958	0.966	0.969	0.960	0.970
		EAP	0.378	0.436	0.937	0.929	0.926	0.913	0.918	0.953	0.956	0.927
	Bayes	MLE	0.385	0.470	0.954	0.959	0.955	0.958	0.966	0.969	0.959	0.970
		EAP	0.377	0.436	0.941	0.931	0.925	0.913	0.918	0.954	0.957	0.928
Medium correlation b/t two general factors												
	MFI	MLE	0.374	0.454	0.953	0.957	0.955	0.958	0.965	0.969	0.959	0.969
		EAP	0.374	0.425	0.935	0.930	0.927	0.914	0.917	0.951	0.956	0.926
	Bayes	MLE	0.374	0.454	0.953	0.958	0.954	0.957	0.966	0.970	0.959	0.970
		EAP	0.370	0.417	0.942	0.935	0.927	0.916	0.932	0.960	0.959	0.930
High correlation b/t two general factors												
	MFI	MLE	0.348	0.410	0.952	0.956	0.956	0.959	0.965	0.968	0.959	0.969
		EAP	0.354	0.394	0.934	0.928	0.925	0.912	0.919	0.950	0.954	0.924
	Bayes	MLE	0.347	0.411	0.952	0.957	0.954	0.958	0.966	0.968	0.959	0.971
		EAP	0.347	0.385	0.942	0.940	0.932	0.920	0.943	0.965	0.957	0.930

Table C.5: Average SE (Higher-order IRT Model (2 primary factors) with two group factors)

Test Length	Item Selection	Scoring	Two group factors					
			G_1	G_2	s_1	s_2	s_3	s_4
	Low correlation b/t two general factors							
	MFI	MLE	0.541	0.592	0.901	0.961	0.872	0.986
		EAP	0.552	0.582	0.847	0.878	0.871	0.957
	Bayes	MLE	0.567	0.594	0.896	0.947	0.880	0.918
		EAP	0.582	0.585	0.819	0.885	0.845	0.961
40	*Medium correlation b/t two general factors*							
	MFI	MLE	0.540	0.574	0.854	0.959	0.891	0.970
		EAP	0.535	0.584	0.872	0.849	0.846	0.938
	Bayes	MLE	0.546	0.570	0.823	0.981	0.920	0.973
		EAP	0.559	0.577	0.834	0.885	0.849	0.935
	High correlation b/t two general factors							
	MFI	MLE	0.528	0.565	0.827	0.941	0.905	0.982
		EAP	0.525	0.565	0.834	0.871	0.892	0.903
	Bayes	MLE	0.538	0.549	0.835	0.920	0.939	0.977
		EAP	0.523	0.561	0.830	0.903	0.889	0.902
	Low correlation b/t two general factors							
	MFI	MLE	0.493	0.495	0.624	0.683	0.610	0.689
		EAP	0.494	0.476	0.607	0.647	0.615	0.695
	Bayes	MLE	0.500	0.494	0.623	0.681	0.610	0.688
		EAP	0.504	0.488	0.581	0.663	0.597	0.695
80	*Medium correlation b/t two general factors*							
	MFI	MLE	0.491	0.488	0.627	0.679	0.617	0.689
		EAP	0.487	0.472	0.609	0.643	0.626	0.697
	Bayes	MLE	0.493	0.488	0.628	0.681	0.618	0.690
		EAP	0.487	0.480	0.593	0.666	0.612	0.699
	High correlation b/t two general factors							
	MFI	MLE	0.453	0.458	0.629	0.678	0.622	0.689
		EAP	0.456	0.455	0.604	0.636	0.631	0.693
	Bayes	MLE	0.452	0.460	0.629	0.679	0.623	0.690
		EAP	0.447	0.455	0.595	0.675	0.626	0.700
	Low correlation b/t two general factors							
	MFI	MLE	0.401	0.401	0.548	0.610	0.535	0.624
		EAP	0.409	0.403	0.552	0.578	0.565	0.647
	Bayes	MLE	0.400	0.401	0.572	0.633	0.560	0.647
		EAP	0.421	0.417	0.525	0.596	0.536	0.649
160	*Medium correlation b/t two general factors*							
	MFI	MLE	0.418	0.417	0.574	0.632	0.571	0.650
		EAP	0.403	0.402	0.555	0.575	0.577	0.648
	Bayes	MLE	0.419	0.417	0.576	0.633	0.570	0.650
		EAP	0.409	0.412	0.535	0.596	0.555	0.654
	High correlation b/t two general factors							
	MFI	MLE	0.386	0.396	0.570	0.629	0.577	0.648
		EAP	0.381	0.391	0.546	0.566	0.580	0.644
	Bayes	MLE	0.371	0.375	0.570	0.629	0.576	0.648
		EAP	0.380	0.393	0.533	0.599	0.568	0.657

Table C.6: Average SE (Higher-order IRT Model (2 primary factors) with four group factors)

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
40												
Low correlation b/t two general factors												
	MFI	MLE	0.647	0.666	0.954	0.979	0.980	0.990	0.985	0.951	0.996	0.974
		EAP	0.648	0.689	0.926	0.950	0.927	0.956	0.917	0.941	0.977	0.915
	Bayes	MLE	0.654	0.673	0.963	0.969	0.980	0.989	0.939	0.979	0.994	0.936
		EAP	0.650	0.689	0.921	0.952	0.927	0.951	0.925	0.945	0.919	0.883
Medium correlation b/t two general factors												
	MFI	MLE	0.644	0.653	0.983	0.969	0.980	0.989	0.939	0.979	0.994	0.936
		EAP	0.636	0.649	0.948	0.939	0.897	0.913	0.967	0.914	0.930	0.912
	Bayes	MLE	0.647	0.656	0.954	0.979	0.980	0.990	0.985	0.951	0.996	0.974
		EAP	0.636	0.645	0.948	0.939	0.897	0.913	0.967	0.914	0.930	0.912
High correlation b/t two general factors												
	MFI	MLE	0.595	0.571	0.958	0.963	0.947	0.985	0.989	0.979	0.981	0.938
		EAP	0.588	0.562	0.951	0.936	0.893	0.917	0.981	0.912	0.976	0.908
	Bayes	MLE	0.591	0.576	0.965	0.972	0.952	0.989	0.983	0.977	0.978	0.937
		EAP	0.577	0.561	0.951	0.935	0.894	0.975	0.982	0.911	0.980	0.908
80												
Low correlation b/t two general factors												
	MFI	EAP	0.556	0.613	0.699	0.725	0.727	0.752	0.747	0.731	0.747	0.724
		MLE	0.559	0.624	0.690	0.683	0.691	0.703	0.699	0.712	0.743	0.677
	Bayes	EAP	0.566	0.632	0.699	0.724	0.726	0.752	0.747	0.731	0.748	0.724
		MLE	0.566	0.633	0.663	0.685	0.685	0.716	0.697	0.704	0.728	0.669
Medium correlation b/t two general factors												
	MFI	EAP	0.562	0.614	0.701	0.727	0.726	0.753	0.748	0.734	0.748	0.725
		MLE	0.555	0.602	0.689	0.689	0.689	0.703	0.712	0.716	0.736	0.681
	Bayes	EAP	0.561	0.613	0.701	0.725	0.725	0.753	0.747	0.734	0.748	0.724
		MLE	0.558	0.605	0.673	0.686	0.683	0.722	0.714	0.701	0.739	0.679
High correlation b/t two general factors												
	MFI	EAP	0.532	0.570	0.700	0.728	0.724	0.752	0.748	0.736	0.749	0.726
		MLE	0.533	0.555	0.686	0.690	0.682	0.706	0.731	0.711	0.738	0.679
	Bayes	EAP	0.534	0.570	0.701	0.728	0.723	0.753	0.749	0.737	0.749	0.726
		MLE	0.529	0.556	0.682	0.687	0.675	0.736	0.730	0.710	0.740	0.679

Test Length	Item Selection	Scoring	Four group factors									
			G_1	G_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
160												
Low correlation b/t two general factors												
	MFI	MLE	0.440	0.504	0.678	0.703	0.686	0.723	0.728	0.710	0.727	0.697
		EAP	0.456	0.508	0.655	0.659	0.629	0.671	0.688	0.687	0.713	0.647
	Bayes	MLE	0.452	0.524	0.677	0.704	0.686	0.724	0.727	0.711	0.726	0.697
		EAP	0.460	0.518	0.635	0.654	0.627	0.679	0.686	0.674	0.702	0.626
Medium correlation b/t two general factors												
	MFI	MLE	0.447	0.511	0.680	0.705	0.682	0.725	0.729	0.716	0.728	0.699
		EAP	0.451	0.498	0.655	0.662	0.624	0.673	0.690	0.692	0.711	0.651
	Bayes	MLE	0.449	0.512	0.680	0.706	0.682	0.723	0.730	0.715	0.727	0.701
		EAP	0.453	0.505	0.642	0.658	0.623	0.681	0.689	0.681	0.713	0.636
High correlation b/t two general factors												
	MFI	MLE	0.419	0.470	0.677	0.706	0.676	0.725	0.731	0.719	0.728	0.700
		EAP	0.434	0.471	0.650	0.662	0.609	0.675	0.691	0.690	0.709	0.647
	Bayes	MLE	0.418	0.469	0.677	0.705	0.675	0.724	0.730	0.719	0.727	0.701
		EAP	0.431	0.471	0.648	0.663	0.608	0.682	0.702	0.691	0.712	0.644

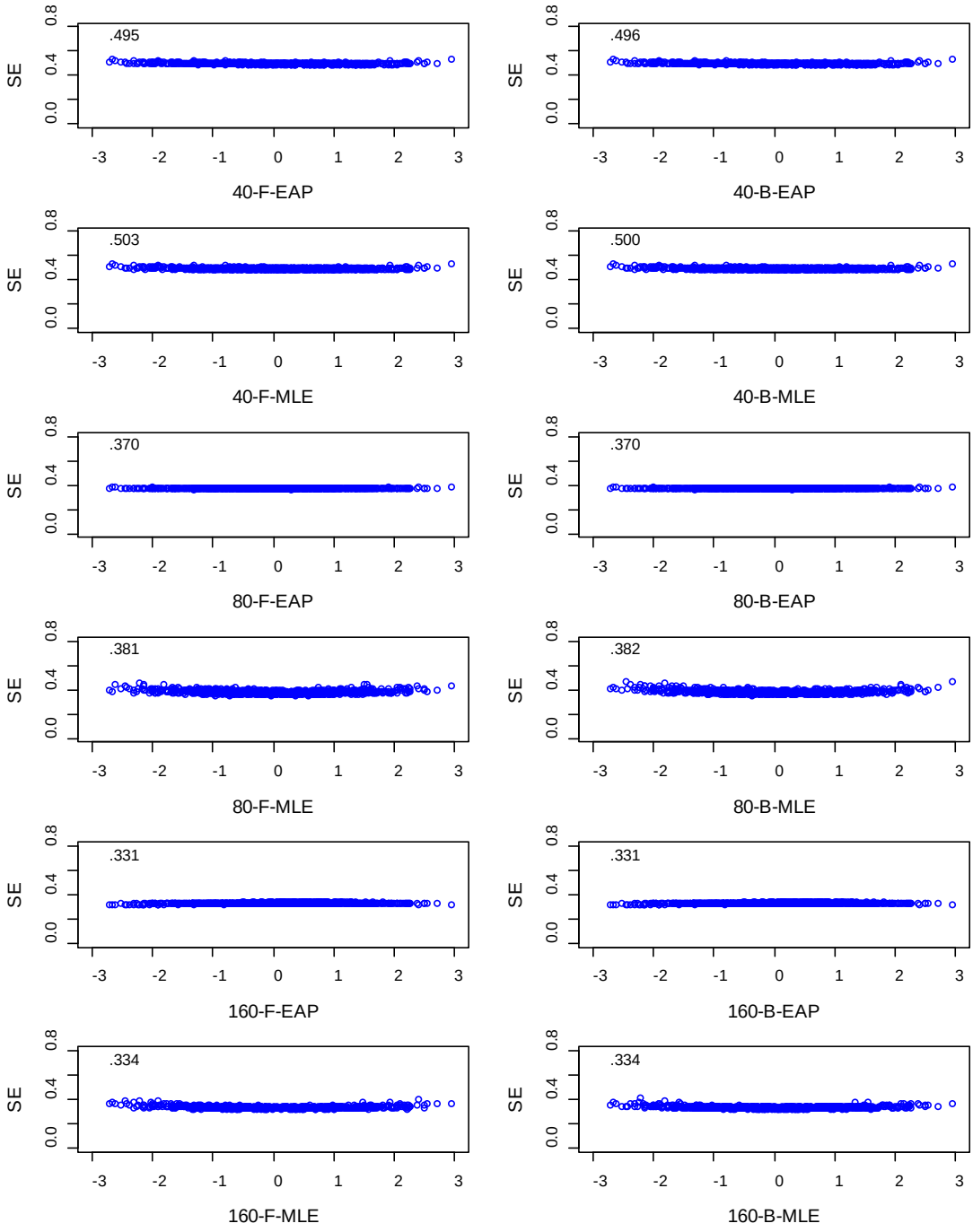


Figure C.1: Average SE (Primary factor for Bifactor IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

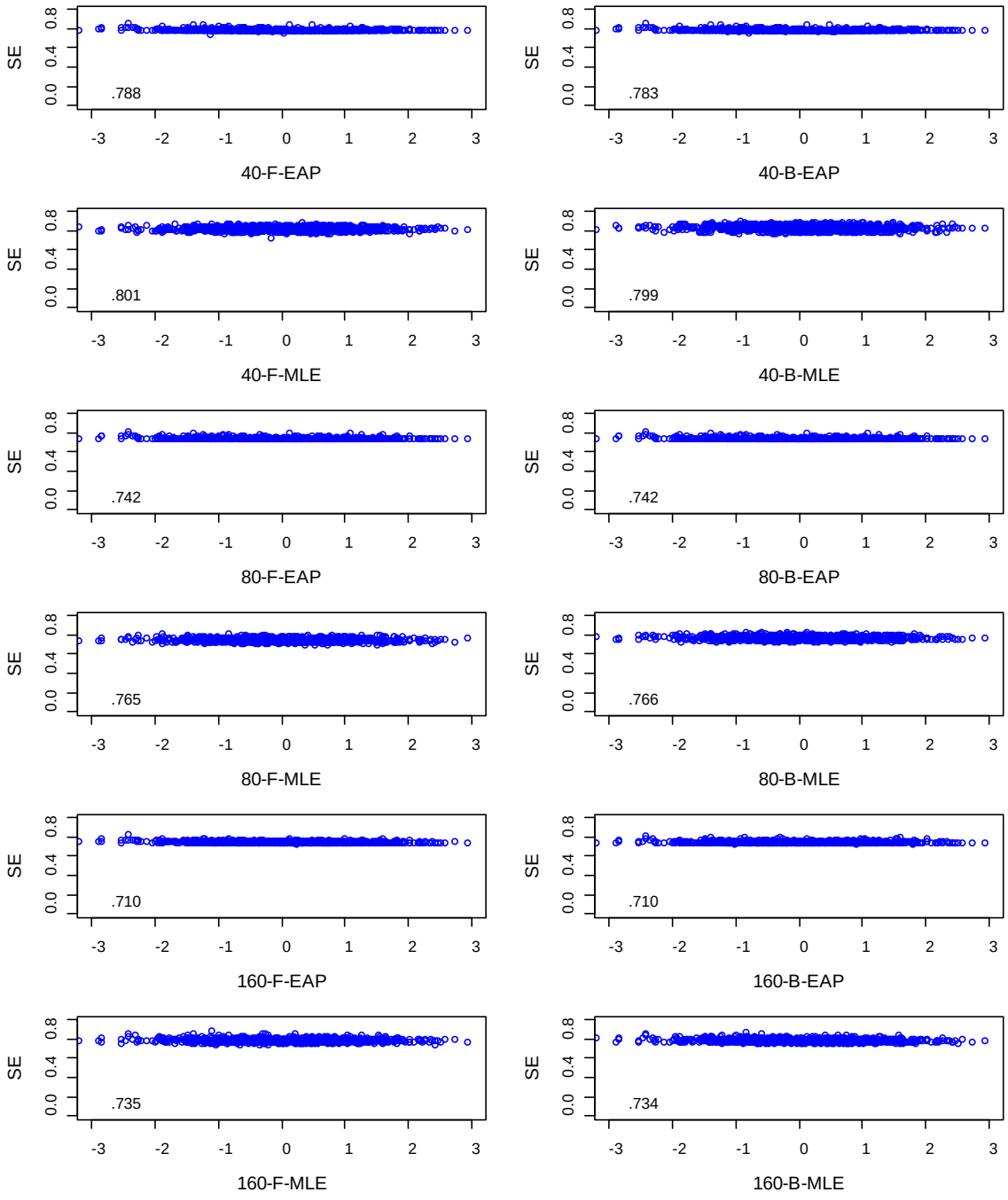


Figure C.2: Average SE (First group factor for Bifactor IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

