Title
A Model-Building Approach to Assessing Q3 Values for Local Item Dependence

Author
Castaneda, Ruben

Publication Date
2017

UNIVERSITY OF CALIFORNIA, MERCED

A Model-Building Approach to Assessing Q3 Values for Local Item Dependence

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychological Sciences

By

Ruben Castaneda

Committee in charge:

Professor Jack L. Vevea, Co- chair

Professor Yang Liu, Co-chair

Professor Sarah Depaoli

April 2017

The dissertation of Ruben Castaneda Sanchez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Sarah Depaoli

Yang Liu

Co-Advisor

Jack Vevea

Chair

University of California, Merced

April 2017

This dissertation is dedicated to my family.

CONTENTS

Acknowledgements

I would like to thank the members of my dissertation committee: Jack L. Vevea, Yang Liu, and Sarah Depaoli. Jack L. Vevea has been my advisor since the moment I started at UC Merced five years ago. I came with an eager mind and hunger for knowledge. He took me in and gave me immense guidance in both my academic and personal life. I would not be here without his support. Yang Liu came into my graduate career later but has provided invaluable insight and perspective on item response theory. I have grown considerably in the short time since he agreed to co-advise me, and I thank him for this. Sarah Depaoli has been part of my committee and part of my academic life in multiple ways; her classes, stories, and words of encouragement have helped me immensely. Thank you, Sarah. I would also like to thank Nicole Zelinsky and my siblings Jannet, Mikey, Tony, and Michelle, as well as my parents, Rubén Castañeda Castillo and Raquel Sánchez.

CHAPTER 1

Introduction and Background

In item response theory (IRT), the local item independence (LII) assumption requires that the probability of responding to a set of items depends only on the latent trait. That is, once the latent trait is accounted for, the item residuals should not be correlated (Embretson & Reise, 2000; Kim, de Ayala, Ferdous, & Nering, 2011; Mcdonald, 1982). Proper test development techniques can minimize the possibility of violating the assumption of local item independence (LII) (Steinberg & Thissen, 1996; Yen, 1993). However, despite such efforts, local item dependence (LID, also known as LD) may still occur. In response to problems with LD, statistical techniques exist that can be applied to test data to ensure proper local item independence (Chen & Thissen, 1997; Ip, 2001; Yen, 1984). Many common LD statistics rely on exploratory methods and multiple pairwise comparisons. Conventional statistics like Yen's Q_3, Pearson's χ^2, and the likelihood-ratio test (G^2) require that researchers test each item pair in a measure. The cumulative type I error rate increases with each item pair tested, unless controlled using an error rate correction technique (Christensen, Makransky, & Horton, 2016; Glas & Falcon, 2003; Houts & Edwards, 2013; Ip, 2001; Kim et al., 2011; Thissen, Steinberg, & Kuang, 2002; Yen, 1984).

In this dissertation, I introduce the technique of modeling Yen's Q_3 values using a meta-analytic approach. This method relies on *a priori* knowledge of or suspicion about the location of LD. I summarize literature on the causes and indicators of LD, and then describe the model-building approach in the context of LD detection. I also demonstrate this method using freely available data online for both a progressive matrices type measure of intelligence (IQ1) and a measure of neuroticism (NPI). Because a large part of this dissertation relies on simulation

studies, I describe the steps and process of Monte Carlo studies. To demonstrate the effectiveness of the model-building approach, I conduct three simulation studies under different conditions and compare the results to existing methods of detecting LD. Finally, I summarize my results and describe the future directions of this project.

Item Response Theory

Item response theory (IRT) consists of a family of models that can be used to estimate latent trait ability (θ) and item characteristics based on a person's responses to a set of items (Embretson & Reise, 2000; Lord & Novick, 1968; Thissen & Steinberg, 2009). To describe individual item characteristics in binary data, researchers are typically interested in up to three or four model parameters, where the fourth, slipping, is rarely estimated in practice. Note that for polytomous data, researchers might estimate more than three parameters (e.g. one or two parameters per response category). First, we want to know the strength of the relationship between the item response and the latent trait. Second, we want to know the difficulty of each item. Finally, we want to know what a student's probability of successfully answering an item is if he or she is guessing and does not know the correct answer. This can arise, for example, if students employ a guessing strategy in educational assessments, or if participants respond randomly in personality testing.

Lord and Novick (1968) use the IRT parameters *a*, *b*, and *c* to denote these ideas. The *a* parameter explains the relationship between an item response and the latent variable, theoretically ranging from $-\infty$ to $+\infty$. The *b* parameter describes the difficulty of the item, also ranging from $-\infty$ to $+\infty$ and, in practice, usually ranging from -3 to 3. Finally, the *c* parameter describes the probability of guessing an item correctly (Birnbaum, 1968; Lord & Novick, 1968). With 4-option

multiple-choice items, a student employing a guessing strategy has a ¼ = .25 chance of answering each item correctly purely by chance. Note that the *c* parameter is free to vary, allowing for the possibility that some distractor options may draw the responders' attention more effectively than other items.

These ideas are also expressed in the form of item characteristic curves (ICCs; Lord & Novick, 1968) or trace lines (Thissen, Steinberg, & Mooney, 1989). In terms of ICCs, for a 2PL model the *a* parameter is the slope and the *b* parameter is the position on the curve where a responder has a 50% chance of endorsing the item. In the 3PL model, guessing is represented by the parameter *c*, the probability of answering the item correctly by chance. In the 3PL model, the interpretation of the *b* parameter also changes; it now represents the probability (1 + *c*)/2 people will answer correctly. If the *c* parameter is 0, the *b* parameter collapses to (1 + 0)/2 or 50%, matching the 2PL model.

