Incorporating Person Covariates and Response Times as Collateral Information to Improve Person and Item Parameter Estimations

Shudong Wang
NWEA

Hong Jiao
University of Maryland

Paper presented at the annual meeting of the National Council on Measurement in Education. April 18-22, 2011, New Orleans, Louisiana.

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206
Shudong.Wang@NWEA.org

I. Introduction

Computerized based testing (CBT) has been widely used in K-12 education assessment. Recording response times (RT) on test items becomes a routine test activity for many large scale state test administrations. Besides RT, the examinee's auxiliary information, such as background (gender, ethnicity), and academic history (course taken, grades received, and scores on other subject area) are also routinely collected in many testing programs.

However, it is still a common practice to estimate person and item parameters based on item response theory (IRT) from item responses alone and ignore both RT and auxiliary or collateral information (CI, Mislevy & Sheehan, 1989), even though it is available for many psychometric applications, such as test scoring, equating and scaling. The purpose of this study is to exploit student RT, background information, and academic information as auxiliary information to improve the precision of parameter estimates in IRT.

It has long been recognized that RT on a test is an important source of information on the student's behavior and research topic of interest in psychophysics and cognitive psychology (Luce, 1983; Roskam, 1997). The RT is also collected routinely in empirical studies such as biological, social, and development and clinical psychology. In fact, more than 27,000 abstracts in PsychInfo database spanning from 1887 to the end of April, 2000 make reference to reaction or response time or latency (Van Zandt, 2002). In educational measurement, however, ignoring RT has been accepted since early part of the 20th century (Baxter, 1931, Mayer, 1960) and this may be partially true due to the difficulty of collecting response time data at the individual item level with paper-and-pencil testing. But this, for most part, is because of the need for the standardization for large-scale standardized norm-referenced tests (SNRT, Anastasi, 1976; Ebel & Frisbie, 1991).

In theory, educational tests are still classified exclusively as either power tests that assume there is no time limit with student achievement ability is as only account for student score, or speed tests that assume RT is the only account for student score and the difficulty of items is not an issue. Under this framework of making an exclusive distinction between speed and power test with ideal assumptions, instead of treating RT as a valuable source of information that may reveal the reason why student s perform in certain way, their behavior, and cognitive style in the testing, RT still be regarded as nuisance variable in traditional testing programs that measure

student achievement. In practice, however, it is very hard or almost impossible to obtain the pure power tests (unlimited time) and pure speed tests (very easy items), which means at least on an individual level, pure measures of ability, uncontaminated by personality and cognitive style reflected by RT, are unattainable (Dennis & Evans, 1996).

Recently, researchers have shown how to use RT as auxiliary information to improve calibration items (Klein Entink, Fox, & van der Linden, 2009; Klein Entink, Kuhn, Hornke, & Fox, 2009; van der Linden, 2010; van der Linden, Klein Entink, & Fox, 2010), to analyze speediness of test (van der Linden, 2009b; van der Linden, Breithaupt, Chuah, & Zhang, 2007), to detect cheating and check test behavior for possible aberrances (van der Linden & Guo, 2008). Currently, there are three different types of models to extract RT related information (van der Linden, 2007). The first type of model focuses on RT exclusively and response scores are not taken into account; the second type of model conducts separate analysis of responses and RT; the third type of model conduct analysis jointly using both information from response and RT. An example of the third type of modeling is a multivariate multilevel approach (Klein Entink, Fox, & van der Linden, 2009) that jointly models dichotomous responses by using regular IRT model and continuous RT by using a lognormal model. For this approach, binary and continuous responses on test items are assumed to be nested with person or examinee and data structures are multivariate clustered data. Modeling such data is necessarily complex, because two types of correlations must be considered correlation between measurements on different variables for each person and correlation between measurements on different variables within a person. Besides using RT as CI, personal information was also used to improve estimation procedures. For example, Hall (2006) and Mislevy (1989) showed that collateral information can reduce the uncertainty in parameter estimations.

The goal of all modeling in test theory is to make test scores more informative, since CI is part of student performance and a byproduct of testing in a computer-based test. It seems natural to use it to improve parameter estimation in many education assessments. Another reason to use RT when estimating parameters is that use of RT does not change the construct measured by a test, but increases the accuracy of estimations. Overall, CI can be used to improve traditional IRT parameter estimations that are based solely on item responses. Despite the fact that many studies have shown that CI can be used to check the quality of the items, improve the design of the tests, monitor their quality during test administrations, diagnose the

response behavior of the test takers, and increase the efficiency of test scoring (Van der Linden, 2010). Few studies have been done on how to use CI to improve the precision of parameter estimation in real applications, the purpose of this study is to compare the parameter estimations with and without the use of CI in a large-scale CBT.

II. Methods and Data

2.1 Instrument

Two Skills Checklist tests (Northwest Evaluation Association, 2009) from Measures of Academic Progress for Primary Grades (MAP™ for Primary Grades, MPG) were used in this study. The goal of MPG is to provide information about specific skills and concepts. The MPG can be used prior to instruction to help teachers determine which skills need the most instructional focus. These tests can be administered as many times as necessary during the school year to give an indication of the student actual learning. The MPG includes Screening tests, diagnostic Skills Checklist tests, and adaptive Survey with Goals tests in Reading and Mathematics (NWEA, 2009). For this study, among 27 mathematics Skill Checklists tests and 13 reading Skill Checklists tests, only Number Sense (NS34) with 34 items was used for mathematics test and only Letter Identification (LI52) with 52 items was used for reading test. All MPG tests are computerized linear tests.

2.2 Sample

All samples are from 2009 administrations across 50 states (including Washington DC) and sample size for each test is about 1454 students who are drawn from 66712 original student pools. For all joint estimations, student test scores were matched (the same students took NS34 and LI52) on tests that are used as CI. Besides RT, student's gender and ethnicity are also identified.

III. Methods

Because we are trying to model clustered data (items within person) with binary (0, 1) and continue responses (response time), a joint model (Cox & Wermuth, 1992; Fitzmaurice, & Laird, 1995; Klein Entink, Fox, & van der Linden, 2009; Snijders & Bosker, 1999) for

multivariate mixed ordinal and continuous responses will be used. The major advantage of joint model is its capability to model the dependence between student ability and speediness across different conditions.

3.1 Joint Model without Person Covariates

The joint model used in this study is the multivariate model under generalized linear mixed models (GLMM) framework. Unlike univariate methods used in GLMM framework that estimate parameters separately with other information, this model simultaneously models binary item response and continuous RT response on test items that nested within person (Klein Entink, et al. (2009). Let Y_{ij} denote the response and T_{ij} denote the RT for person i =1, …, N on item j=1,.., J, then the probability that person i answer item j correctly is $p(Y_{ij}=1|\theta_i) = \frac{1}{1+exp[-(\theta_i - b_j)]}$, which is Rasch model and where θ_i is person ability parameter, b_j is item difficulty parameter, and $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$; the log-response time of person i on item j follow a normal model: $T_{ij} = -\zeta_i + \lambda_j + \varepsilon_{\zeta ij}$ and where ζ_i is person speed parameter, λ_j is time intensity parameter, and $\varepsilon_{\zeta ij} \sim N(0, \sigma_j^2)$. If let $\mu_{1ij} = p(Y_{ij}=1|\theta_i)$ and $\mu_{2ij} = E(T_{ij}|\zeta_i)$, the joint mode is given by

$$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \theta_i - b_j \quad (1)$$

$$\text{Log}(\mu_{2ij}) = \eta_{2ij} = -\zeta_i + \lambda_j \quad (2)$$

where $(\theta_i, \zeta_i) \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix} \quad (3)$$

The covariance matrix $\boldsymbol{\Sigma}$ models the variability in ability and speed ($-\zeta_i$) of the person and allows for dependence between them by parameter ρ in the model. If ρ>0, this means that person who answer faster than average on the test are expected to have above- average ability; if ρ<0, this means that person who answer faster than average on the test are expected to have below- average ability. If two data source are independent, then ρ = 0, which means independence between ability and speed, but this doesn't mean the independent between the response and RTs.

3.2 Joint Model with Person Covariates

Although some researches (Adams, Wilson & Wu, 1997; Mislevy, 1987; Rijmen, Tuerlinckx, De Boeck & Kuppens, 2003) have shown examples of univariate approach of using person covariates in the latent regression. A few studies have focused estimation parameters jointly with CI. According to Klein Entink, Fox, & van der Linden (2009) and Snijders & Bosker (1999), regression of random effects (person ability and speed) on covariates can be formulated as:

$$\theta_i = \gamma_{10} + CV_{i1}\,\gamma_{11} + CV_{i2}\,\gamma_{12} + CV_{i3}\,\gamma_{13} + \ldots + CV_{im}\,\gamma_{1m} + e_{1i}, \quad (4)$$

$$\zeta_i = \gamma_{20} + CV_{i1}\,\gamma_{21} + CV_{i2}\,\gamma_{22} + CV_{i3}\,\gamma_{23} + \ldots + CV_{im}\,\gamma_{2m} + e_{2i}, \quad (5)$$

where covariates (CV, 1,2, …, to *m*) could be either categorical or continue as fixed effects in this regression model, γ_{1m} and γ_{2m} are regression coefficients of CVs. Here $(e_{1i}, e_{2i})^t$ ~MVN(0, $\boldsymbol{\Sigma}$) and $\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2\end{bmatrix}$. For example, if we would like to know the effect of RT of each item, totally scores of LI52, gender, ethnicity as CI for each individual students on student person and NS34 test item parameter estimations, we could get following joint model by substituting (4) and (5) into (1) and (2),

$$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + LI52_i\,\gamma_{11} + gender\,\gamma_{12} + ethnicity\,\gamma_{13} - b_j + e_{1i}, \quad (6)$$

$$\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + LI52_i\,\gamma_{21} + gender\,\gamma_{22} + ethnicity\,\gamma_{23} + \lambda_j + e_{2i}, \quad (7)$$