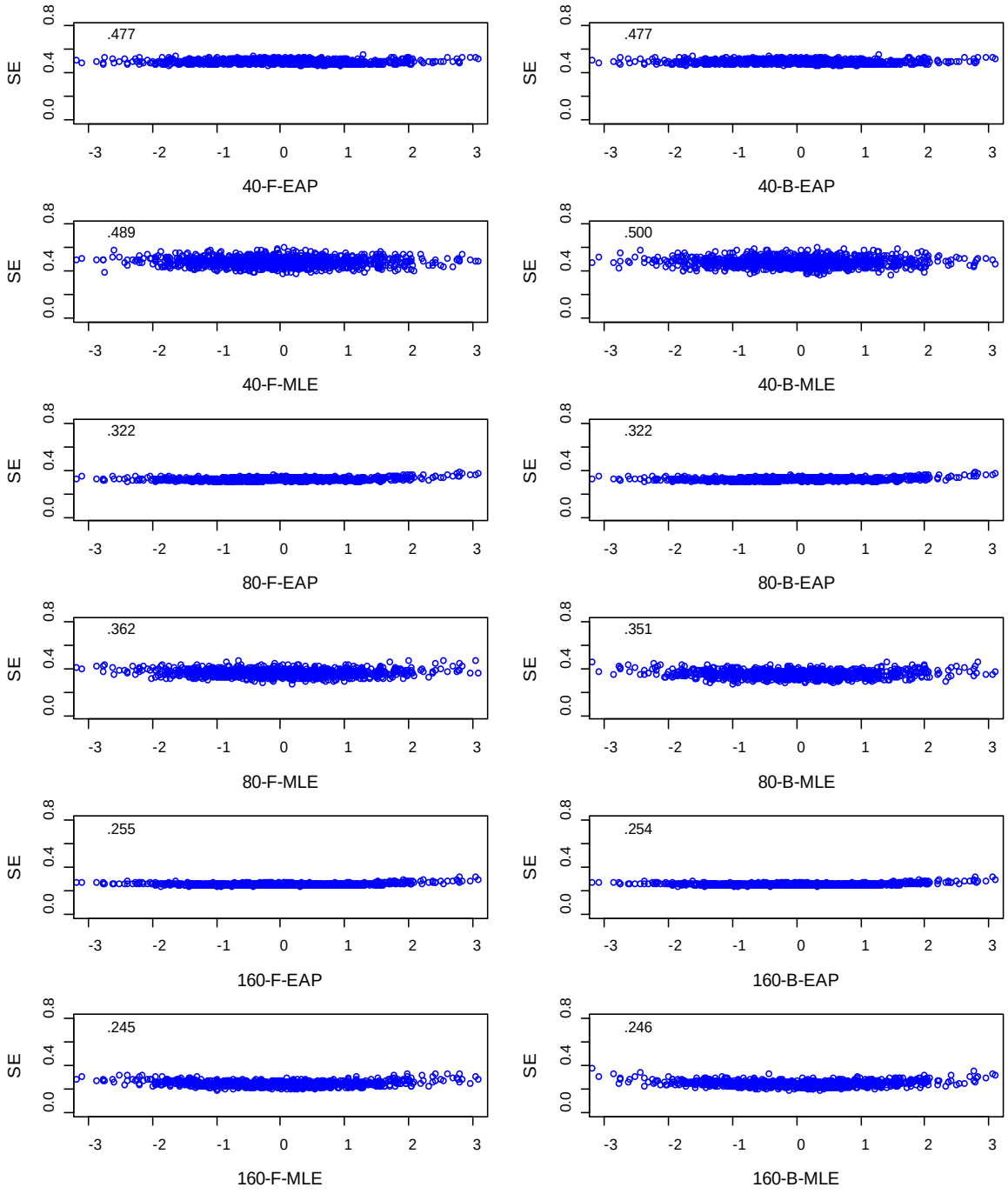


Figure C.3: Average SE (Primary factor for Bifactor IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

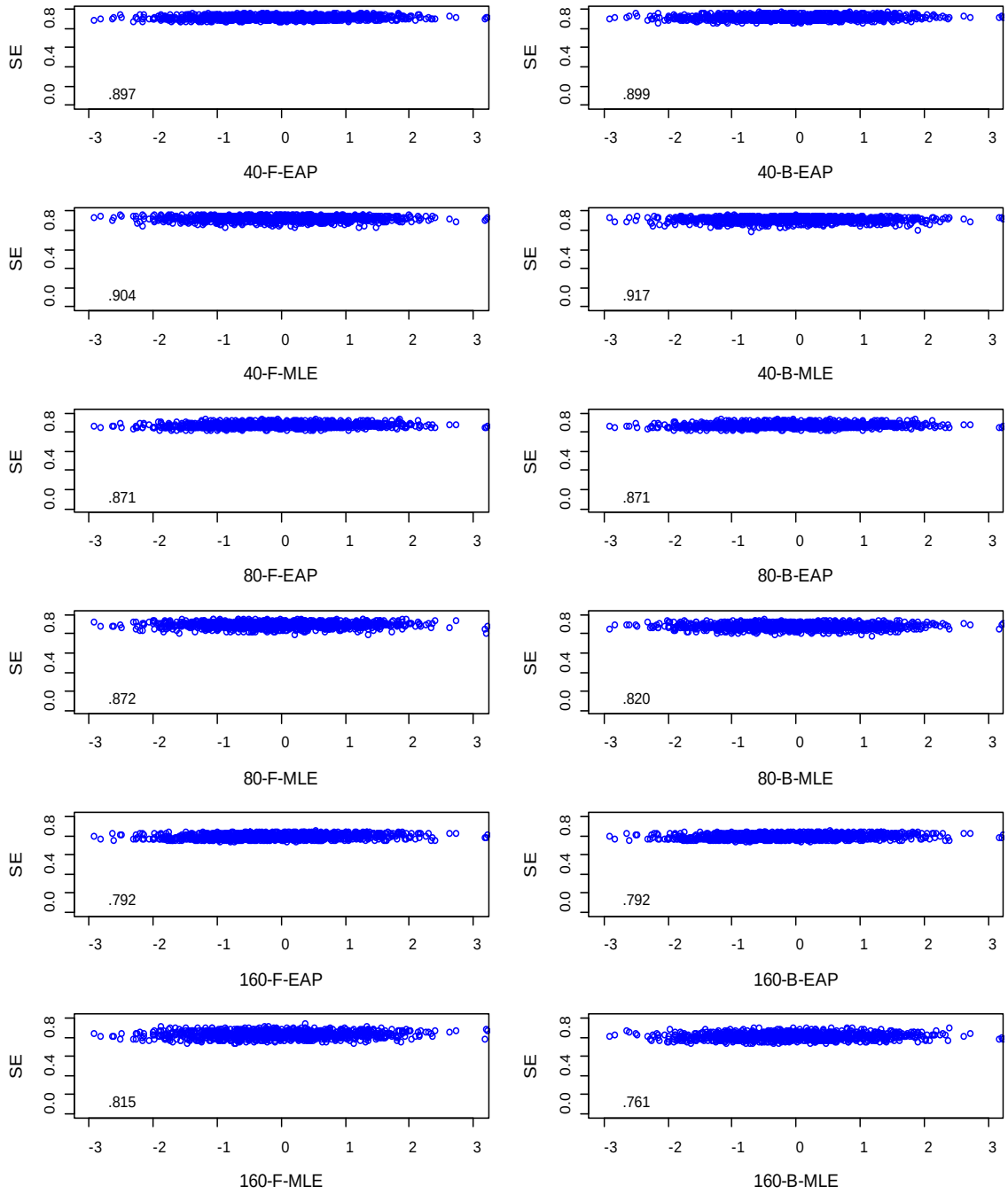


Figure C.4: Average SE (First group factor for Bifactor IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

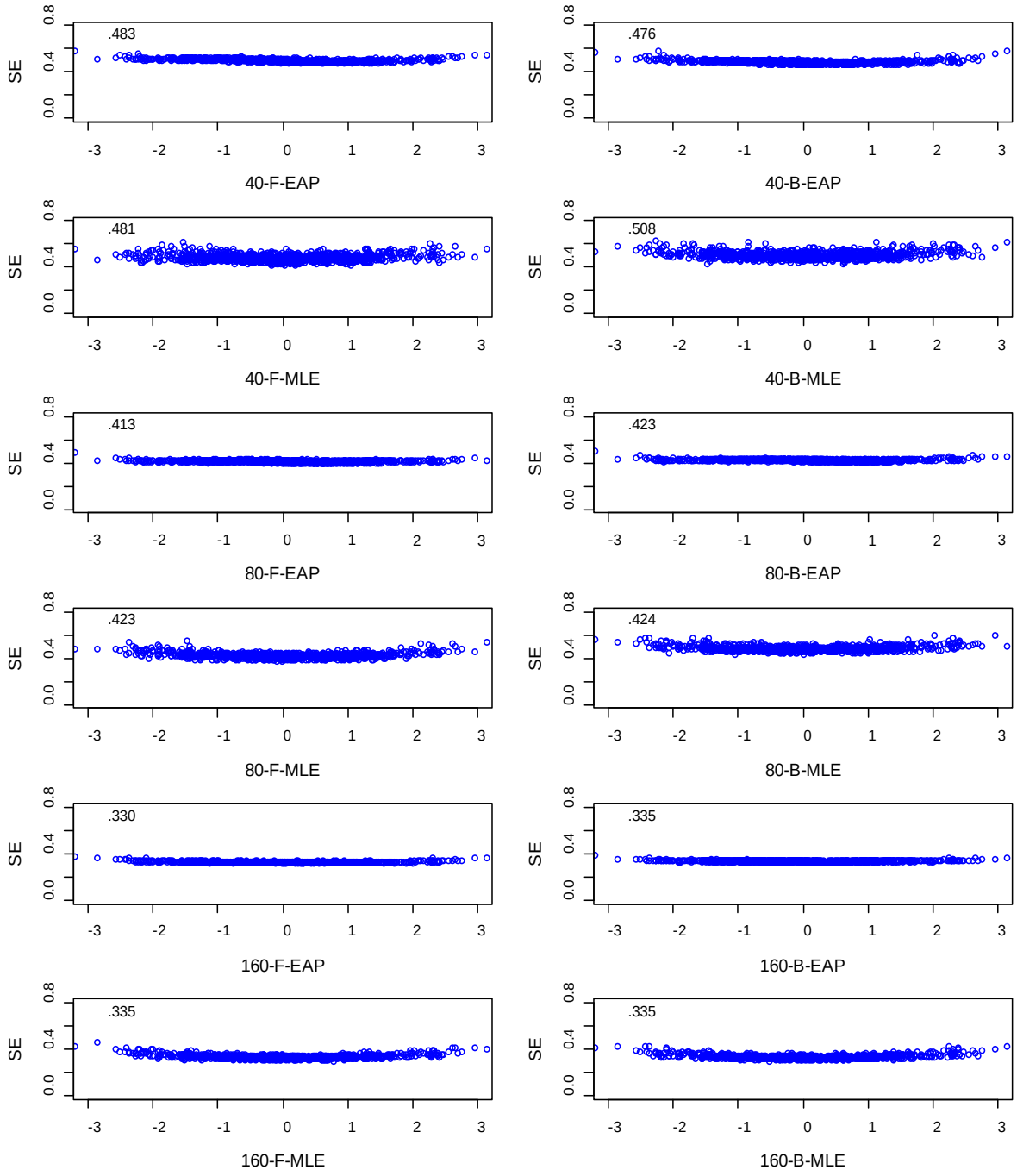


Figure C.5: Average SE (Primary factor for Higher-order IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

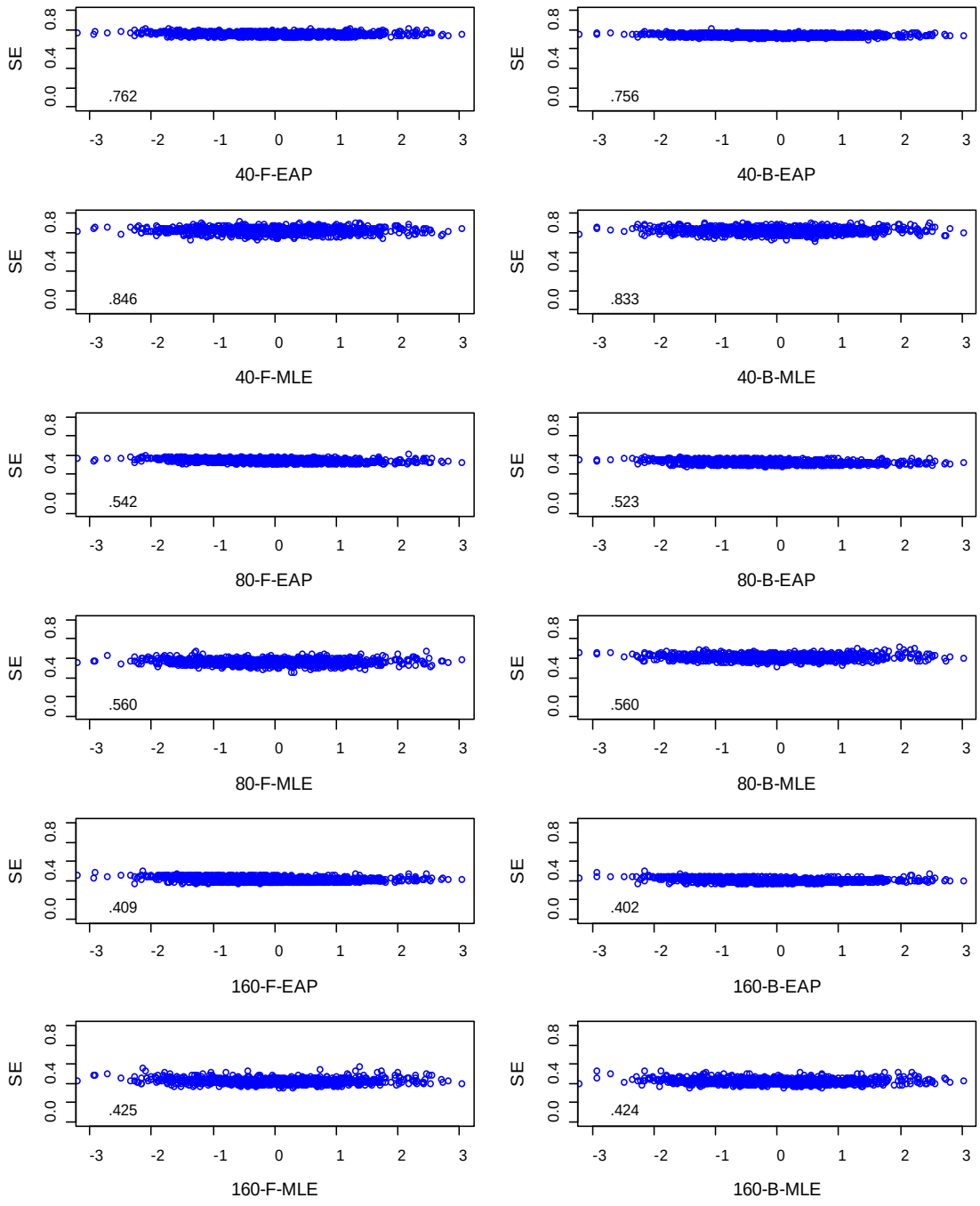


Figure C.6: Average SE (First group factor for Higher-order IRT model with two group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

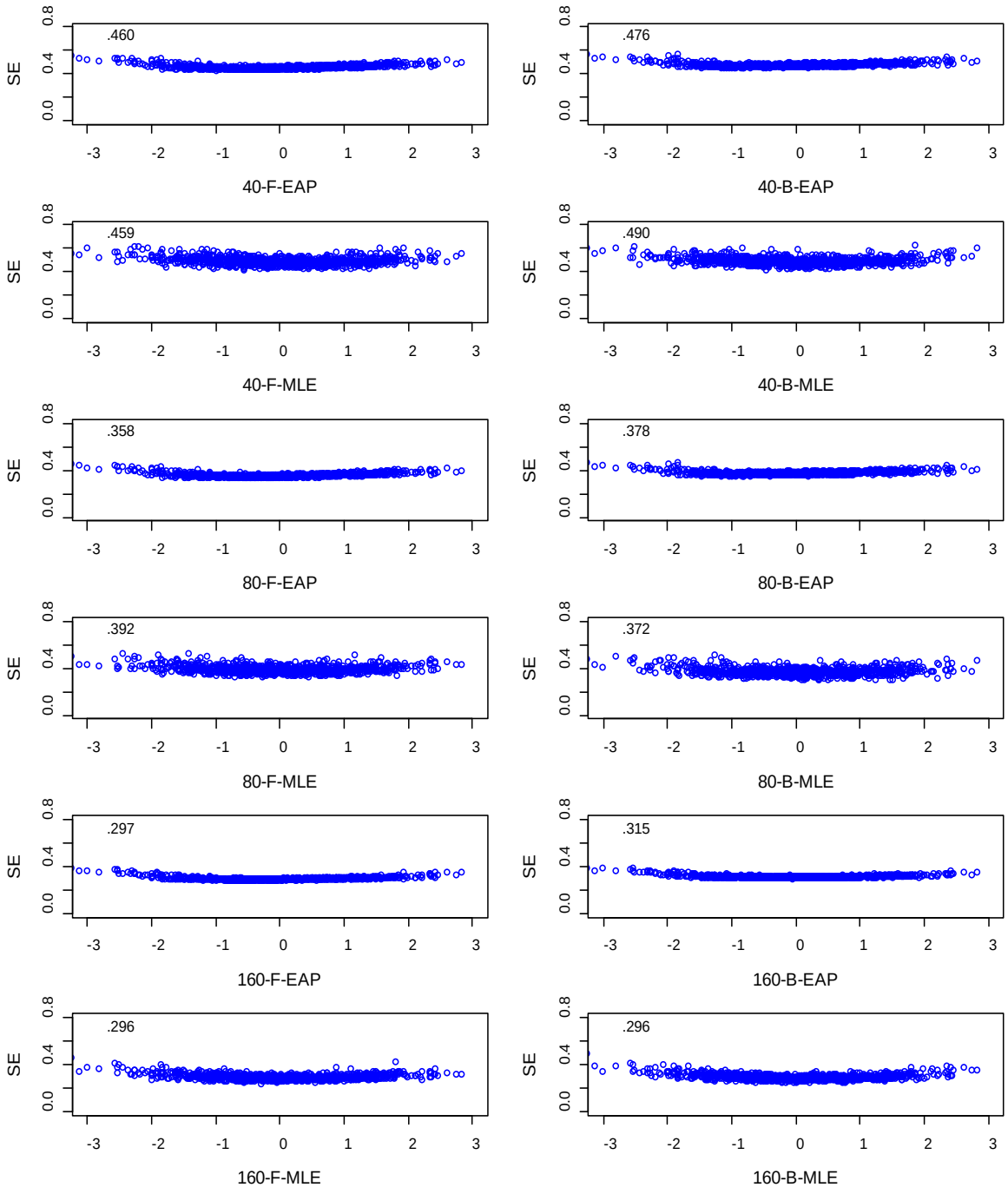


Figure C.7: Average SE (Primary factor for Higher-order IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

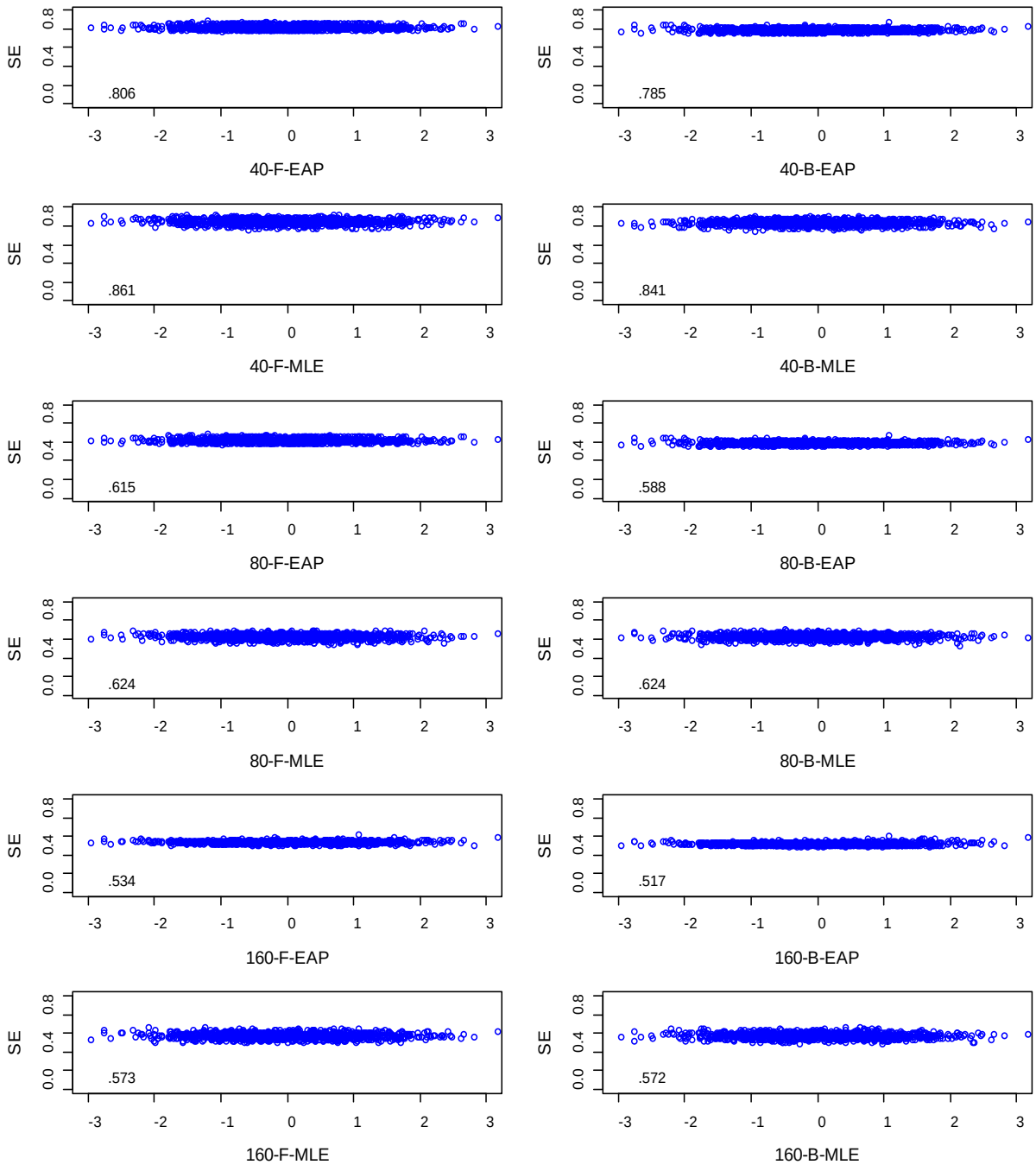


Figure C.8: Average SE (First group factor for Higher-order IRT model with four group factors)

Note. 40-F-EAP: 40 items – Fisher item selection method – EAP scoring method
160-B-MLE: 160 items – Bayesian item selection method – MLE scoring method

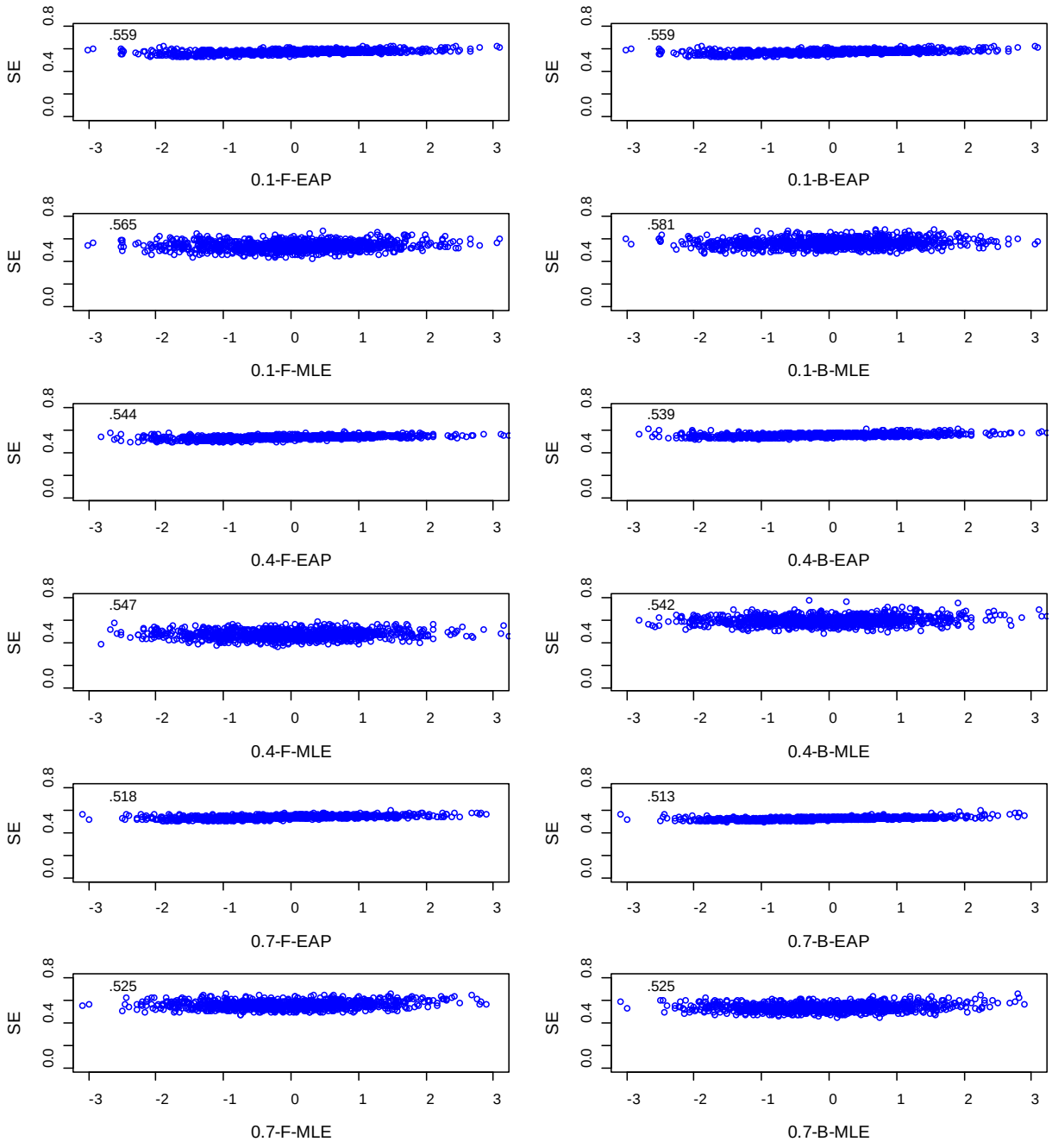


Figure C.9: Average SE (First primary factor for Two-tier IRT model with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

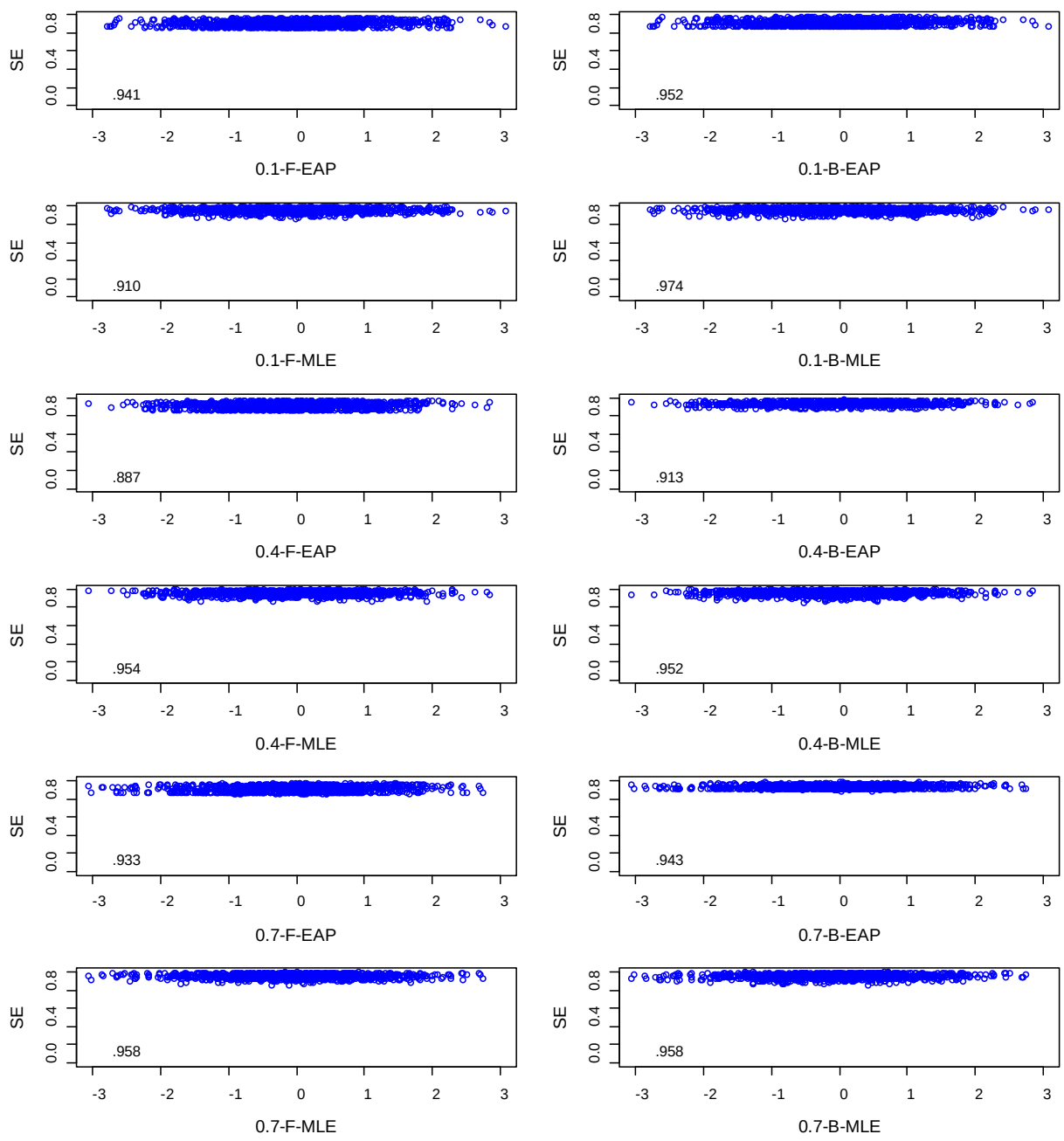


Figure C.10: Average SE (First group factor for Two-tier IRT model with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

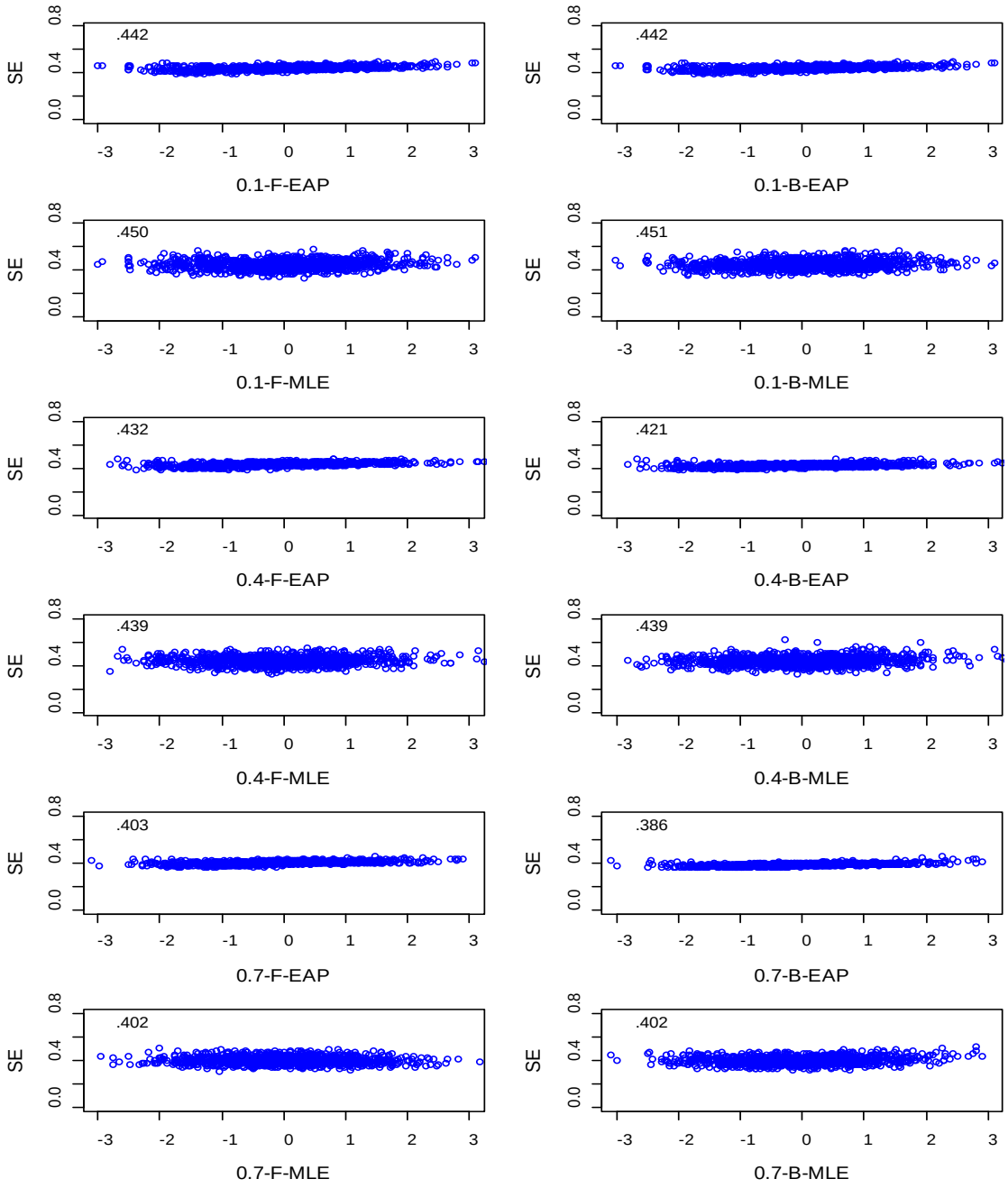


Figure C.11: Average SE (First primary factor for Two-tier IRT model with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