Similar, the effect of RT of each item, totally scores of NS34, gender, ethnicity as CI for each individual students on student person and LI54 test item parameter estimations

$$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + NS34_i\,\gamma_{11} + gender\,\gamma_{12} + ethnicity\,\gamma_{13} - b_j + e_{1i}, \quad (8)$$

$$\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + NS34_i\,\gamma_{21} + gender\,\gamma_{22} + ethnicity\,\gamma_{23} + \lambda_j + e_{2i}, \quad (9)$$

3.3 Estimation /Calibration

The SAS procedure PROC GLIMMIX (GLIMMIX Procedure Documentation, 2005) was used to jointly analyze a continuous and binary outcome outcomes. The GLIMMIX procedure

fits statistical models to data with correlations or nonconstant variability and where the response is not necessarily normally distributed. These models are known as GLMM. The limitation of GLMIMXED is when number of items and observations become large or in case of many missing, then the GLIMMIX sometimes cannot converge. Although there are two ways to model the dependence between response and RT for the same student (see, GLIMMIX Procedure Documentation, 2005), in this study, the dependence ρ between θ_i and ζ_i was directly estimated and correlations matrix was obtained by estimated **G** matrix in the Generalized linear mixed models (GLMM) in matrix form:

$$\mathrm{E}[\mathbf{Y}|\boldsymbol{\theta}] = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta}) = g^{-1}(\eta)$$

and the inverse link function is defined as $g^{-1}(\eta) = \mu$ and a linear predictor can contain random effect:

$$\eta = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta}$$

Where $\boldsymbol{\theta} \sim \mathrm{N}(0,\boldsymbol{G})$. The estimation method used throughout this study is restricted maximum likelihood estimation methods (as the default estimation method of GLIMMIX).

3.4 Model Selection

Models described in equation (6) to (9) are full models in which all predictors, items (I), raw score (R), gender (G), and ethnicity (E) are included. A variety of nested models or reduced models (special case of full model) can be statistically tested by constraining predictors in the full models. Because the GLMMIX use quasi-likelihood, the goodness of fit (GOF) statistics used to test model fit is the Likelihood-Ratio (LR) statistics that involves quasi-likelihood instead of likelihood. The statistical test hypothesis is

H_0: Reduced model is true against H_1: Full (current) model is true.

The LR statistics is

$$\begin{aligned} \mathrm{G} = \chi^2 &= -2 \text{ log (pseudo) likelihood of reduced model} \\ &\quad - (-2 \text{ log (pseudo) likelihood of current model}) \\ &= -2\,\mathrm{LL_r} - (2\mathrm{LL_c}) \\ &= -2\ln(\mathrm{LL_r}/2\mathrm{LL_c}) \end{aligned} \tag{10}$$

and the degree of freedom d is the difference of number predictors in two models, the p-value is $p(\chi^2_d \geq G)$. Table 1 lists all nested models that are jointly tested for both NS34 and LI52 tests.

3.5 Real Data Analysis

Because the purposes of the study is to see how effectiveness of CI on test parameters calibration, different baseline models will be used for the comparison purpose. Because of time limited in this study, models that jointly estimated but do not include dependence between bivariate variables (person ability and speed) are not tested, the results of bivariate without dependence modeling should be the same as the results from two separated univariate modeling approach. In this study, all real data analyses are based on joint-models approaches and different nested models are tested. For example in Table 1, models 3 is the model 4 without raw-score.

IV. Results and Discssions

4.1 Model Fit

The summary of fit statistics for both NS34 and LI52 are shown in Table 2. Two type GOF test results are presented in the Table 2. First type of tests (G1) uses model without CI as base model (I); second type of tests (G2) uses previous nested model as base model.

The results for NS34 show that G1 indicates that alternative hypotheses (H_1) are true for model 2, 4, and 5 and the same is true for G2. These results mean that (1), models with R, R and G, and all CI variables are better than the base model 1; and (2), current model is better than reduced model when adding R to model1, R and G to model 3, and all CI variables to model 4.

The results for LI52 show that G1 indicates that alternative hypotheses (H_1) are true for model 7, 8, and 9 and the same is true for G2 except for model 8. These results mean that (1), models with R, G, and R and G are better than the base model 6; and (2), current model is better than reduced model when adding R to model 6, R and G to model 8.

4.2 Person Parameter Estimation

Table 3 presents means and standard deviations of theta and eta parameter estimations and standard error of estimations for both NS34 and LI52 tests. Figures 1 and 2 include the scatter plots of theta parameters for NS34 and LI52. The results show that the distribution of theta parameters of NS34 is negative skewed and the distribution of theta parameters of LI52 is close to the normal distribution, which means either the items in NS34 are very easy or students

have high ability for NS34 test. The difference among models have a little effect on theta parameter estimations for both NS34 and LI52 tests.

Figure 3 and 4 contain the scatter plots of zeta parameters for NS34 and LI52. The results show that the distribution of zeta parameters of NS34 is close to normal distribution and the distribution , while the distribution of parameters in LI52 is a little bit of positively skewed. The difference among models have more effect on zeta parameter estimations for both NS34 and LI52 tests than that on theta parameter estimations.

4.2 Relationship Between Person Parameters

The correlations $\rho(\theta_i, \zeta_i)$ between theta and zeta across different models are listed in the Table 2. The negative sign of ρ means that higher ability students tend to response the item fast. The range of ρ for NS34 test under different models are from -0.18389 to -0.16582; The range of ρ for LI52 test under different models are from -0.34984 to -0.29045. Figures 5 and 6 contain the scatter plots of person theta and speed parameters for base model (1) of NS34 Test and base model (6) of LI52 test. From these correlation results, it is clear that the student ability measured by LI52 test tend to be affected more by the speed than the ability measured by NS34 test.

4.3 Item Parameter Estimation

Table 4 presents means and standard deviations of b and lambda parameter estimations and standard error of estimations for both NS34 and LI52 tests. Figures 7 and 8 include the scatter plots of theta parameters for NS34 and LI52. Both results show that the distributions of item difficulty parameters are positively skewed. Figure 9 and 10 contain the scatter plots of item time intensity parameters for NS34 and LI52. Both results show that the distributions of item time intensity parameters are negatively skewed. For both b- and lambda-parameters, the model effects are small.

4.4 Relationship Between Item Parameters

Pearson correlation coefficients $\rho(b_i, \lambda_i)$ of item parameters (b and lambda) estimations across different models for NS34 and LI52 are listed in Table 5. Figures 11 and 12 contain scatter plots of item difficulty and time intensity parameters across models for NS34 Test. It is clear that models have little impact on correlations between b and lambda. Comparing to the

$\rho(\theta_i, \zeta_i)$ that has higher absolute values for LI52 test than for NS34 test, $\rho(b_i, \lambda_i)$ has higher absolute values for NS34 test than for LI52 test.

4.5 Overall Conclusions

Currently, the common practices in educational assessment is still treat CI as by-products even though that CIs are all simultaneously collected (or available) during the test administration along student responses. Many researchers (Mislevy & Sheehan, 1989; van der Linden, Klein Entink, & Fox, 2010) from the psychometric field suggested that the benefit role or contribution of CI in improving both the accuracy and the bias of item and person parameter estimates should not be ignored. Other researchers (Dennis & Evans, 1996) from psychology fields believed that incorporating cognitive elements that may be reflected in RT in standard psychometric model (psycho-metric model) will greatly enhance the quality of current educational assessments, and they called for developing statistical models that models not only the cognitive ability, but also personality or cognitive style. The notion that most educational tests are pure power tests is inconsistent with reality because giving student unlimited time to finish any test rarely happens in practice, so the RT has more or less impact on student's performance. The ultimate goal here to using RT as CI is to make test that is fair to all candidates by accounting for the difference of personality or cognitive style among test takers, just like current practice in educational test that have to account for the gender and ethnicity difference through differential item functioning (DIF) analysis.

The results in this study show that in general, models with CI fit better than models without CI or reduced model with less CI. Besides the benefit of RT (for more information on advantages of using RT, see van der Linden, Klein Entink, & Fox, 2010) as one type of CIs in data –model fit, RT does not alternate the construct being measured in NS34 and LI52 and only increase the accuracy of estimates. However, construct measured by a given test can be improved by the construct measured by different test. As matter of fact, student's mathematics score (NS34) can improve item parameter estimation for students using their reading scores (LI52). It is important to emphasize that the improvement in estimates of item parameters using RT as CI in estimate ability parameters are not compared to any base model because separated analysis of theta and eta is not conducted because the limited time in this study.

Besides to joint in person (person as random effect in GLMM), it is also possible to conduct analysis on joint in items.

V. Scientific Significance of the Study

The advantages of IRT methods over traditional methods allow substantive researchers and testing practitioners to solve many difficult problems in activities such as scoring, equating, scaling, computerized adaptive testing, bias analysis, and so forth. Because these activities are important components in K-12, licensure, certification, and admission tests, accurately estimating IRT parameters plays a dominant role in psychometric research. For decades, researchers and practitioners have made a great deal of efforts to study a variety of methods to increase parameter accuracy, but only recently can, researchers start focusing on improving parameter estimations by using joint model that could incorporate RT and students information as CI. Given that many tests are currently administrated by computers and recorded RT as much other personal information is usually thrown out as test byproduct, how to use such available (may be valuable) information to improve the quality of many high stake tests has become urgent issues for states and test industries. So far in practice, the consequence of not using CI in estimation is ignored and few studies have focused on this issue; the present study attempts to provide empirical evidence on the consequence of ignoring CI on improvement parameter estimation.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Anastasi, A. (1976) *Psychological Testing* (pp. 32–34). New York: Macmillan.

Baxter, B. (1931). An experimental analysis of the contributions of speed and level in an intelligence test. *The Journal of Educational Psychology, 22*, 285-296.

Cox, D.R. and N. Wermuth, 1992. Response models for mixed binary and quantitative variable. *Biometrika, 79 (3)*: 441-461.