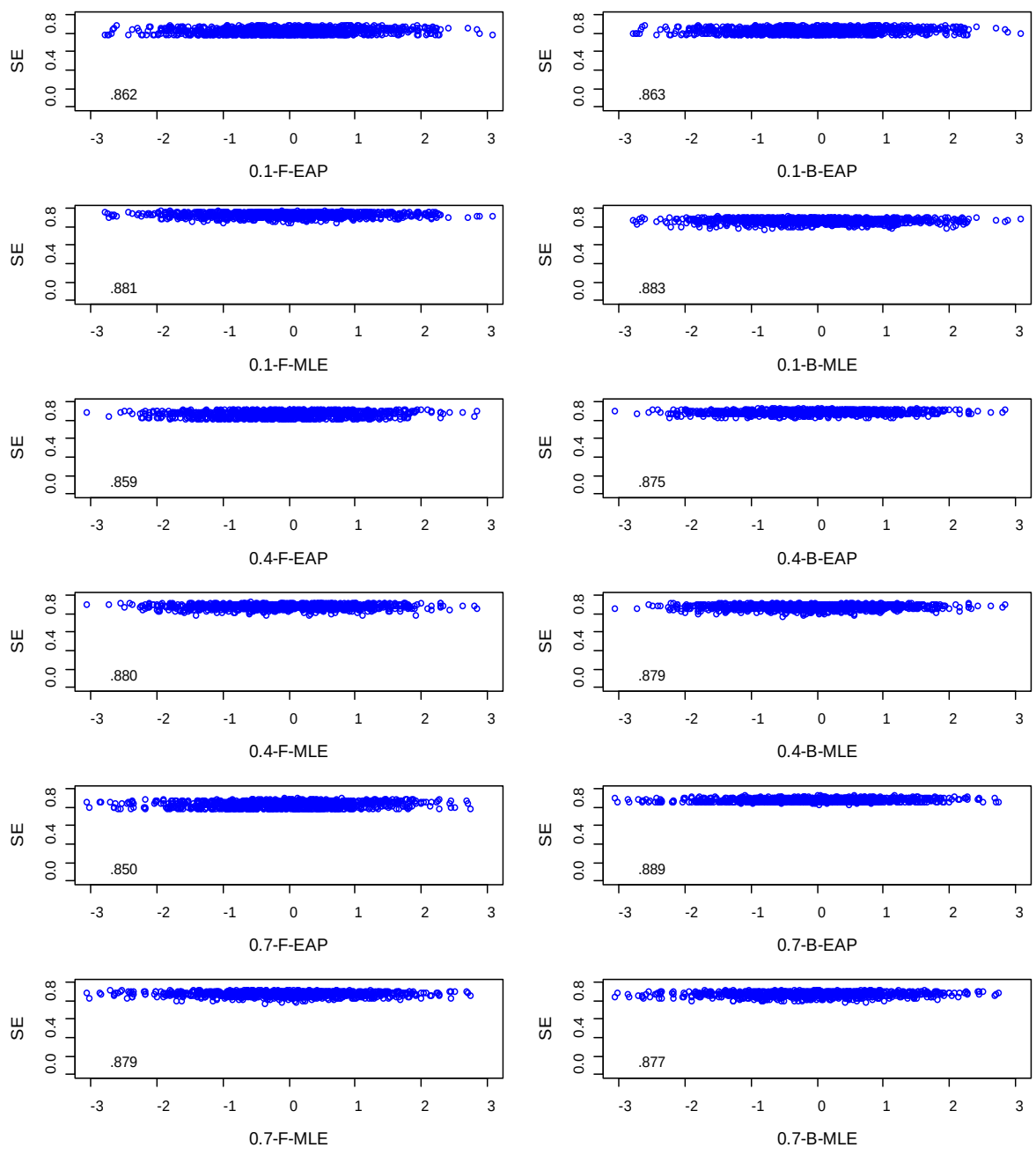


Figure C.12: Average SE (First group factor for Two-tier IRT model with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

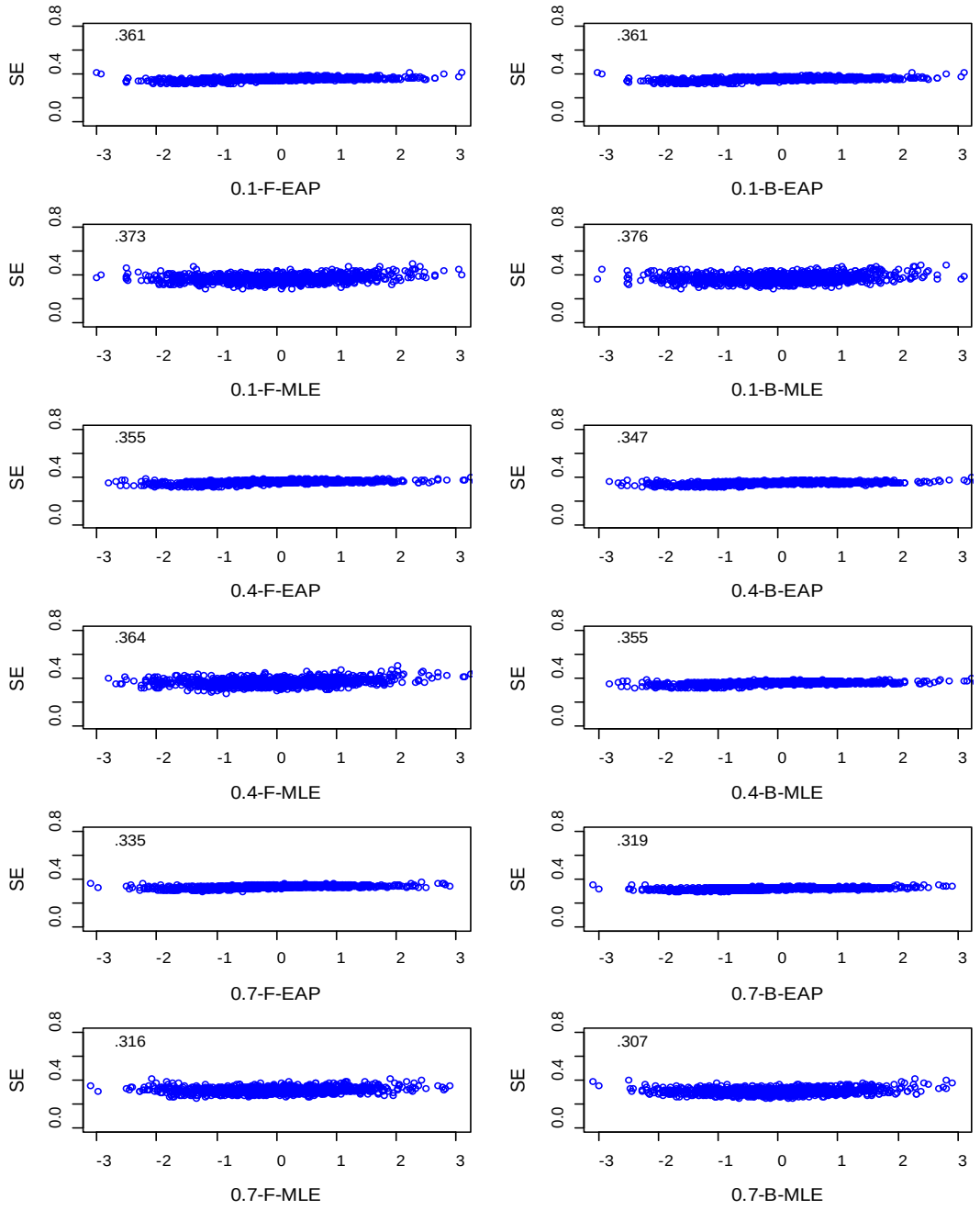


Figure C.13: Average SE (First primary factor for Two-tier IRT model with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

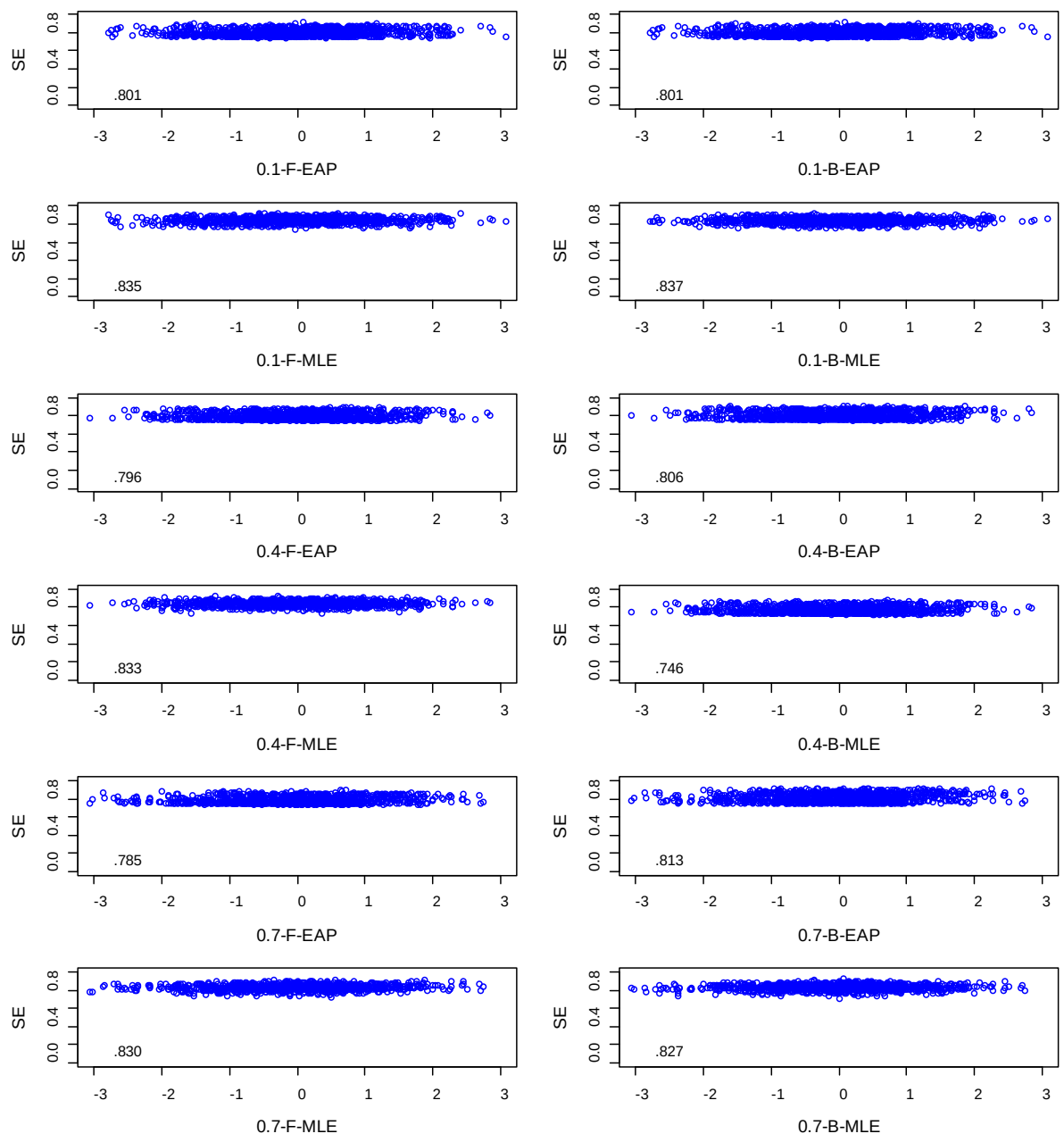


Figure C.14: Average SE (First group factor for Two-tier IRT model with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

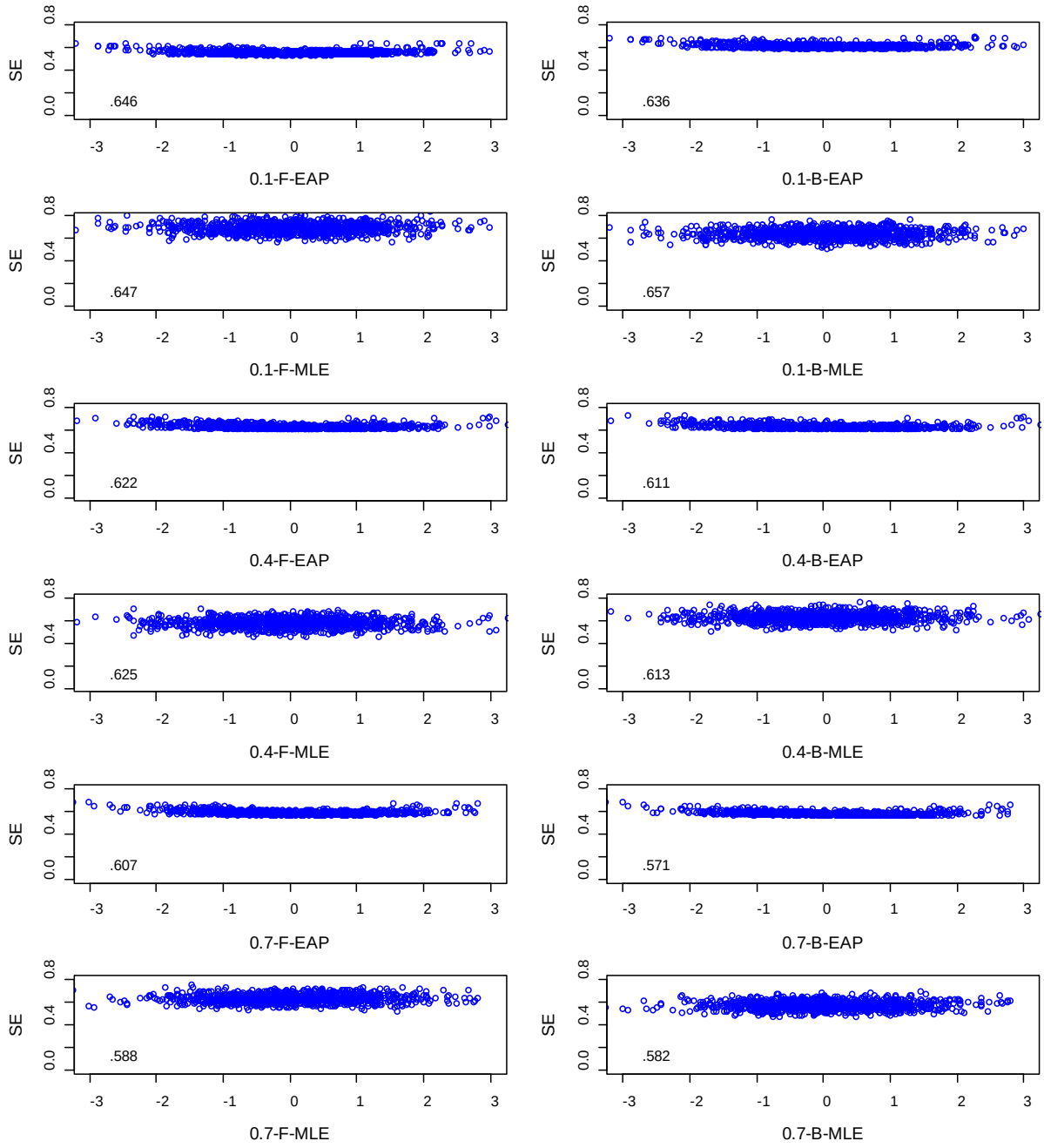


Figure C.15: Average SE (First primary factor for Two-tier IRT model with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

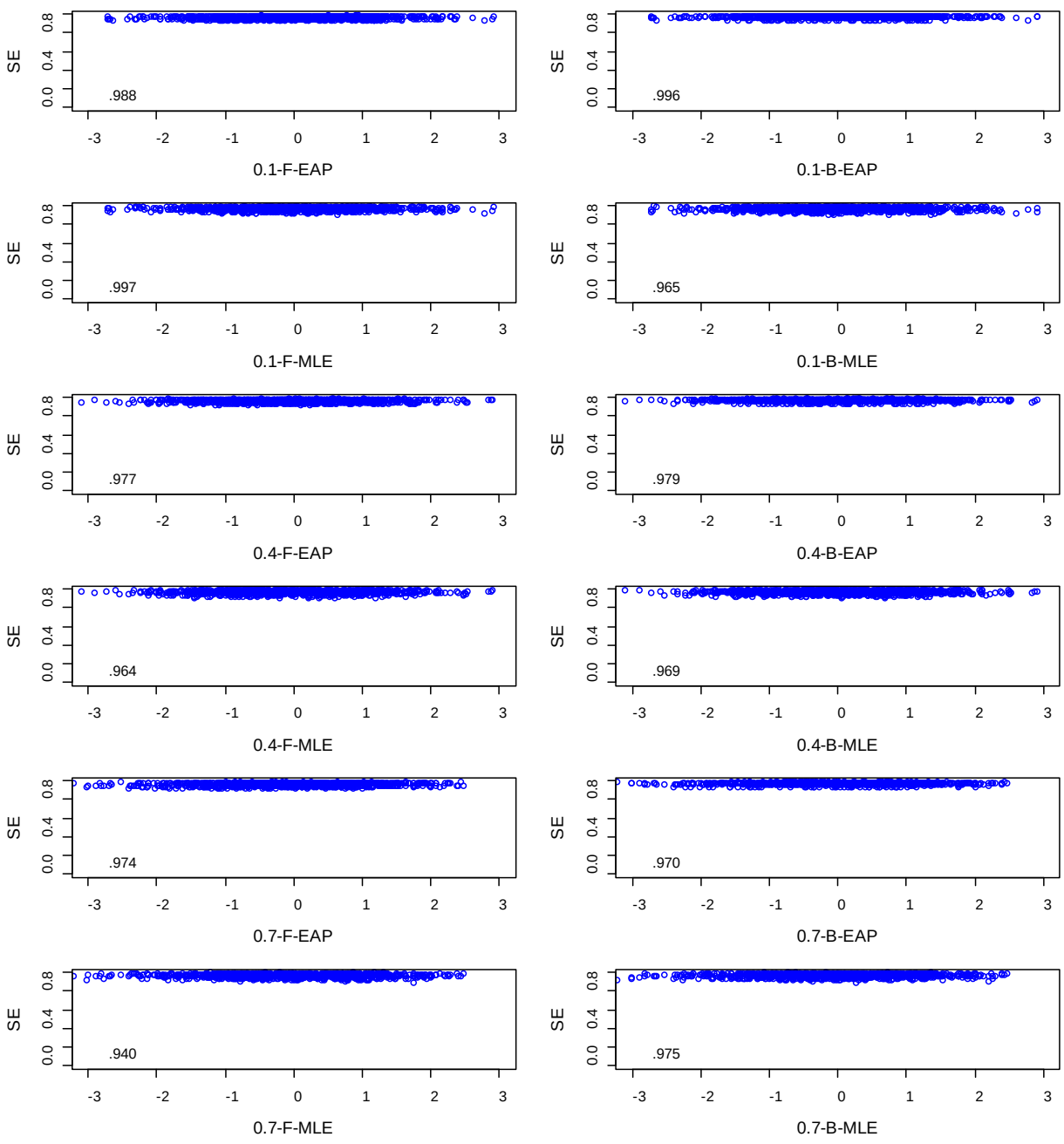


Figure C.16: Average SE (First group factor for Two-tier IRT model with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