Dennis, I. & Evans, J. (1996). The speed-error trade off problem in psychometric testing. *British Journal of psychology, 87*, 105-129.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*, pp. 286–288. Englewood Cliffs, NJ: Prentice-Hall.

Fitzmaurice, G.M., & Laird, N.M. (1995). Regression models for bivariate discrete and continuous outcomewith clustering. *J. Am. Stat. Assoc., 90*: 845-852.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software, 20*, 1-14.

GLIMMIX Procedure Documentation. (2005). "*The GLIMMIX Procedure, Nov. 2005*", SAS Institute.

Hall, E. (2006). *Using Collateral Item and Examinee Information to Improve IRT Item Parameter Estimation*. Unpublished Dissertation. University of Iowa.

Klein Entink, R.H., Fox, J.-P., & van der Linden, W.J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers, *Psychometrika, 74,* 21-48.

Klein Entink, R. H., Kuhn, J.-T., Hornke L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times, *Psychological Methods, 14*, 54-75.

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

Luce, R. D. (1983). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Northwest Evaluation Association. (2009, April). *Technical manual for Measures of*

Academic Progress and Measures of Academic Progress for Primary Grades. Porttland, OR: Author.

Mislevy, R.J., & Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661-679.

Myers, C. T., (1960). Symposium: The effects of time limits on test scores. *Educational and Psychological Measurement, 20*, 221 -222.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185-205.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis, an introduction to basic and advanced multilevel modeling*. London: Sage Publishers.

Roskam, E. E. (1997). Models for speed and time limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York: Springer.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *73*, 287–308.

van der Linden, W. J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25–41.

van der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement, 47*, 92-114.

van der Linden, W. J. (2010). IRT Parameter Estimation With Response Times as Collateral Information. *Applied Psychological Measurement, 34*, 327-347

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365–384.

Van der Linden, W.J., Klein Entink, R.H., Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*.

Van Zandt, T.(2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.) *Stevens' Handbook of Experimental Psychology (3rd Edition), Volume 4: Methodology in Experimental Psychology (pp. 461-516)*. New York: Wiley Press.

Table 1. Summary of Nested Joint-Models Tested in This Study

Test	Model	Predictor	Model Formulation
NS34	1	Item (I)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + \lambda_j + e_{2i}$,
	2	Item (I), Raw-Score (R)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + NS34_i\,\gamma_{11} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + NS34_i\,\gamma_{21} + \lambda_j + e_{2i}$,
	3	Item (I), Gender (G)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + gender\,\gamma_{12} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + gender\,\gamma_{12} + \lambda_j + e_{2i}$,
	4	Item (I), Raw-Score (R) Gender (G)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + NS34_i\,\gamma_{11} + gender\,\gamma_{12} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + NS34_i\,\gamma_{21} + gender\,\gamma_{22} + \lambda_j + e_{2i}$,
	5	Item (I), Raw-Score (R) Gender (G), Ethnicity (E)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + NS34_i\,\gamma_{11} + gender\,\gamma_{12} + ethnicity\,\gamma_{13} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + NS34_i\,\gamma_{21} + gender\,\gamma_{22} + ethnicity\,\gamma_{23} + \lambda_j + e_{2i}$,
LI52	1	Item (I)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + \lambda_j + e_{2i}$,
	2	Item (I), Raw-Score (R)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + LI52_i\,\gamma_{11} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + LI52_i\,\gamma_{21} + \lambda_j + e_{2i}$,
	3	Item (I), Gender (G)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + gender\,\gamma_{12} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + gender\,\gamma_{12} + \lambda_j + e_{2i}$,
	4	Item (I), Raw-Score (R) Gender (G)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + LI52_i\,\gamma_{11} + gender\,\gamma_{12} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + LI52_i\,\gamma_{21} + gender\,\gamma_{22} + \lambda_j + e_{2i}$,
	5	Item (I), Raw-Score (R) Gender (G), Ethnicity (E)	$\text{Logit}(\mu_{1ij}) = \eta_{1ij} = \gamma_{10} + LI52_i\,\gamma_{11} + gender\,\gamma_{12} + ethnicity\,\gamma_{13} - b_j + e_{1i}$, $\text{Log}(\mu_{2ij}) = \eta_{2ij} = \gamma_{20} + LI52_i\,\gamma_{21} + gender\,\gamma_{22} + ethnicity\,\gamma_{23} + \lambda_j + e_{2i}$,

Table 2. Summary Results of Joint-Model Fit Statistics and Descriptive Statistics

Test	Model	Predictor[*1]	$\rho(\theta_i, \zeta_i)$	-2LL	G1=Δ(-2LL) [*2]	$p\ (\chi^2_x \geq G1)$ [*3]	G2=Δ(-2LL) [*4]	$p\ (\chi^2_1 \geq G2)$
NS34	1	I	-0.16582	591515.3				
	2	I, R	-0.18389	590839.1	-676.2	<0.001	-676.2	<0.001
	3	I, G	-0.16656	591611.8	96.5	<0.001	772.7	<0.001
	4	I, R, G	-0.18335	590918.7	-596.6	<0.001	-693.1	<0.001
	5	I, R, G, E	-0.16681	590781.8	-733.5	<0.001	-136.9	<0.001
LI52	6	I	-0.34411	340680.7				
	7	I, R	-0.29045	340618.8	-61.9	<0.001	-61.9	<0.001
	8	I, G	-0.34984	340678.9	-1.8	>0.25	60.1	<0.001
	9	I, R, G	-0.29643	340622.6	-58.1	<0.001	-56.3	<0.001
	10	I, R, G, E	-0.29052	340719.4	38.7	<0.001	96.8	<0.001

Note:

*1: I - Item, R-Raw Score, G - Gender, E – Ethnicity.

*2: Reduced model used here is the model with I only.

*3: The degree of freedom x for χ^2_x has range from 1 to 4.

*4: Reduced model used here is previous reduced model.

Table 3. Summary Statistics of Person Parameter Estimation for NS34 and LI52

Test	Model	Predictor	Variable	N	Mean	Std Dev
NS34	1	I	theta	1454	0.00	0.77
			SE_theta	1454	0.45	0.04
			Zeta	1454	0.00	0.28
			SE_Zeta	1454	0.08	0.00
	2	I, R	theta	1454	0.00	0.76
			SE_theta	1454	0.45	0.04
			Zeta	1454	0.00	0.28
			SE_Zeta	1454	0.08	0.00
	3	I, G	theta	1454	0.00	0.77
			SE_theta	1454	0.45	0.04
			Zeta	1454	0.00	0.28
			SE_Zeta	1454	0.08	0.00
	4	I, G, R	theta	1454	0.00	0.76
			SE_theta	1454	0.45	0.04
			Zeta	1454	0.00	0.28
			SE_Zeta	1454	0.08	0.00
	5	I, G, R, E	theta	1454	0.00	0.77
			SE_theta	1454	0.45	0.04
			Zeta	1454	0.00	0.27
			SE_Zeta	1454	0.08	0.00
LI52	6	I	theta	1454	0.00	1.54
			SE_theta	1454	0.97	0.36
			Zeta	1454	0.00	0.28
			SE_Zeta	1454	0.06	0.00
	7	I, R	theta	1454	0.00	1.55
			SE_theta	1454	0.98	0.36
			Zeta	1454	0.00	0.27
			SE_Zeta	1454	0.06	0.00
	8	I, G	theta	1454	0.00	1.54
			SE_theta	1454	0.97	0.36
			Zeta	1454	0.00	0.28
			SE_Zeta	1454	0.06	0.00
	9	I, G, R	theta	1454	0.00	1.55
			SE_theta	1454	0.98	0.36
			Zeta	1454	0.00	0.27
			SE_Zeta	1454	0.06	0.00
	10	I, G, R, E	theta	1454	0.00	1.55
			SE_theta	1454	0.98	0.36
			Zeta	1454	0.00	0.26
			SE_Zeta	1454	0.06	0.00

Table 4. Summary Statistics of Item Parameter Estimation for NS34 and LI52

Test	Model	Predictor	Variable	N	Mean	Std Dev
NS34	1	I	b	34	-1.98	1.69
			SE_b	34	0.11	0.05
	2	I, R	b	34	-1.85	1.69
			SE_b	34	0.12	0.05
	3	I, G	b	34	-1.96	1.69
			SE_b	34	0.11	0.05
	4	I, G, R	b	34	0.33	1.69
			SE_b	34	0.11	0.05
	5	I, G, R, E	b	34	0.33	1.69
			SE_b	34	-	-
	1	I	Lambda	34	-2.42	0.51
			SE_Lambda	34	0.02	0.00
	2	I, R	Lambda	34	-2.29	0.51
			SE_Lambda	34	0.06	0.00
	3	I, G	Lambda	34	-2.40	0.51
			SE_Lambda	34	0.02	0.00
	4	I, G, R	Lambda	34	-0.11	0.51
			SE_Lambda	33	0.02	0.00
	5	I, G, R, E	Lambda	34	-0.11	0.51
			SE_Lambda	33	0.02	0.00
LI52	6	I	b	52	-4.12	0.41
			SE_b	52	0.15	0.01
	7	I, R	b	52	-4.43	0.41
			SE_b	52	0.16	0.01
	8	I, G	b	52	-4.10	0.41
			SE_b	52	0.15	0.01
	9	I, G, R	b	52	-2.17	0.41
			SE_b	52	0.15	0.01
	10	I, G, R, E	b	52	-2.17	0.41
			SE_b	52	-	-
	6	I	Lambda	52	-2.02	0.04
			SE_Lambda	52	0.01	0.00
	7	I, R	Lambda	52	-2.34	0.04
			SE_Lambda	52	0.05	0.00
	8	I, G	Lambda	52	-2.00	0.04
			SE_Lambda	52	0.02	0.00
	9	I, G, R	Lambda	52	-0.08	0.04
			SE_Lambda	51	0.02	0.00
	10	I, G, R, E	Lambda	52	-0.08	0.04
			SE_Lambda	51	0.02	0.00