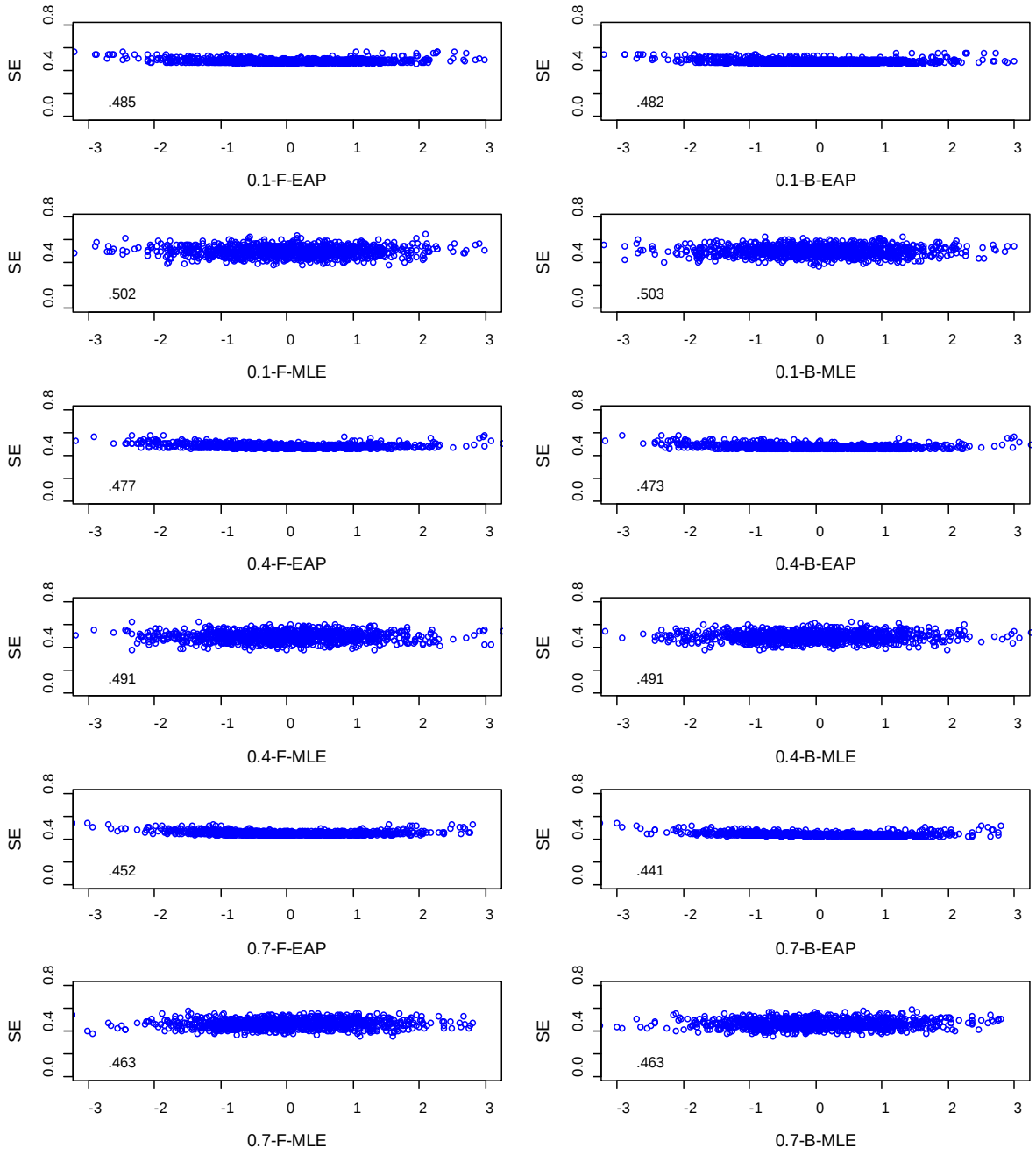


Figure C.17: Average SE (First primary factor for Two-tier IRT model with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

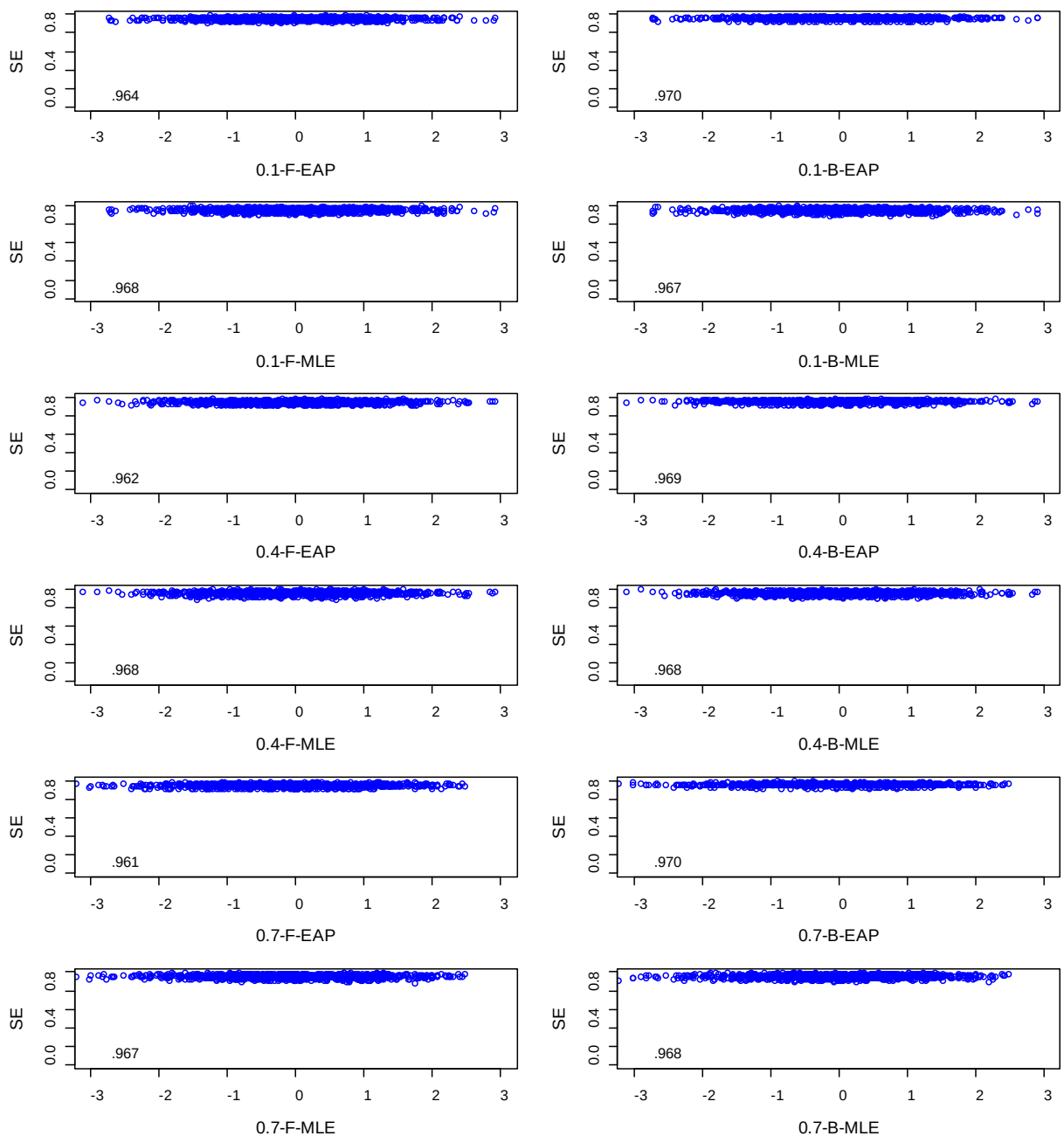


Figure C.18: Average SE (First group factor for Two-tier IRT model with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

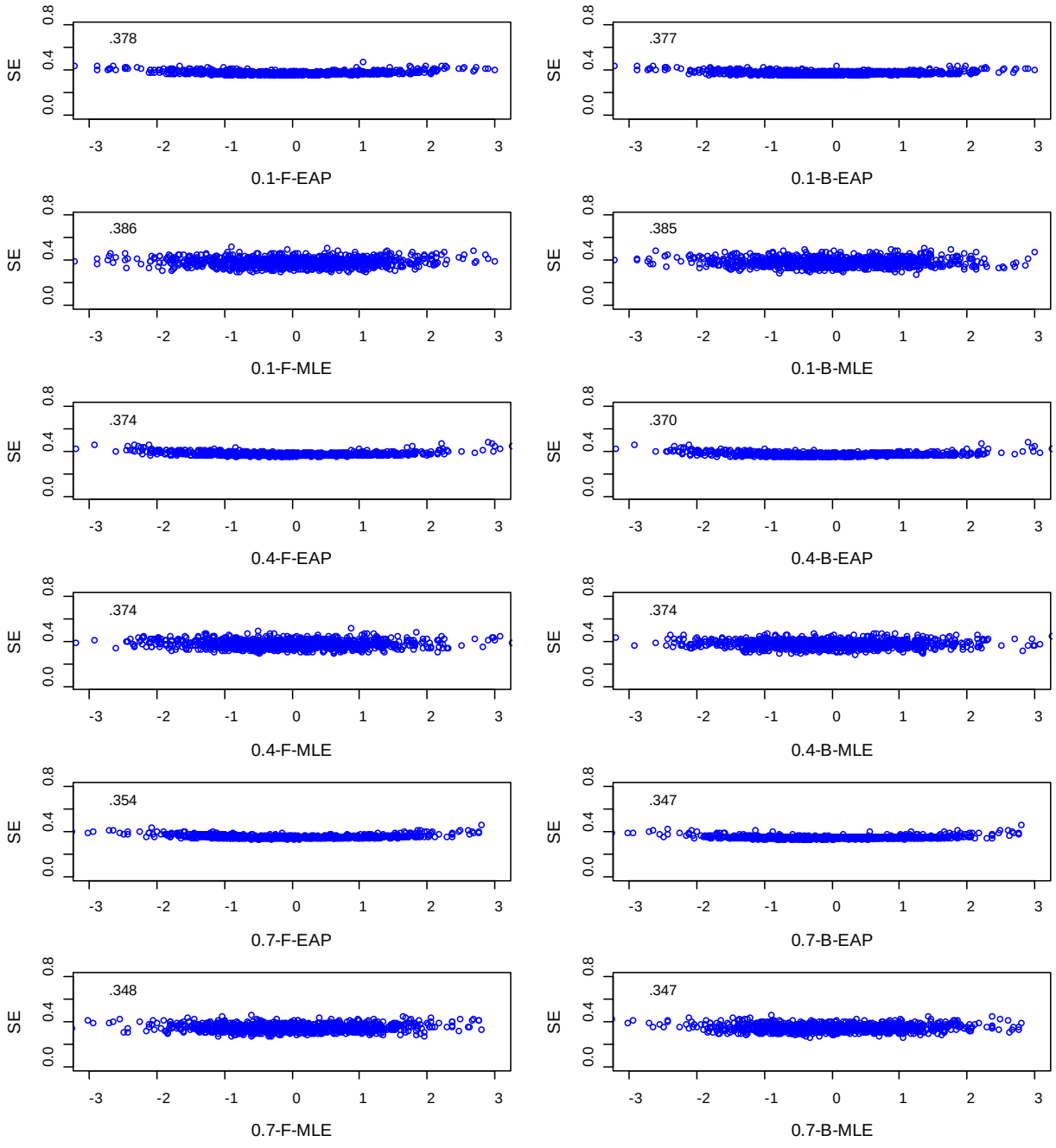


Figure C.19: Average SE (First primary factor for Two-tier IRT model with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

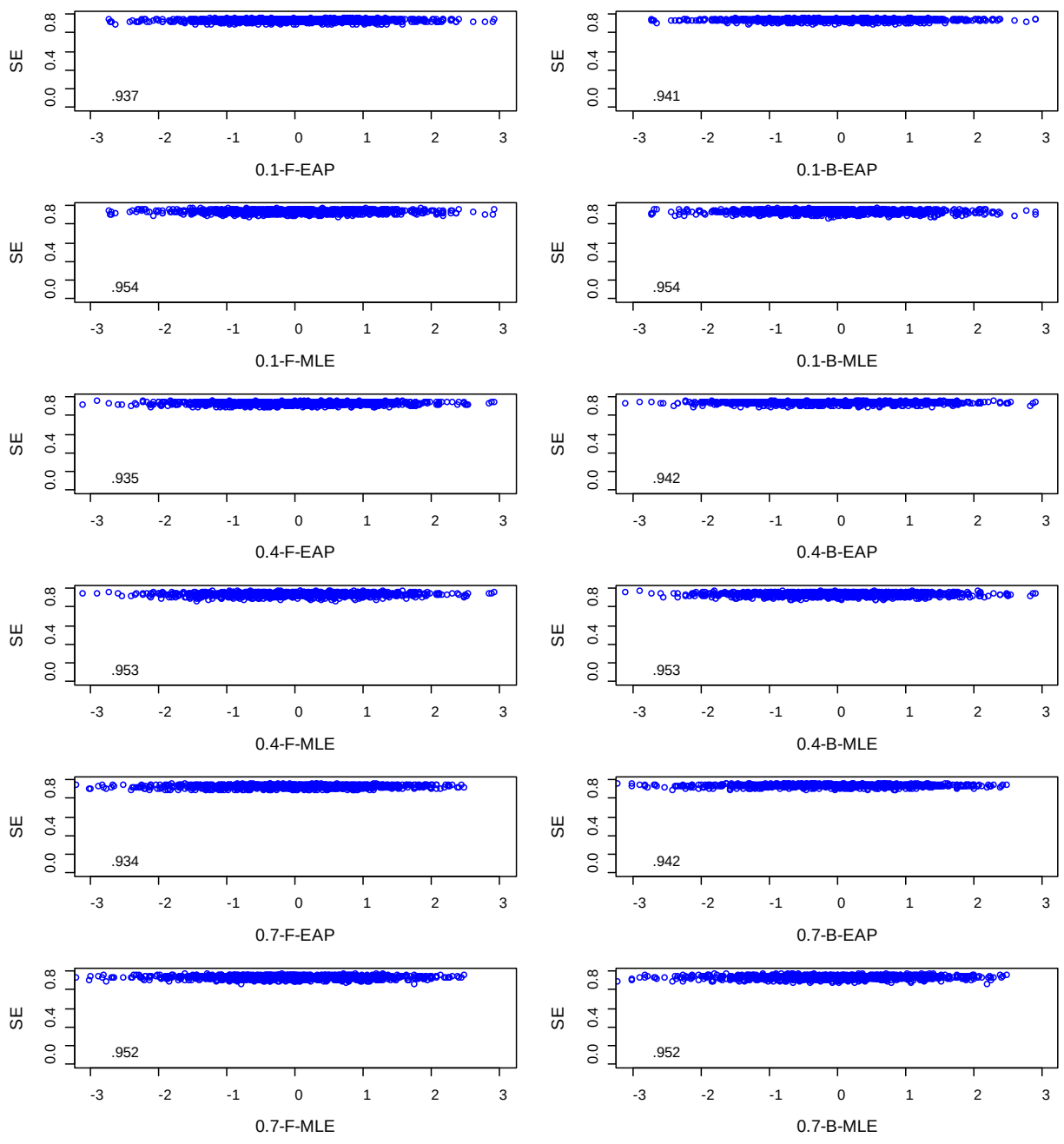


Figure C.20: Average SE (First group factor for Two-tier IRT model with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring method

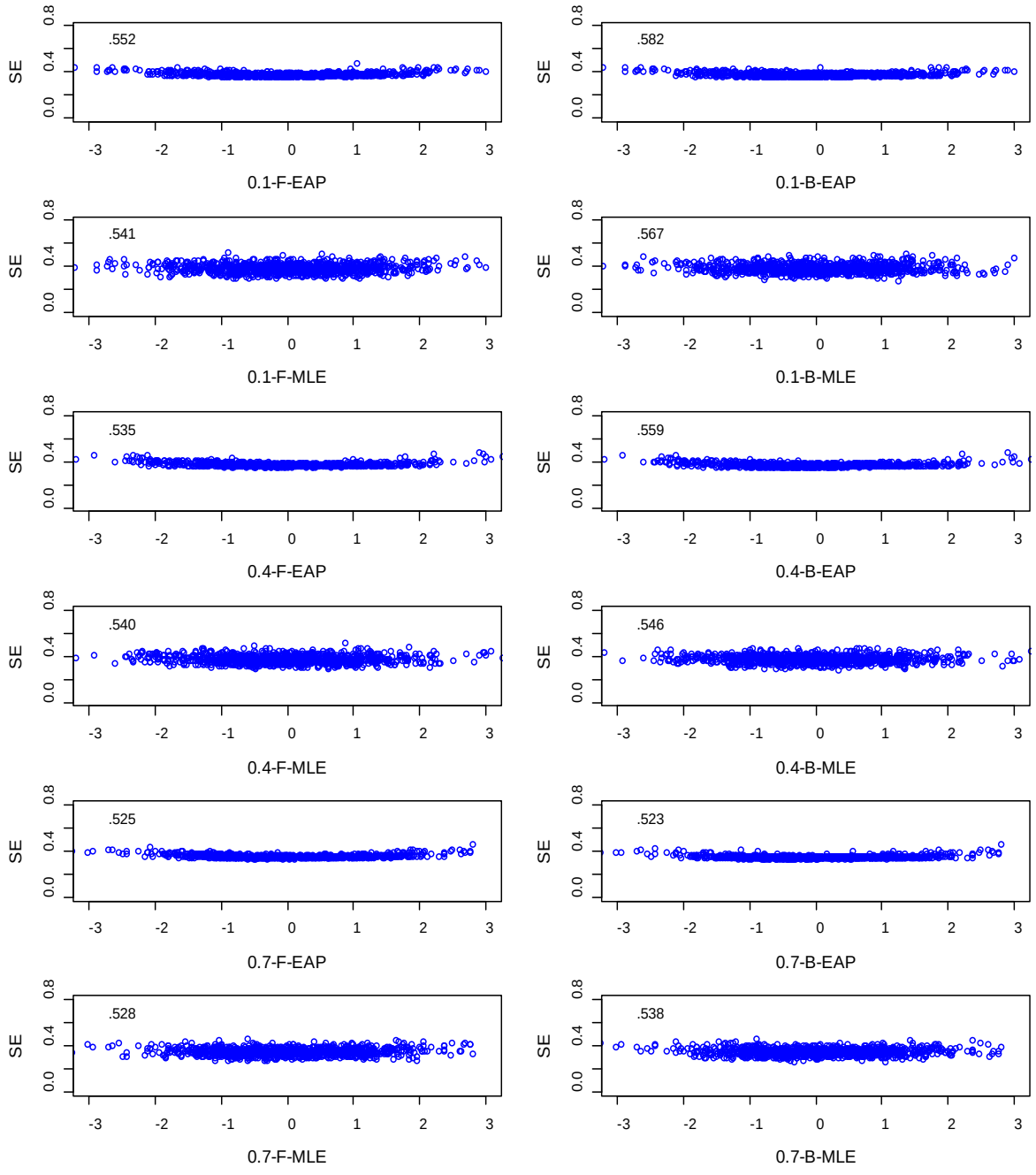


Figure C.21: Average SE (First primary factor for Higher-order IRT model (2 primary factors) with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

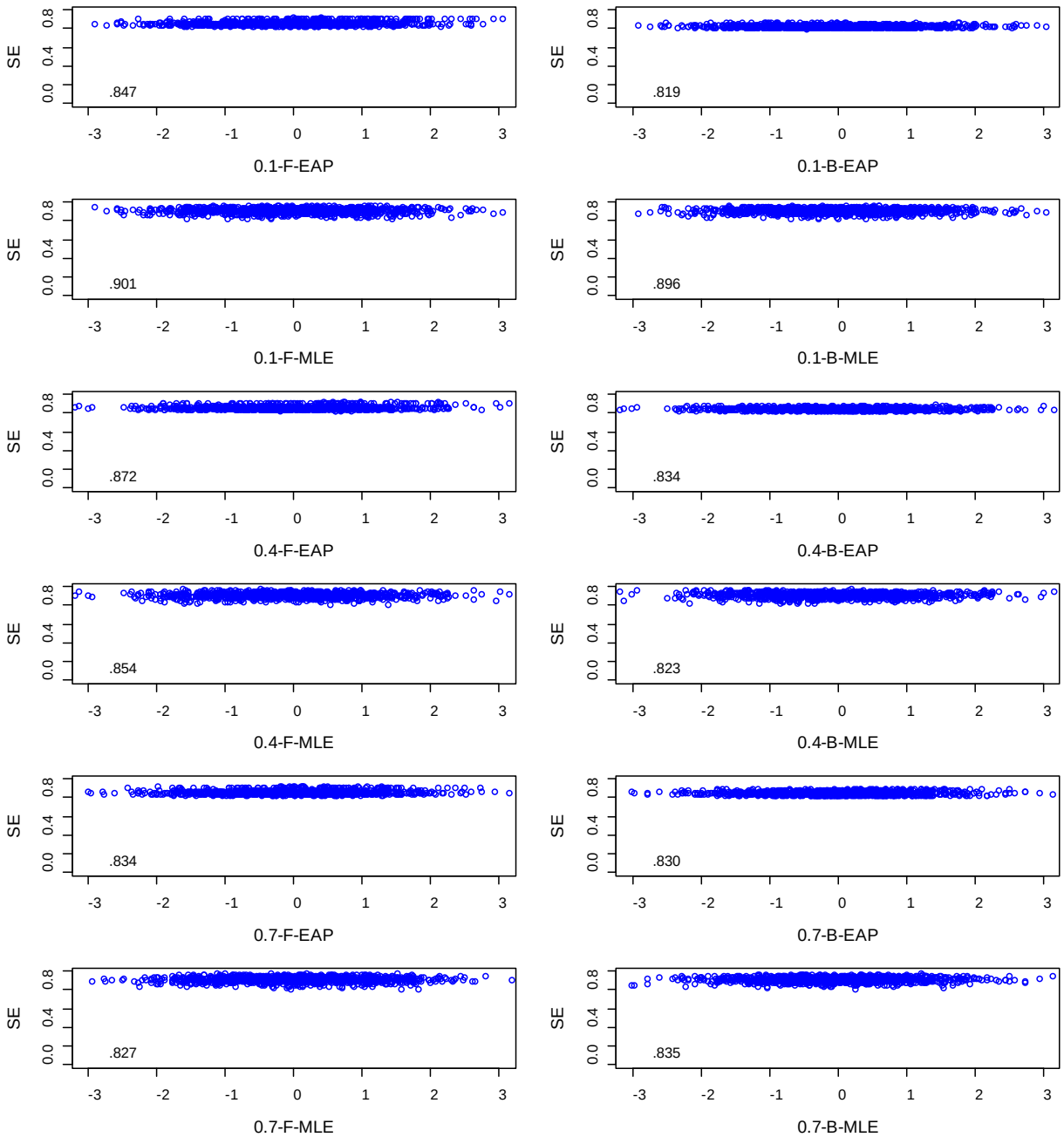


Figure C.22: Average SE (First group factor for Higher-order IRT model (2 primary factors) with two group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

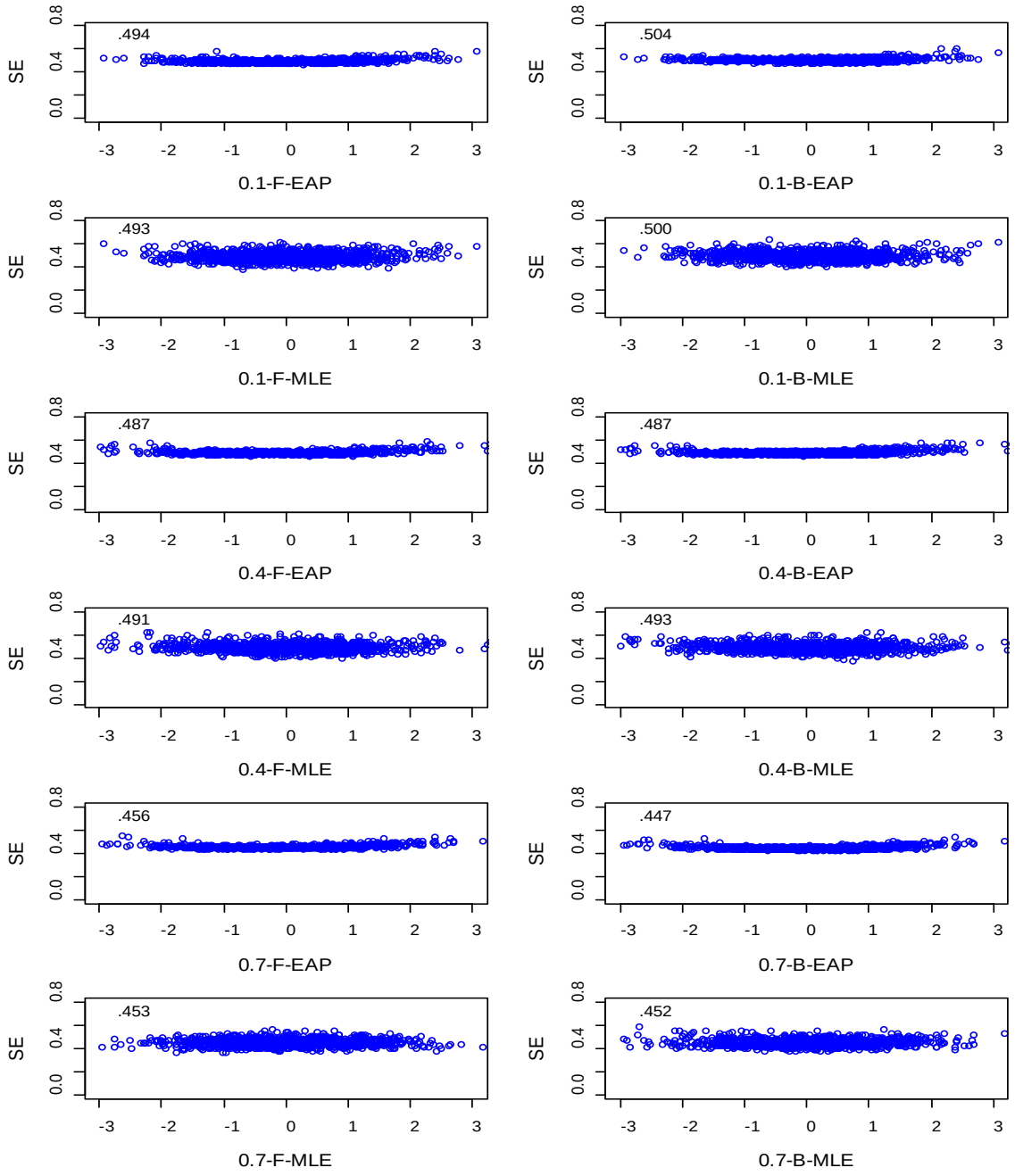


Figure C.23: Average SE (First primary factor for Higher-order IRT model (2 primary factors) with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

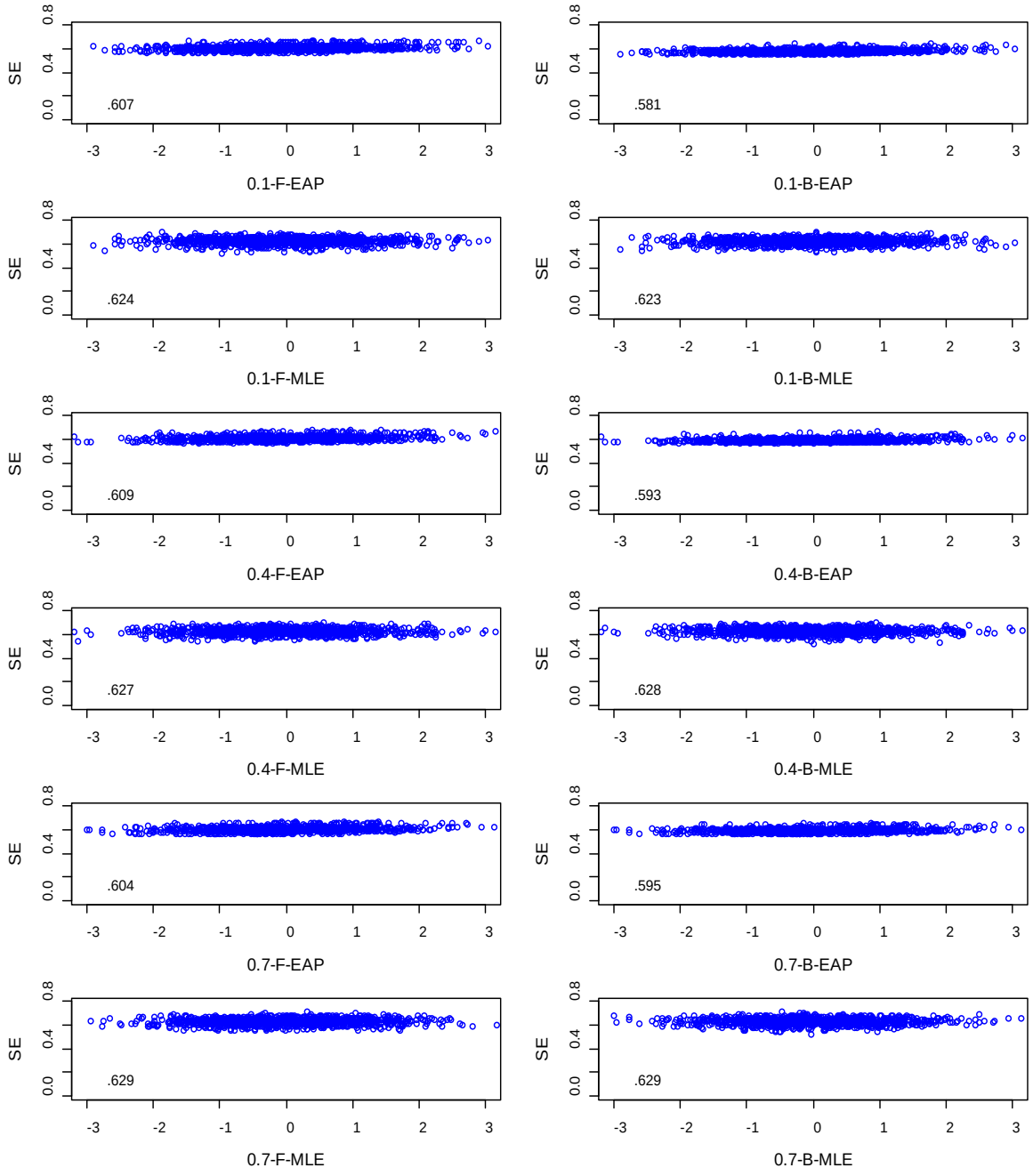


Figure C.24: Average SE (First group factor for Higher-order IRT model (2 primary factors) with two group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