Table 5. Pearson Correlation Coefficients of Item Parameters (b and lambda) Estimations across Models for NS34 and LI52

Test	Variable	b_I	b_IR	b_IG	b_IRG	b_IRGE	Lambda_I	Lambda_IR	Lambda_IG	Lambda_IRG	Lambda_IRGE
NS34	b_I	1.00000	1.00000	1.00000	1.00000	1.00000	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945
	b_IR	1.00000	1.00000	1.00000	1.00000	1.00000	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945
	b_IG	1.00000	1.00000	1.00000	1.00000	1.00000	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945
	b_IRG	1.00000	1.00000	1.00000	1.00000	1.00000	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945
	b_IRGE	1.00000	1.00000	1.00000	1.00000	1.00000	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945
	Lambda _I	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IR	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IG	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IRG	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IRGE	-0.72945	-0.72945	-0.72945	-0.72945	-0.72945	1.00000	1.00000	1.00000	1.00000	1.00000
LI52	b_I	1.00000	1.00000	1.00000	1.00000	1.00000	-0.28759	-0.28759	-0.28759	-0.28759	-0.28759
	b_IR	1.00000	1.00000	1.00000	1.00000	1.00000	-0.28762	-0.28762	-0.28762	-0.28762	-0.28762
	b_IG	1.00000	1.00000	1.00000	1.00000	1.00000	-0.28759	-0.28759	-0.28759	-0.28759	-0.28759
	b_IRG	1.00000	1.00000	1.00000	1.00000	1.00000	-0.28763	-0.28763	-0.28763	-0.28763	-0.28763
	b_IRGE	1.00000	1.00000	1.00000	1.00000	1.00000	-0.28762	-0.28762	-0.28762	-0.28762	-0.28762
	Lambda _I	-0.28759	-0.28762	-0.28759	-0.28763	-0.28762	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IR	-0.28759	-0.28762	-0.28759	-0.28763	-0.28762	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IG	-0.28759	-0.28762	-0.28759	-0.28763	-0.28762	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IRG	-0.28759	-0.28762	-0.28759	-0.28763	-0.28762	1.00000	1.00000	1.00000	1.00000	1.00000
	Lambda _IRGE	-0.28759	-0.28762	-0.28759	-0.28763	-0.28762	1.00000	1.00000	1.00000	1.00000	1.00000

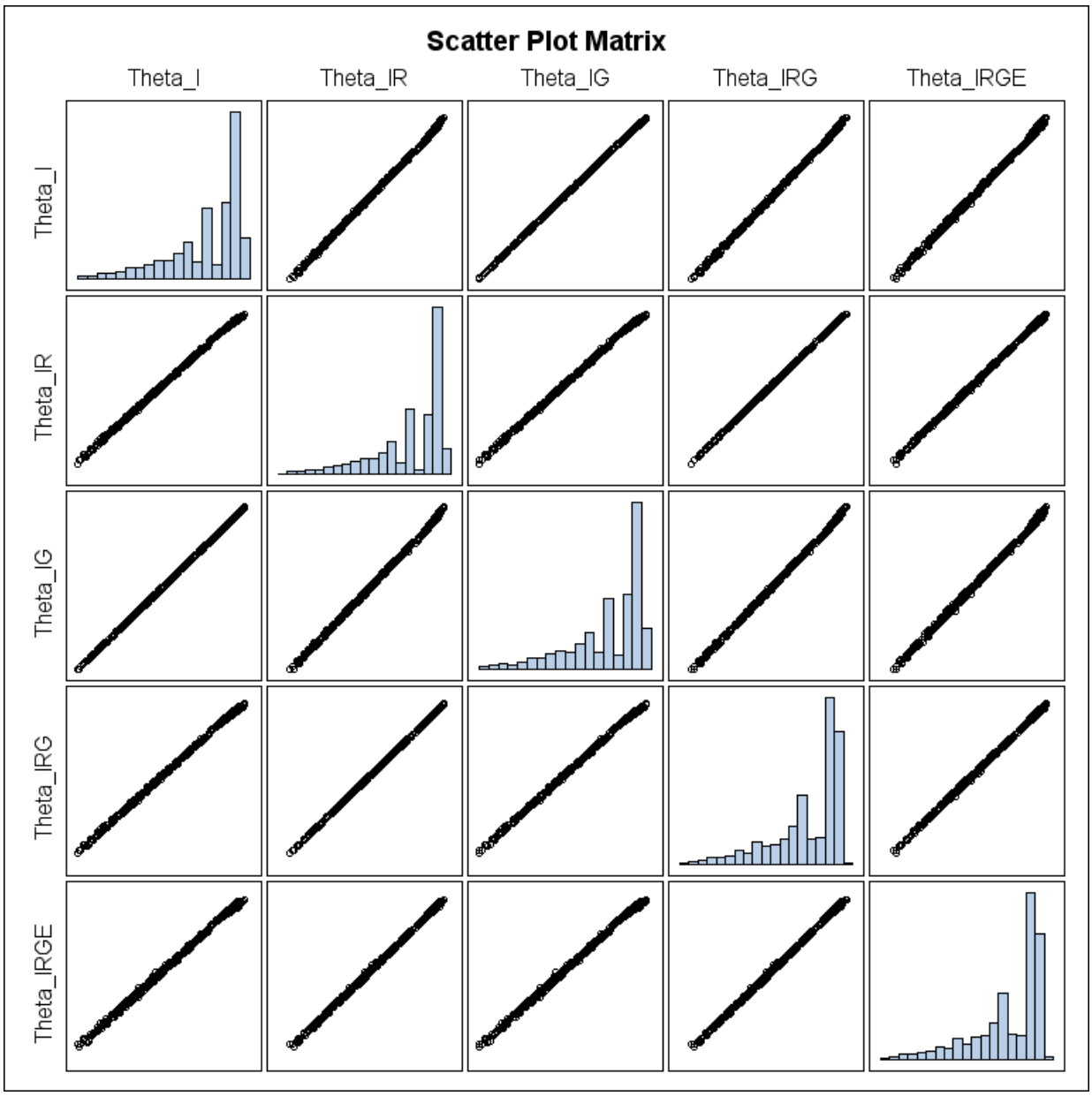

Figure 1. Scatter Plots of Person Ability Parameters Among Different Models for NS34 Test

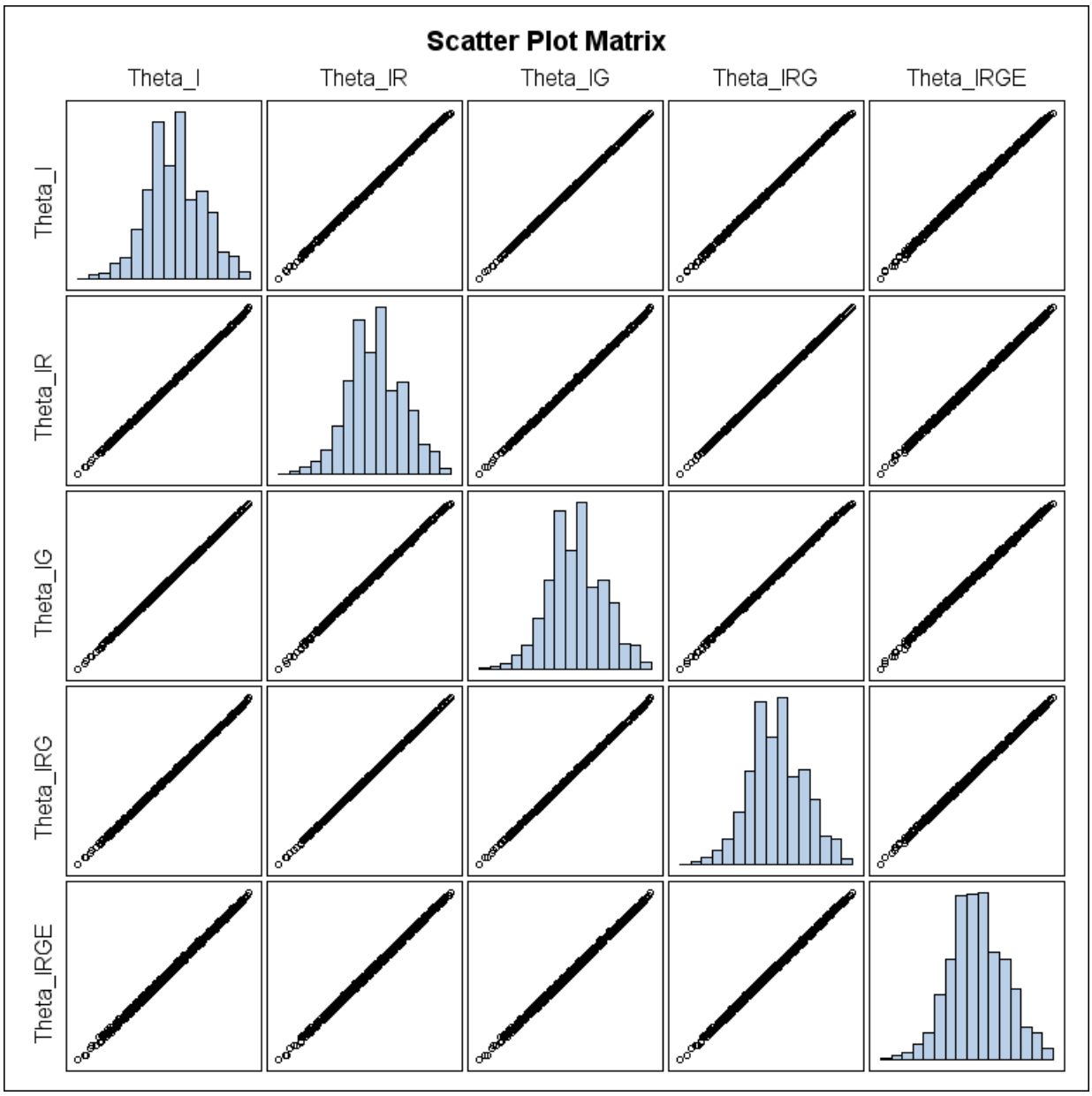

Figure 2. Scatter Plots of Person Ability Parameters Among Different Models for LI52 Test

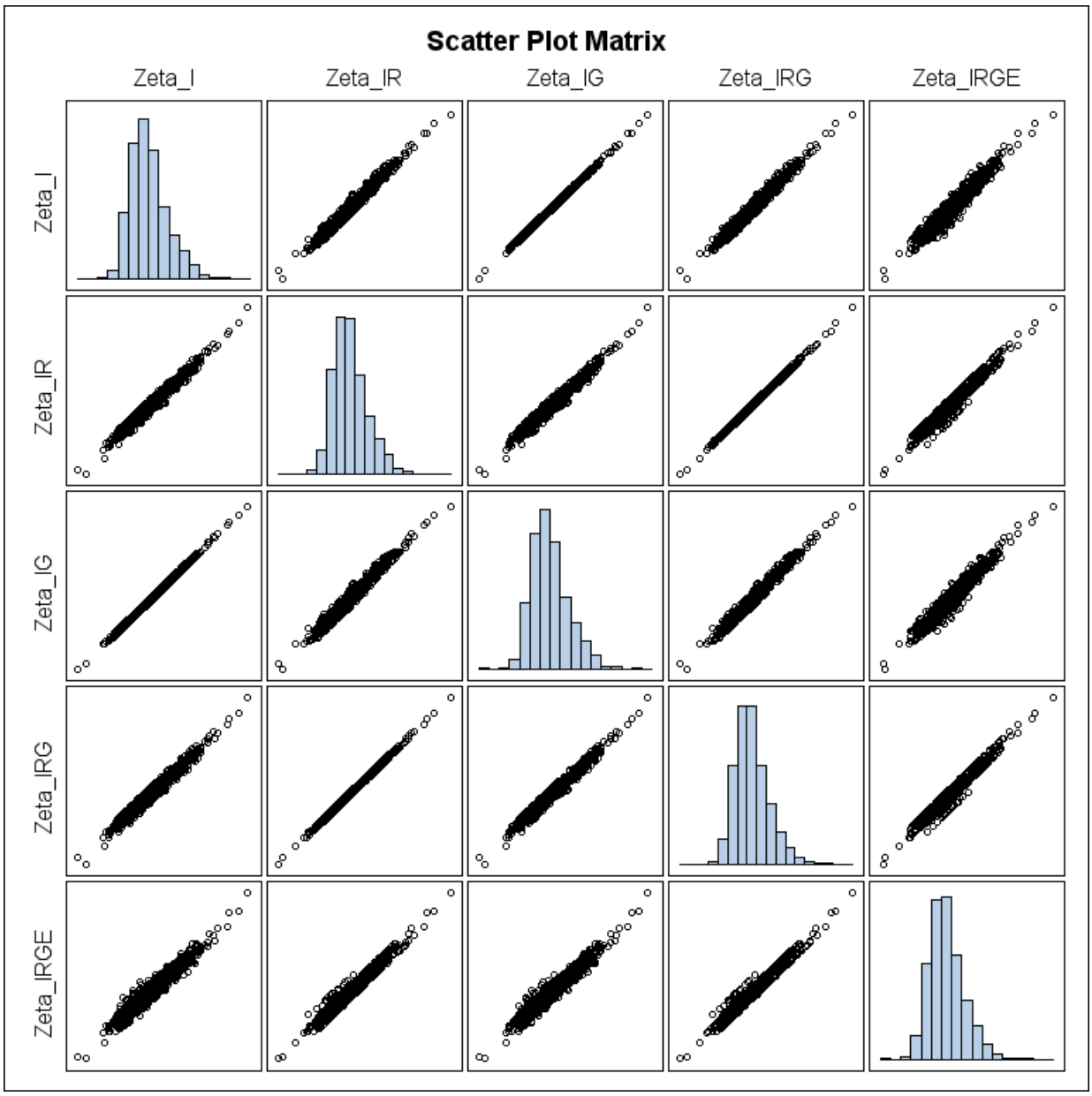

Figure 3. Scatter Plots of Person Speed Parameters Among Different Models for NS34 Test

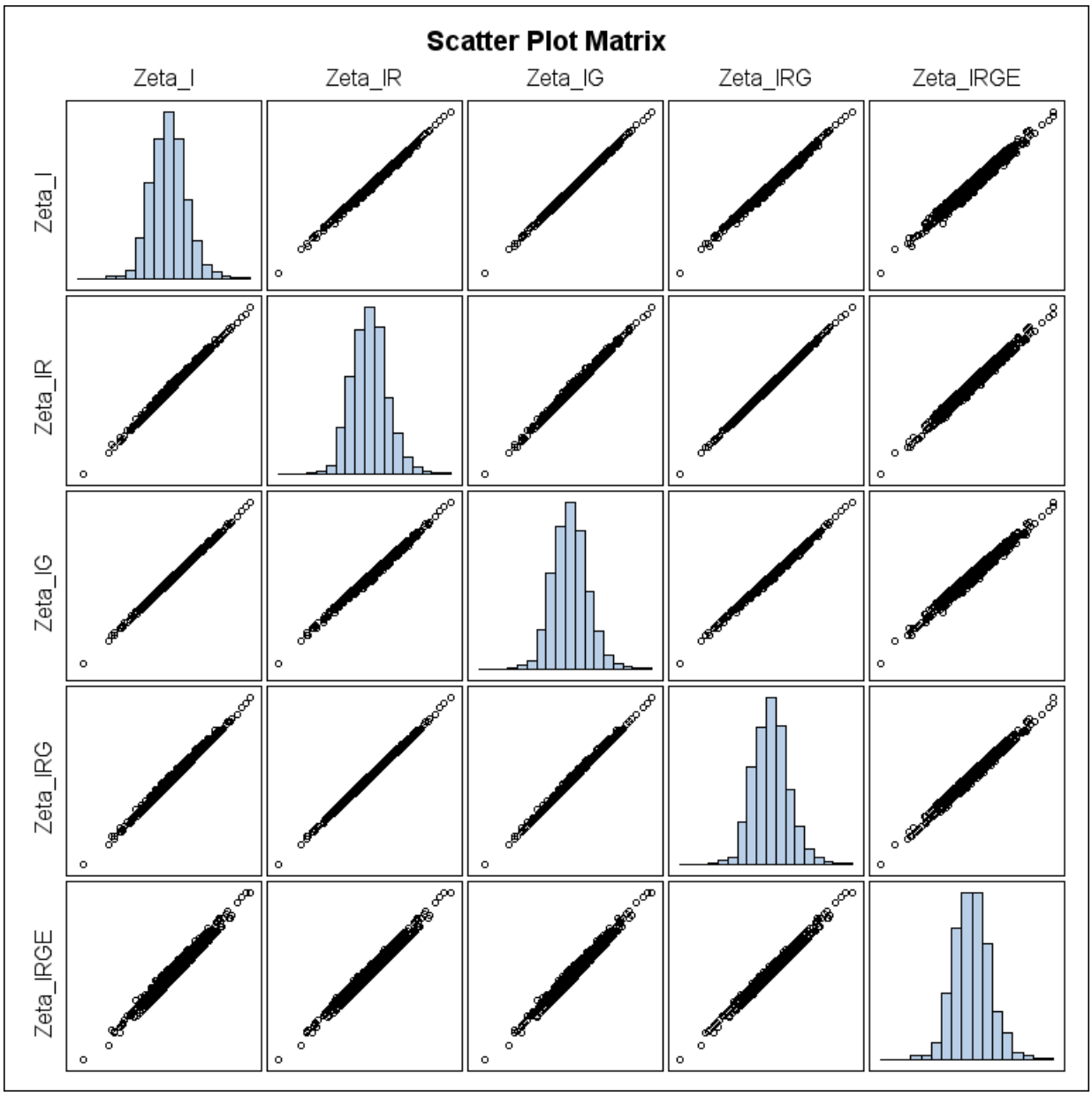

Figure 4. Scatter Plots of Person Speed Parameters Among Different Models for LI52 Test