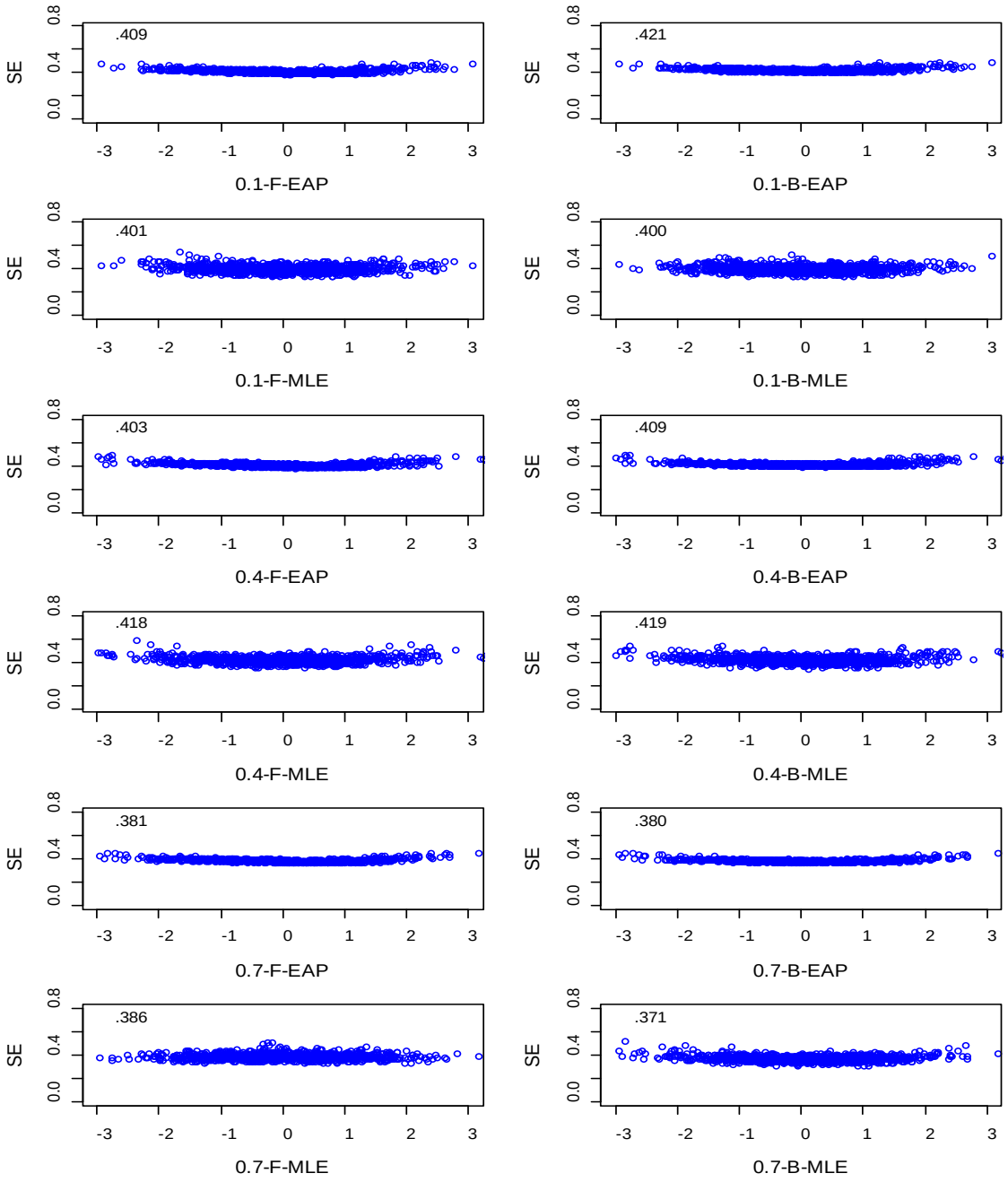


Figure C.25: Average SE (First primary factor for Higher-order IRT model (2 primary factors) with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

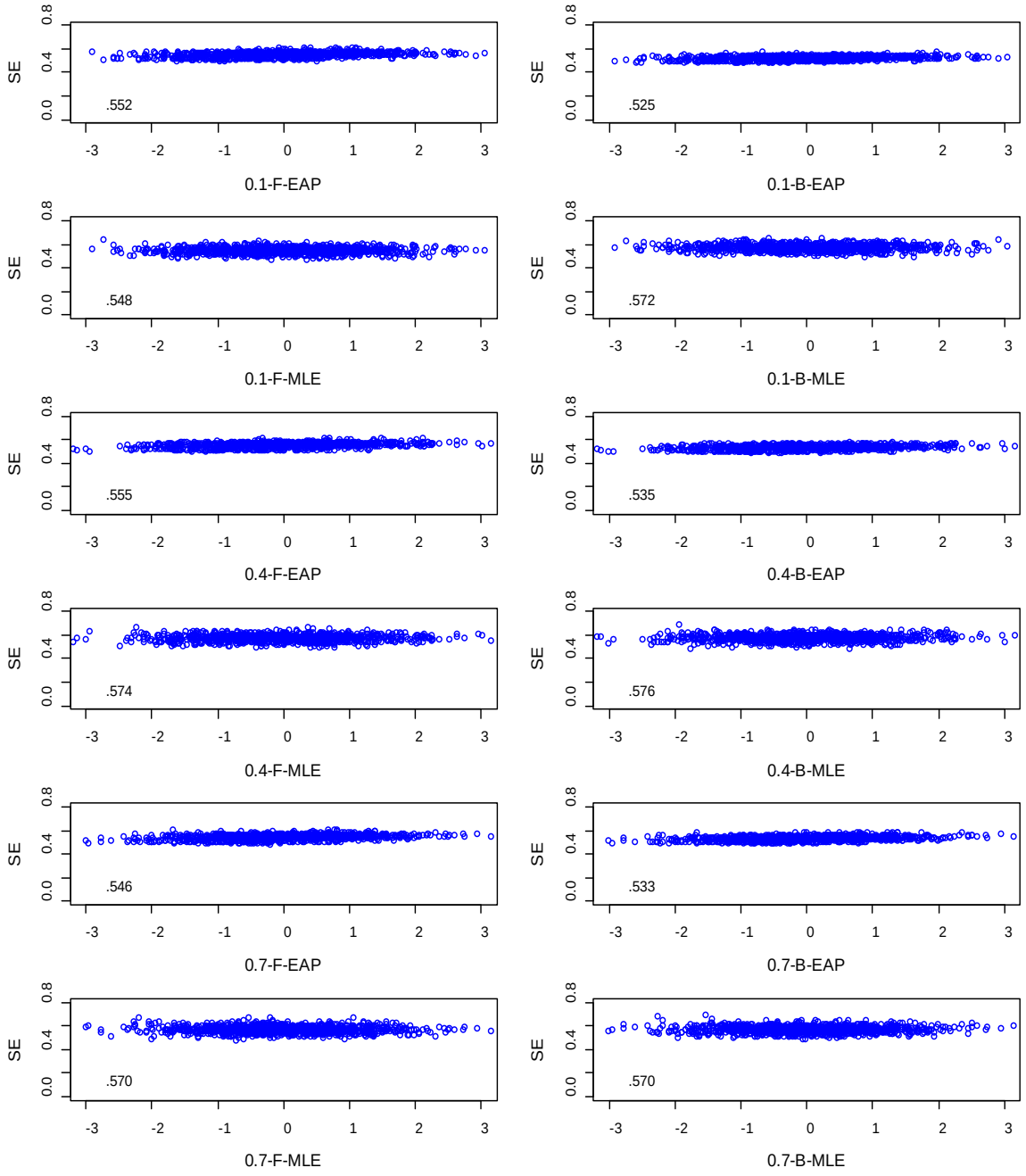


Figure C.26: Average SE (First group factor for Higher-order IRT model (2 primary factors) with two group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

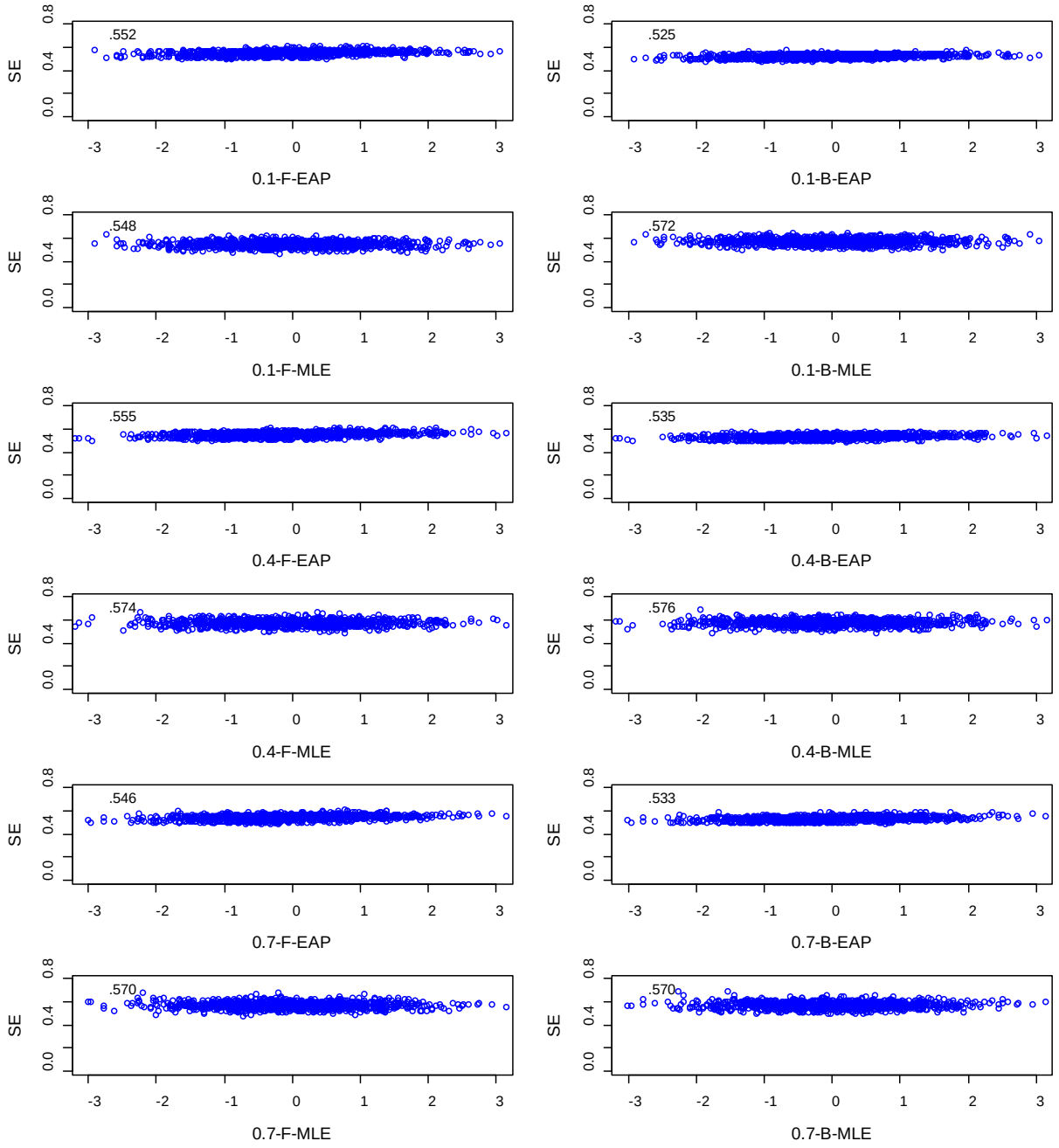


Figure C.27: Average SE (First primary factor for Higher-order IRT model (2 primary factors) with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

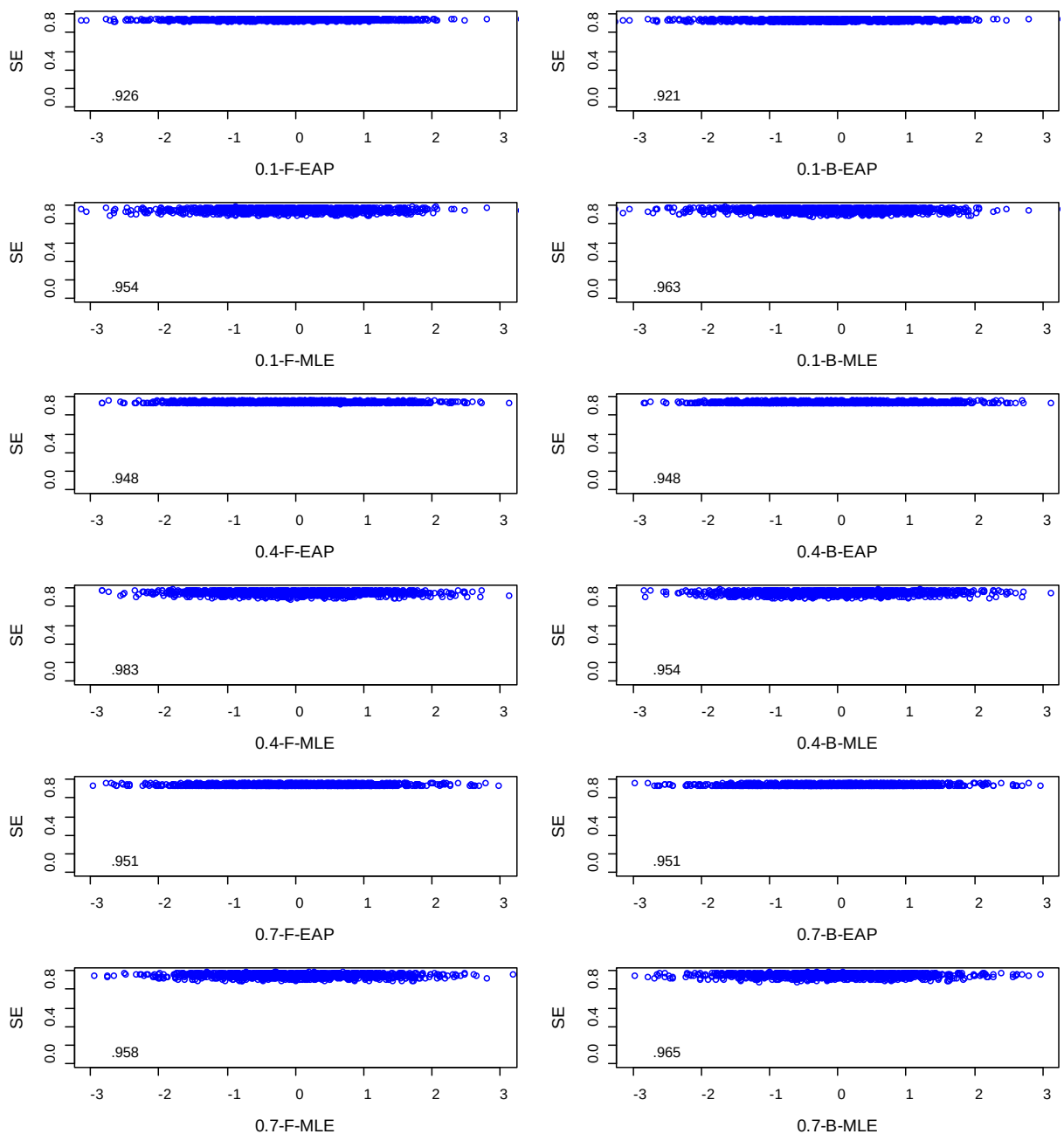


Figure C.28: Average SE (First group factor for Higher-order IRT model (2 primary factors) with four group factors (40 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

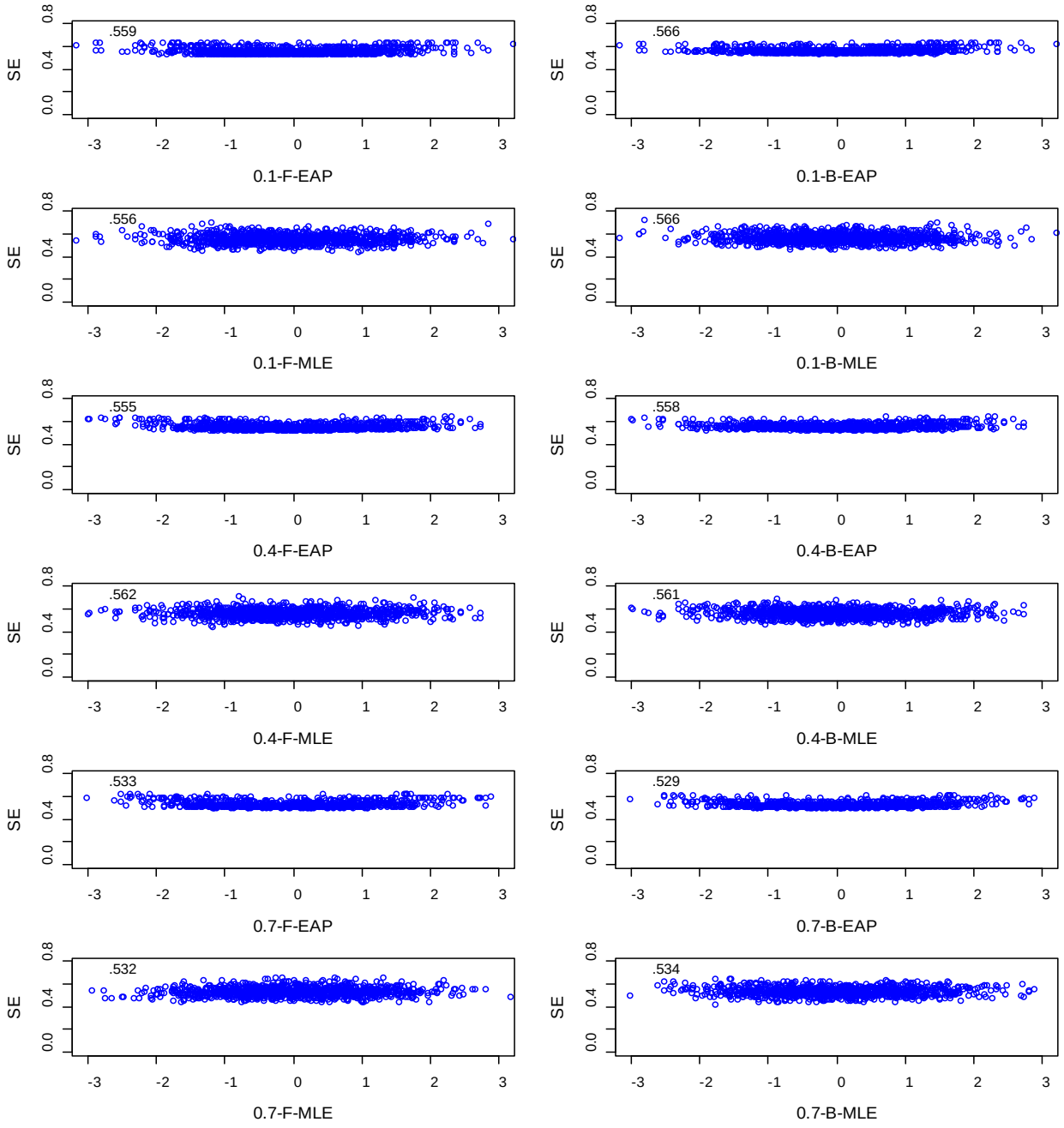


Figure C.29: Average SE (First primary factor for Higher-order IRT model (2 primary factors) with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