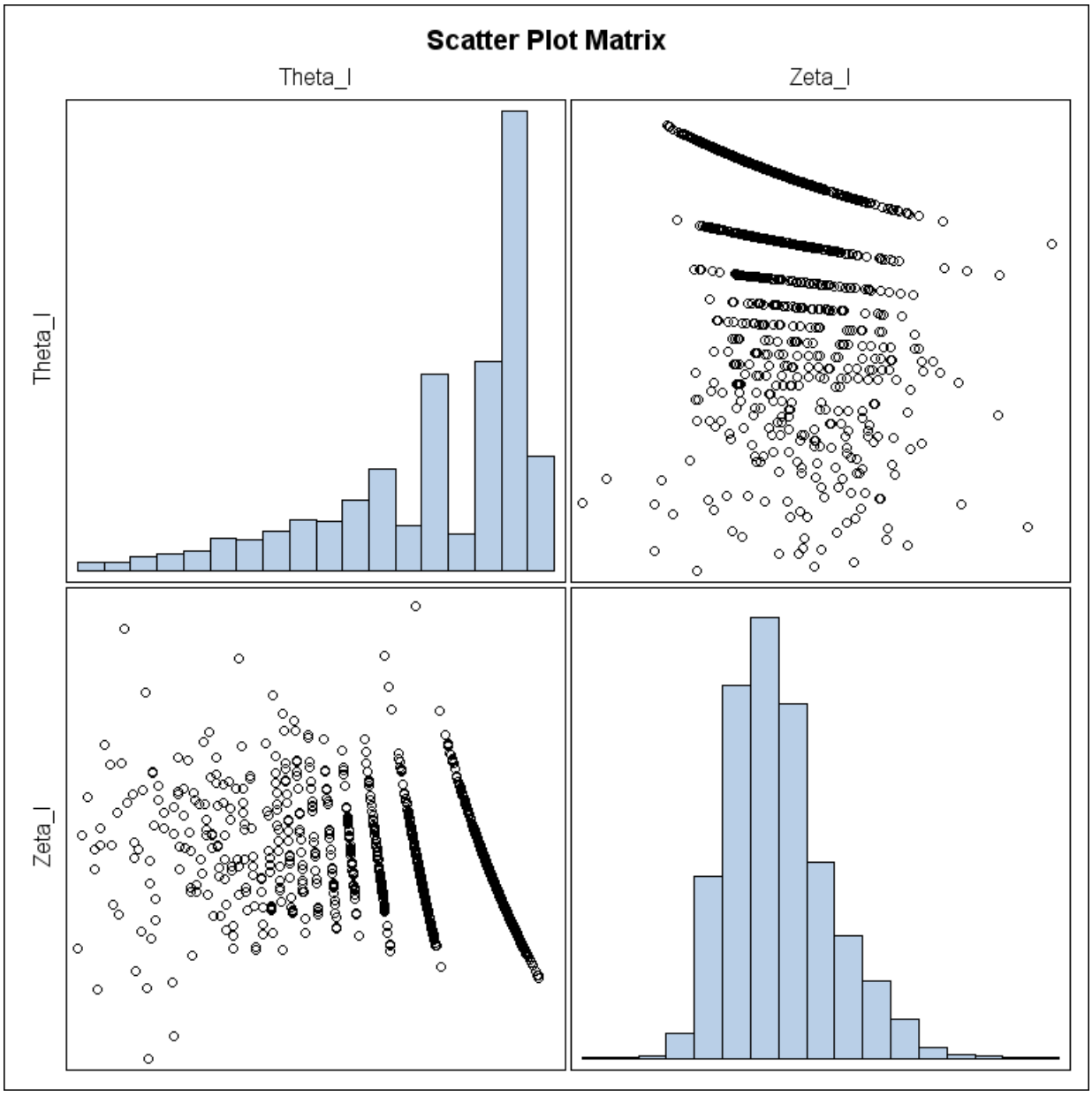

Figure 5. Scatter Plots of Person Theta and Speed Parameters from Base Model (1) for NS34 Test.

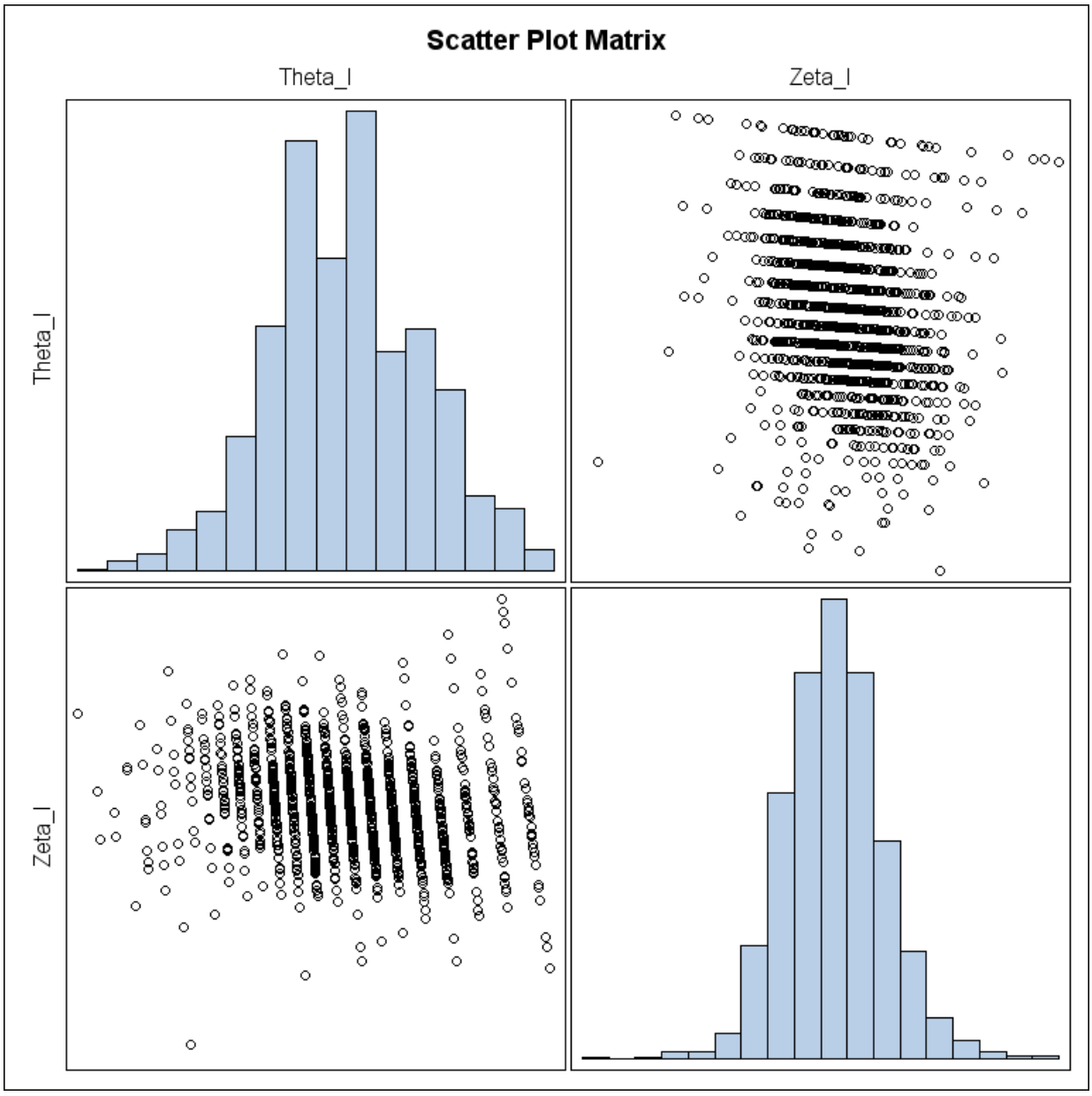

Figure 6. Scatter Plots of Person Theta and Speed Parameters from Base Model (6) for LI52 Test.

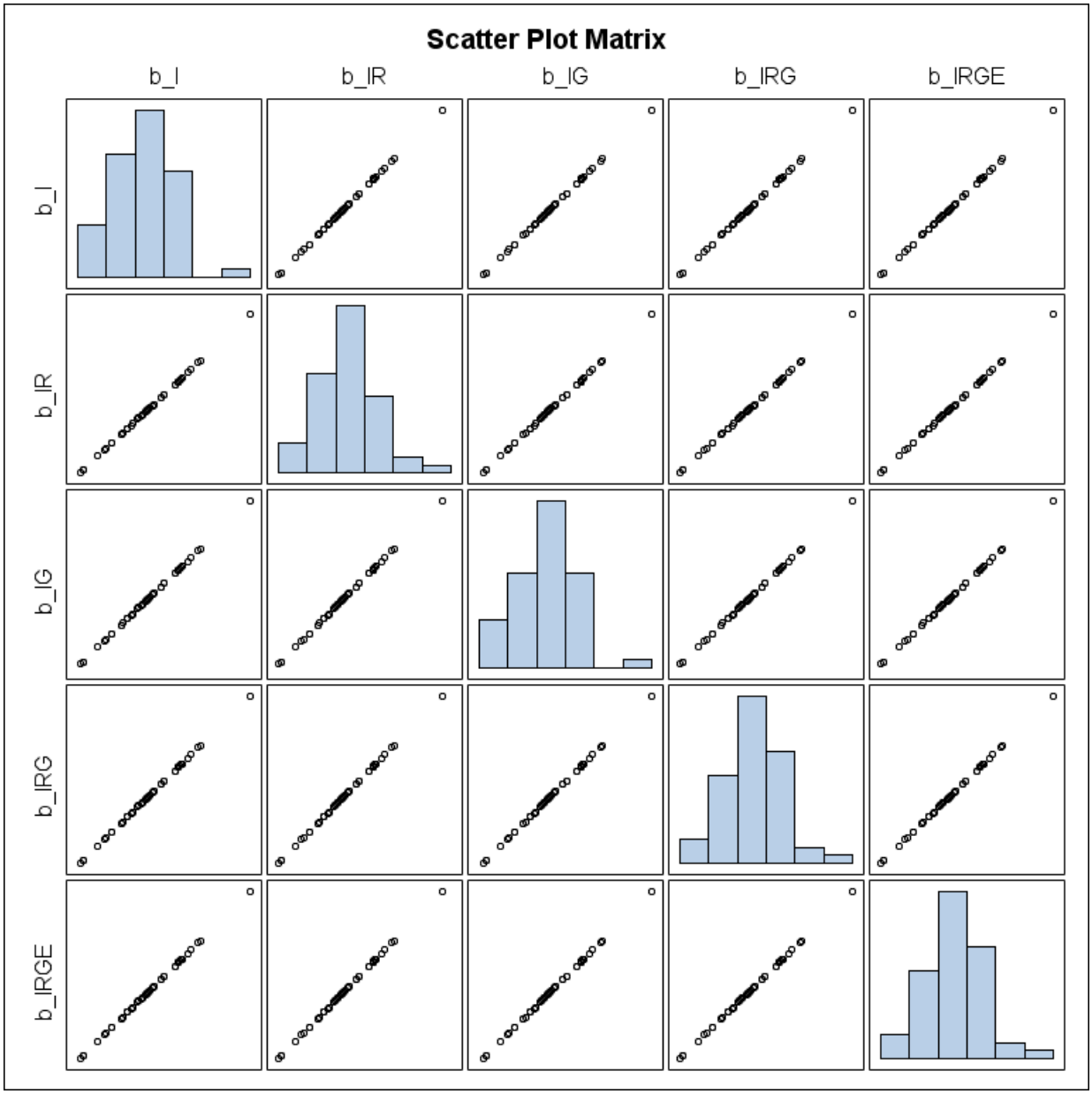

Figure 7. Scatter Plots of Item Difficulty Parameters Among Different Models for NS34 Test

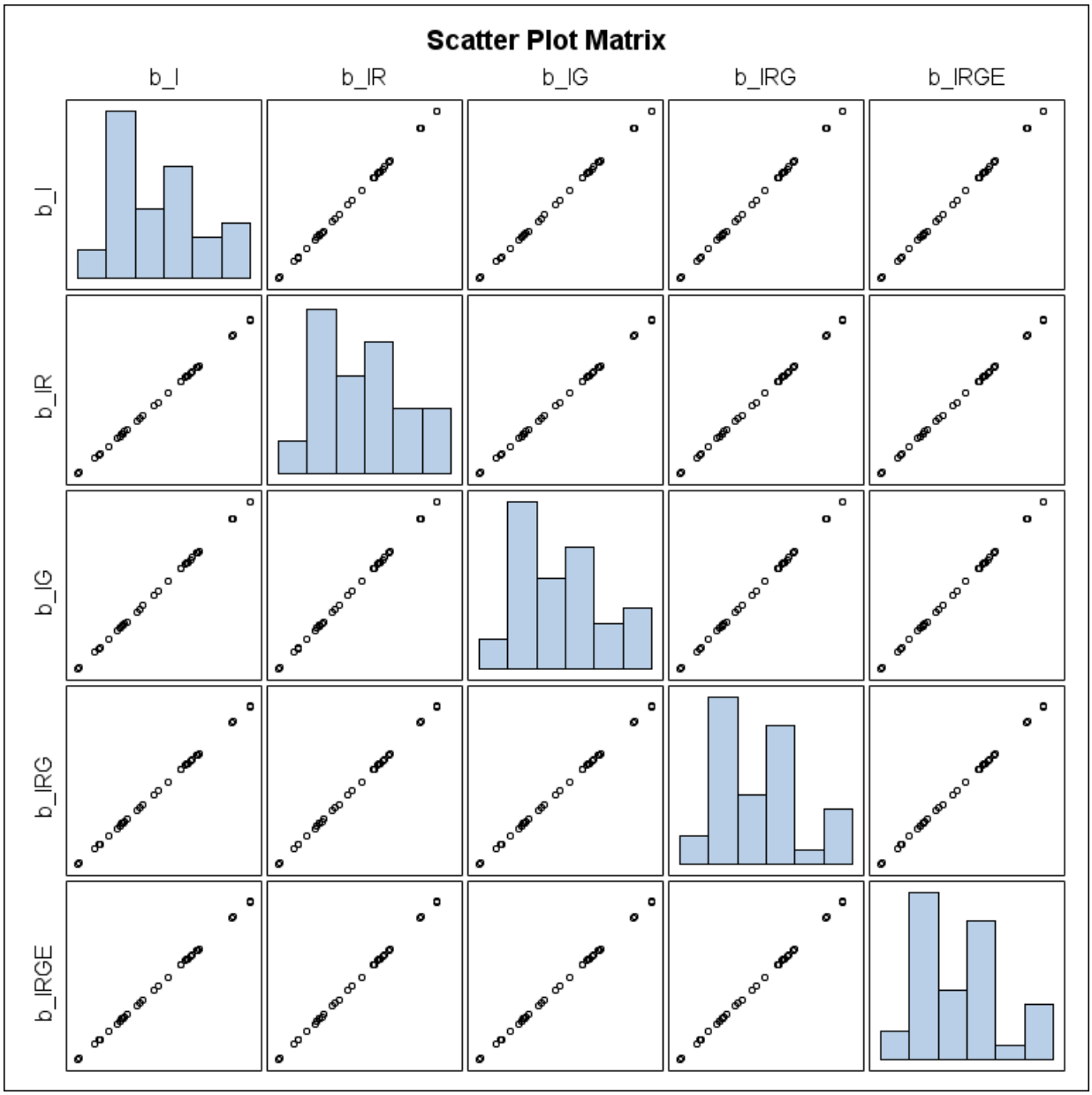

Figure 8. Scatter Plots of Item Difficulty Parameters Among Different Models for LI52 Test

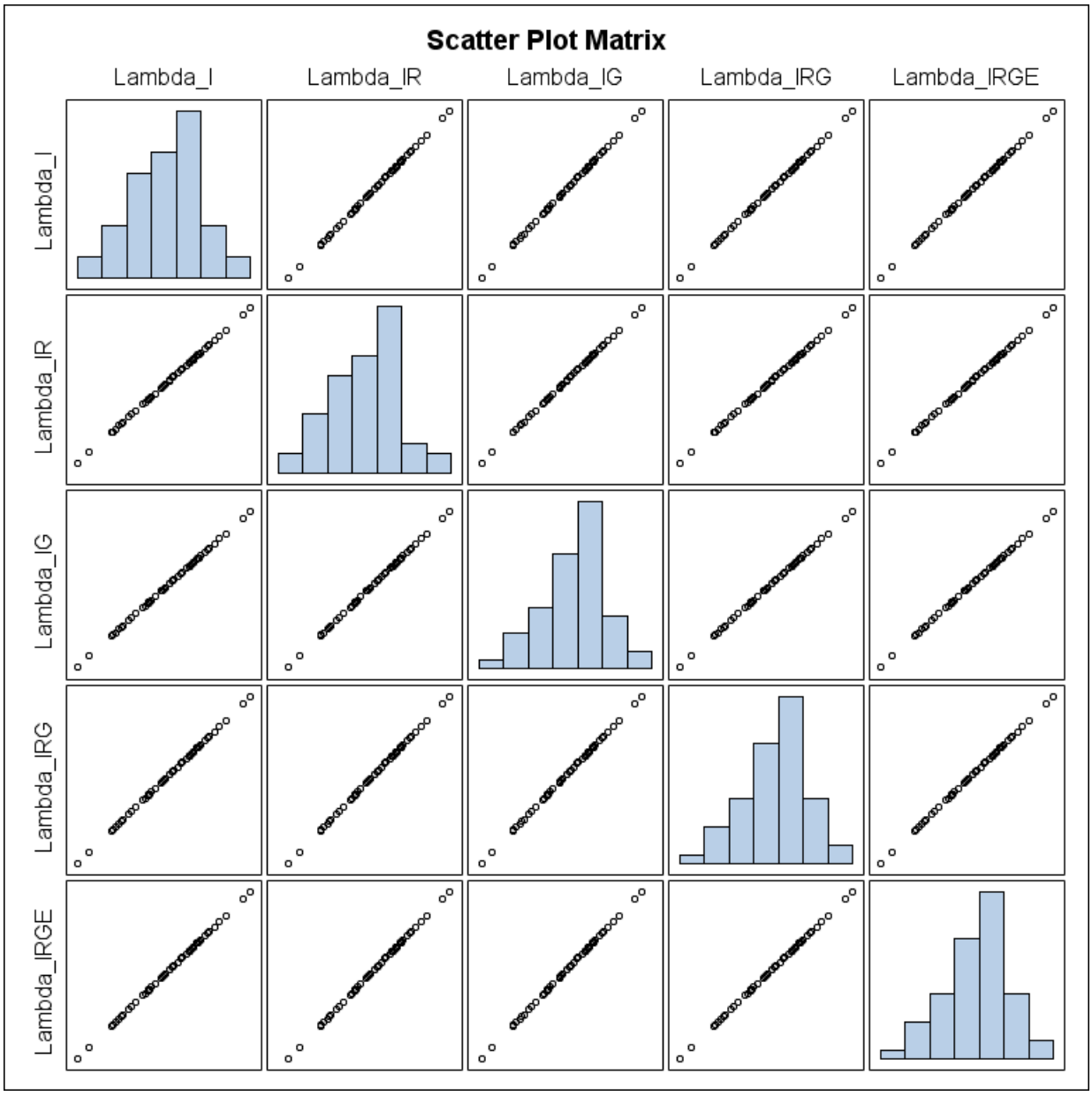

Figure 9. Scatter Plots of Item Time Intensity Parameters Among Different Models for NS34 Test