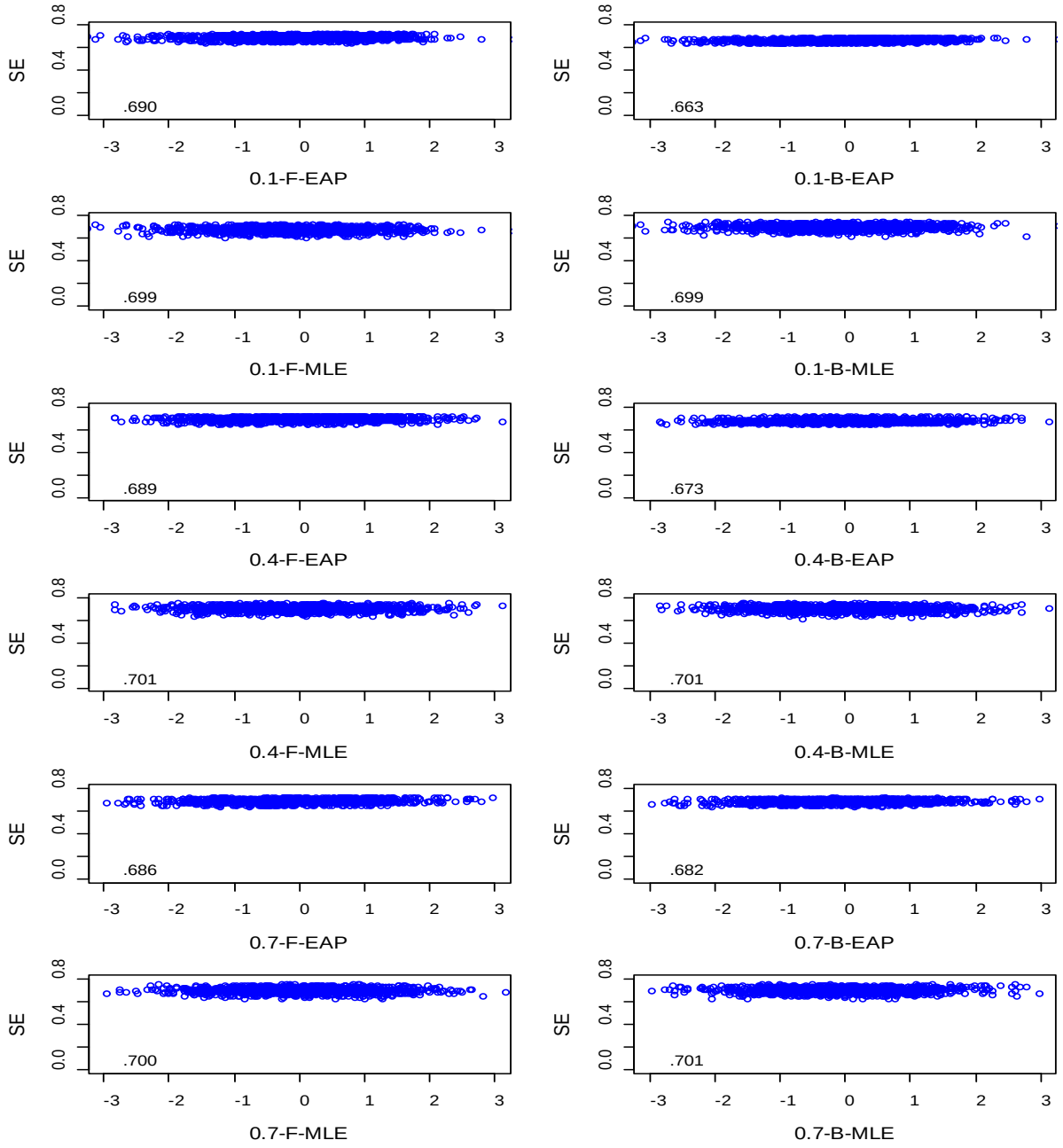


Figure C.30: Average SE (First group factor for Higher-order IRT model (2 primary factors) with four group factors (80 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

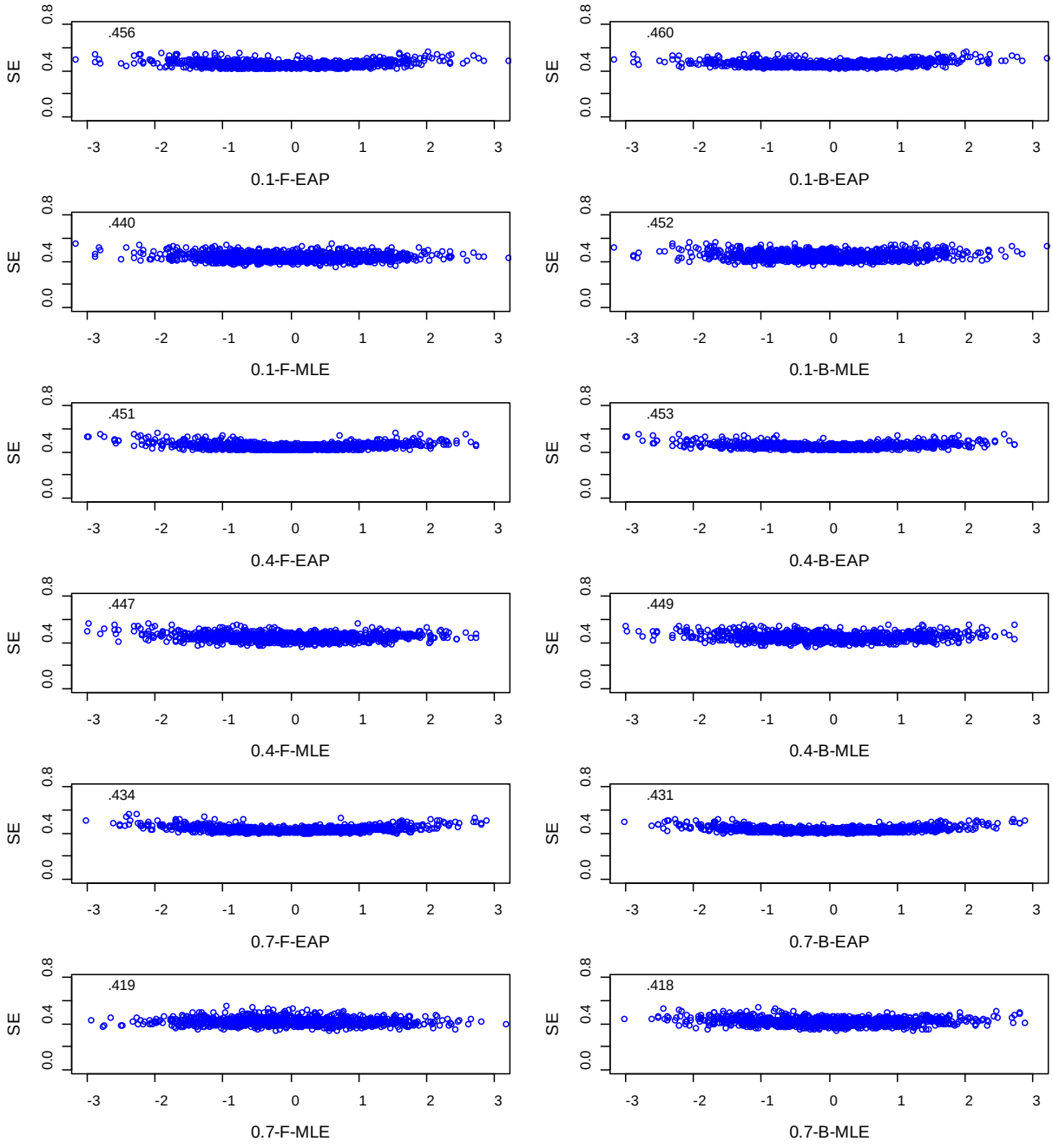


Figure C.31: Average SE (First primary factor for Higher-order IRT model (2 primary factors) with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

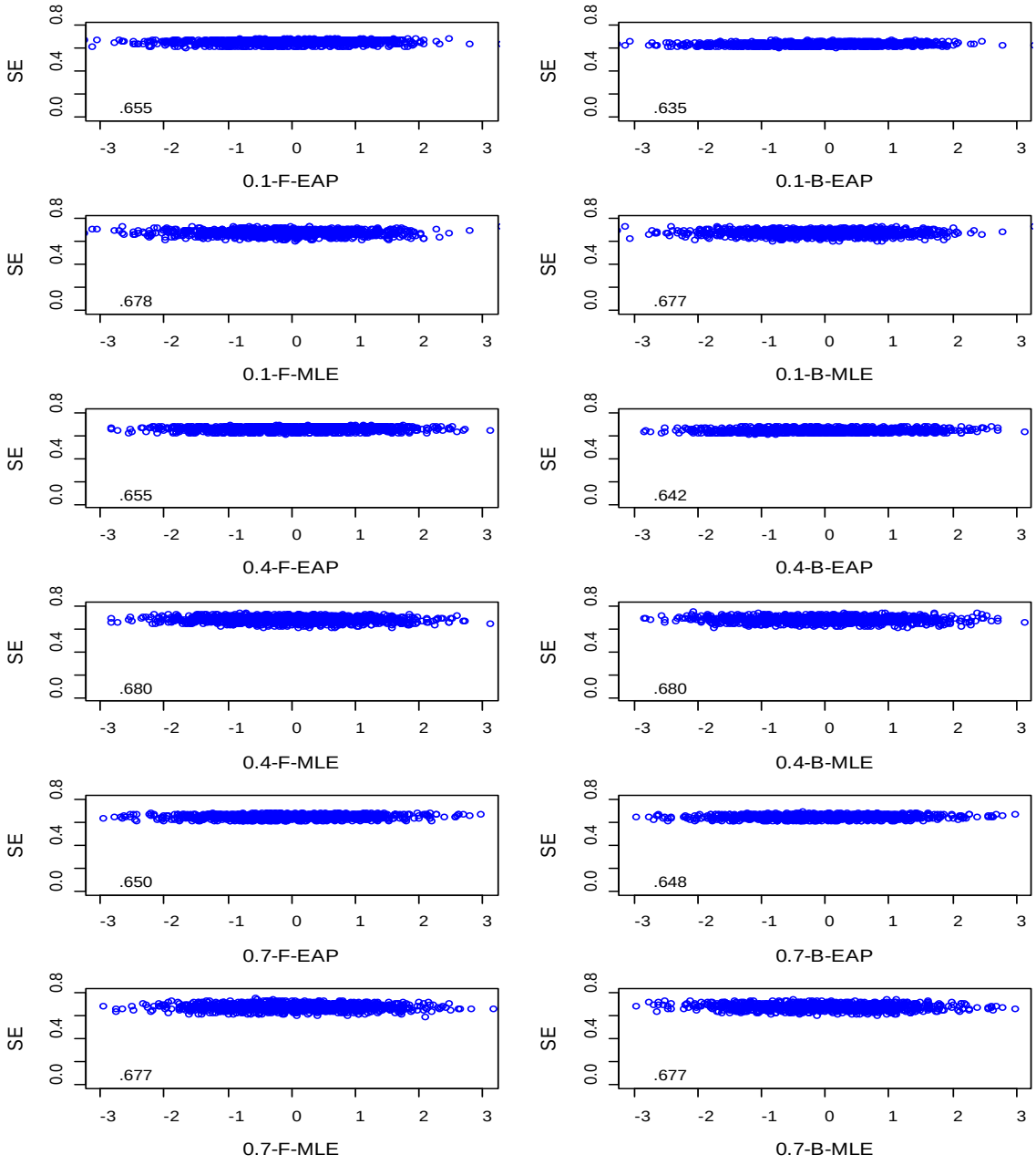


Figure C.32: Average SE (First group factor for Higher-order IRT model (2 primary factors) with four group factors (160 items))

Note. 0.1-F-EAP: 0.1 Correlation between two general factors – Fisher item selection method – EAP scoring method

0.7-B-MLE: 0.7 Correlation between two general factors – Bayesian item selection method – MLE scoring

Bibliography

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15(1),* 13–24.

Baker, F. B. (1992). *Item response theory parameter estimation techniques.* New York: Marcel Dekker.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541-562.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York, NY: Springer.

Bergstrom B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates.

Bloxom, B. M., & Vale, C. D. (1987, June). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29-51.

Bock, R. D. & Mislevy, R. J. (1988). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bolt, D. M., & Lall, V. F. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models Using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27,* 395 - 414.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581-612.

Cai, L. (2013). flexMIRT® version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16,* 221-248.

Carroll, J. B. (1993). *Human cognitive abilities.* Cambridge, England: Cambridge University Press.

Chang, S. W., Ansley, T. N., & Lin, S. H. (2000). *Performance of item exposure control methods in computerized adaptive testing: Further explorations.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Chen, F. F., West. S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189-225.

Choi, S. W., Grady, M. W., && Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71,* 37-53.

CTB/McGraw-Hill. (2002). *InView Technical Bulletin.* Monterery, CA: Author.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27(5),* 335-356.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115-130.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika. 69(3),* 333-353.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R Rao & S. Sinharay (Eds.) *Handbook of Statistics, 26,* (pp. 979-1030). Amsterdam: Elsevier.

Dodd, B. G., & DeAyala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied psychological measurement. 19,* 5-22.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika. 57,* 423-436.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., and Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Arch Gen Psychiatry, 69(11),* 1104-1112.

Gibbons R.D., Bock R.D., Hedeker D., Weiss D., Segawa E., Bhaumik D.K., Kupfer D., Frank E., Grochocinski V., Stover A. (2007). Full-Information Item Bi-Factor Analysis of Graded Response Data. *Applied Psychological Measurement, 31,* 4-19.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). *Using computerized adaptive testing to reduce the burden of mental health assessment.* Psychiatric Services. Vol. 59, No. 4.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1). Norwood, NJ: Ablex.

Green, B. F. (1983). Adaptive Testing by Computer. In R. B. Ekstrom (Ed.), *Measurement, technology, and individuality in education: New directions for testing and measurement,* No 17. San Francisco, CA: Josey-Bass.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage.

Huang, H.-Y., Chen, P.-H., & Wang, W.-C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement, 36(8),* 689-706.

Immekus, J. C., Gibbons, R. D., & Rush, A. J. (2007). Patient-reported outcomes measurement and computerized adaptive testing: An application of post-hoc simulation to a diagnostic screening instrument. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3),* 258-272.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lee, Y., Ip, E., & Fuh, C. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement, 68(2),* 215–232.

Lord, F. M. (1977). A broad-range test of verbal ability. *Applied Psychological Measurement, 1,* 95-100.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23,* 157–162.

Luecht, R. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20(4),* 389–404.

Matin, L., & Adkins, D. C. (1954). A second-order factor analysis of reasoning abilities. *Psychometrika, 19,* 71-78.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examination general test. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 117-135). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor. *Structural Equation Modeling, 4,* 193-211.

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74(2),* 273–296.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,*351-356.

Parshall, C. G., Spray, J., Kalohn, J., Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

R Development Core Team (2008*). R: A language and environment for statistical computing.* R Foundation for statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL. http://www.R-project.org.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8,* 3-3.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. van der Linden, & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer-Verlag.

Reckase, M.D. (2009). *Multidimentional item response theory.* New York: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York: Guilford.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph,* No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph,* No. 18.

Sawaki, Y., Sticker, L. J., & Andreas, H. O. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26,* 5-30.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Segall, D. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 57–75). New York, NY: Springer.

Segall D. O., & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Service Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 35- 65). Mahwah, NJ: Lawrence Erlbaum Associates.

Seo, D. (2011). *Application of the Bifactor Model to Computerized Adaptive Testing.* Unpublished doctoral dissertation, University of Minnesota.

Silvey, S. D. (1980). *Optimal design*. London, UK: Chapman & Hall.

Stocking, M. L., & Lewis, C. (1995a) . *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b) . *Controlling Item exposure conditional on ability in computerized adaptive testing* (Research Report 95-24). Princeton, NJ: Educational Testing Service.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393-408.

The Glossary of Education Reform (2013). Retrieved November 11, 2013, from http://edglossary.org/formative-assessment/

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: a primer* (pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69,* 778-793.

U.S. Department of Education (2012). *Race to the Top Assessment: Smarter Balanced Assessment Consortium Year One Report,* Washington, DC.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error variance criterion. *Journal of Educational and Behavioral Statistics, 24(4),* 398–412.

van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Zeitschrift für Psychologie / Journal of Psychology, 216(1),* 3–11.

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, W.-C., & Huang, S.-Y. (2011). Computerized classification testing under the one-parameter logistic response model with ability-based guessing. *Educational and Psychological Measurement, 71,* 925-941.

Wang W.-C., Chen P.-H., & Cheng Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9,* 116-136.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing,* URL. http://www.psych.umn.edu/psylabs/catcentral/pdffiles/cat07weiss&gibbons.pdf

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized testing to educational problems. *Journal of Educational Measurement, 21(4),* 361–375.

Weiss, D.J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 272-285.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Testing scoring, item analysis, and item factor analysis.* Chicago, IL: Scientific Software International, Inc.

Yao L., & Boughton K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31,* 83-105.

Yung, Y. F., Thissen, D., McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113-128.