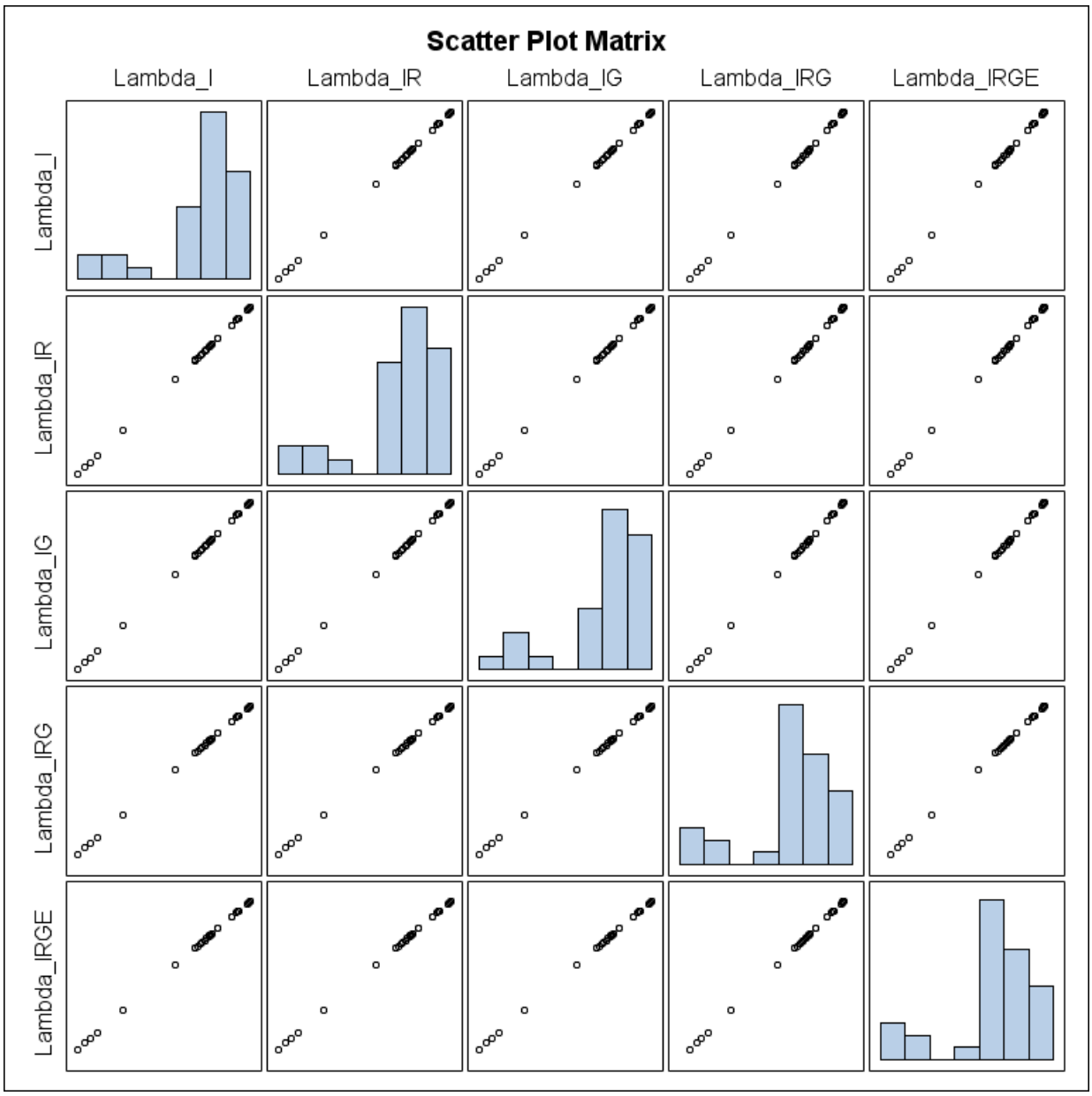

Figure 10. Scatter Plots of Item Time Intensity Parameters Among Different Models for LI52 Test

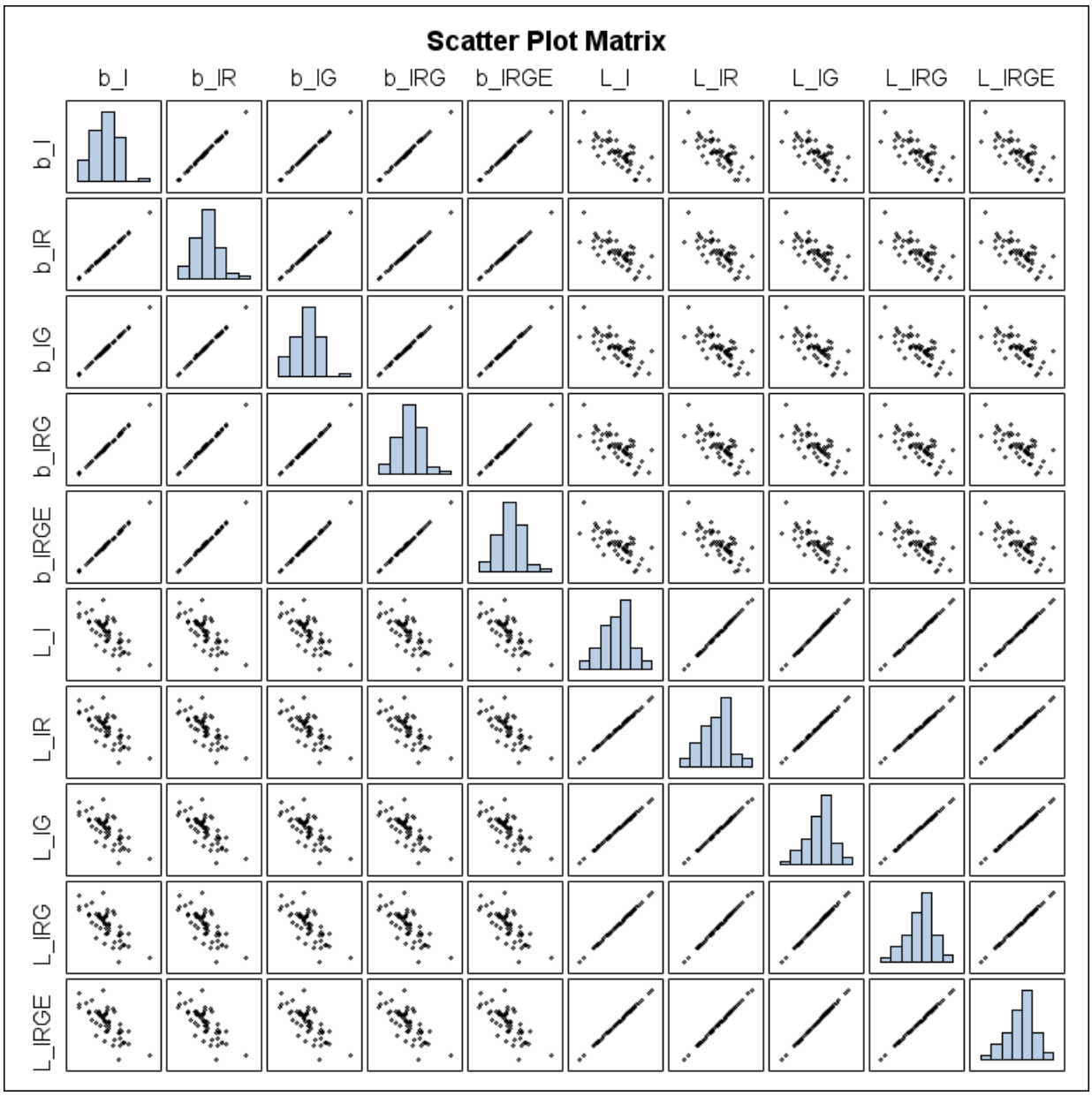

Figure 11. Scatter Plots of Item Difficulty and Time Intensity Parameters across Models for NS34 Test (L_ represents Lambda_).

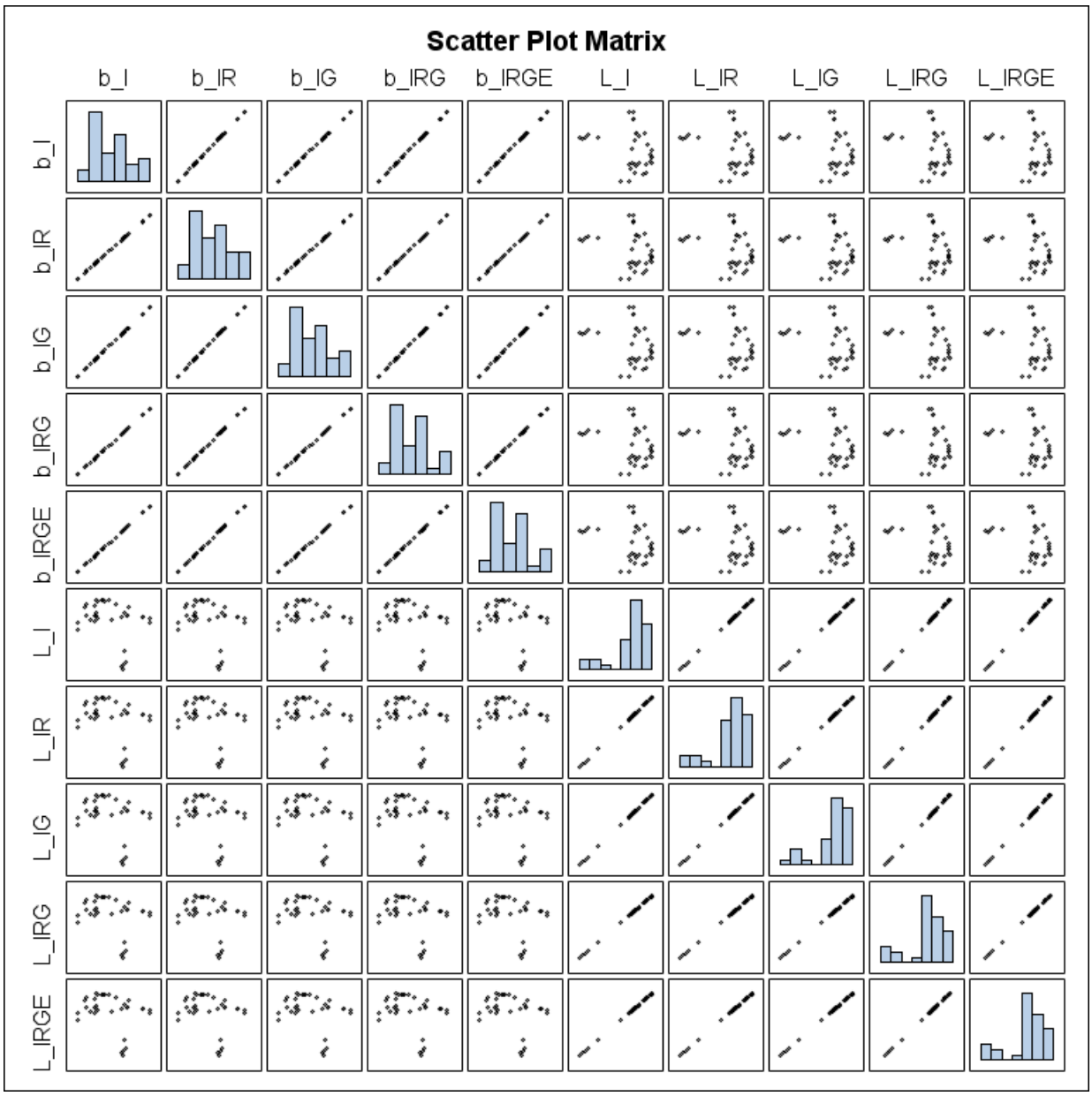

Figure 12. Scatter Plots of Item Difficulty and Time Intensity Parameters across Models for LI52 Test (L_ represents Lambda_).