IRT models for binary data include the Rasch model (Rasch, 1960), the one-parameter logistic (1PL) model, and the two-parameter logistic (2PL) and three-parameter logistic (3PL) models (Birnbaum, 1968; Lord & Novick, 1968). The Rasch model and the 1PL model are similar in that they only estimate item difficulty. The difference between these models is that the slope parameter, *a,* is fixed to 1 for all items in the case of the Rasch model and, rather than assuming that the underlying latent distribution is standard normal, its variance is estimated. The Rasch model, the probability of a response Y for person *i* and item *j* is:

$$P\left(Y_{ij} = 1 \middle| \theta_i\right) = \frac{1}{1+\exp(-1(\theta_i - b_j))}. \quad (1)$$

The 1PL model takes the same form but replaces 1 with the *a* parameter, where *a* is the estimated slope constrained to be equal across all items. The distribution of ability in the 1PL model is assumed to be standard normal.

The 2PL model frees the slope parameter, estimating one unique slope per item. This allows the model to specify different relationships between each item and the underlying latent trait. The 2PL model is:

$$P\left(Y_{ij} = 1 \middle| \theta_i\right) = \frac{1}{1+\exp(-a_j(\theta_i - b_j))}. \tag{2}$$

Finally, the 3PL model adds a guessing parameter that accounts for possible guessing behavior. The 3PL model is:

$$P\left(Y_{ij} = 1 \middle| \theta_i\right) = c_j + (1 - c_j)\ \frac{1}{1+\exp(-a_j(\theta_i - b_j))}. \tag{3}$$

Researchers can apply each of these models to a multitude of psychological measures. The researcher can consider all models and choose which one to apply (Embretson & Reise, 2000). Factors to consider include the sample size, the construct of interest, and, ultimately, model fit (Harwell, Stone, & Hsu, 1996; Kean & Reilly, 2014; Koning, Sijtsma, & Hamers, 2002; Thissen & Steinberg, 2009). As the number of parameters increases, the models need a greater number of observations to produce a valid estimate (Thissen, Reeve, Bjorner, & Chang, 2007). A smaller number of observations can result in the model failing to converge to a stable solution or yielding imprecise parameter estimates. Often, even if convergence is achieved, the parameter estimates have larger standard errors, indicating greater uncertainty. Regarding general application, a Rasch or 1PL model is appropriate if the researcher expects all items to have the same relationship to the latent variable (e.g. a set of multiplication items). The 2PL model is more suitable if different items probe the construct of interest at different levels (e.g. some items require complex math, while some require only addition). The 3PL is applicable if participants may respond to an item correctly even when their latent ability is lower than the item difficulty predicts.

Assumptions of IRT

The assumptions of IRT vary depending on the model used to estimate the item parameters. For traditional likelihood-based unidimensional IRT models, there are three basic assumptions (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Items should have a monotonic relationship with their underlying latent trait. Researchers should specify the appropriate dimensionality; for example, in the case of unidimensional models, items should only load onto a single latent trait. Finally, item responses should be independent of one another, conditional on ability.

In IRT, these assumptions are strong and difficult to avoid (Hambleton & Jones, 1993; Yen, 1993). Mcdonald (1982) proposes a weaker assumption for item independence in that variables are uncorrelated, after accounting for the latent trait, result in meeting this assumption. It is accepted that the assumptions for IRT will be violated to some degree in one way or another; however, the degree to which these assumptions are violated can be minimized by using proper test development practices and statistical methods for detecting these violations.

Monotonicity is the idea that as ability increases, so too should the probability of answering an item correctly (Holland, 1981; Junker & Sijtsma, 2000). This relationship allows us to model item responses with an S-shaped curve, which in turn allows for easier interpretation of item parameters. This S-shaped curve in IRT is called the item response function (IRF). The IRF is often modeled as a logistic curve (Koning et al., 2002). Violations of the assumption of monotonicity can lead to model misfit and inaccurate slope parameters (Yen, 1993), and such violations can result from misrepresenting psychological constructs. For example, imagine a multiple-choice question that is scored as correct/incorrect with an overwhelmingly high distractor item that catches overthinkers (Beier & Oswald, 2012). Students with lower ability

levels may select the right answer, but responders with a higher ability may choose the distractor because the right answer seems too obvious. Situations like this are easily avoidable by using basic assumption checking methods, like those developed by Junker and Sijtsma (2000). Simply regressing the item response onto a total score or regressing the response onto $\sum_i^n x_i - x_{\neq i}$ can provide insight into the items' monotonicity. In this equation *i* is the individual item, *n* is the total number of items, and *x* is the item's score (0,1).

The assumption of unidimensionality (which must be met for unidimensional IRT) indicates that a single latent trait explains the relationships among all of the items (McDonald, 1981). Generally, the assumption is that researchers must specify the correct number of dimensions; however, this dissertation focuses on single-dimension IRT models. In practice, most educational tests include a broad set of questions designed to probe a certain domain of knowledge. For example, an elementary school math test might include questions about addition, subtraction, multiplication, and division, all of which fall under the broad knowledge domain of arithmetic (a unidimensional trait). Adding word problems to this test, however, would add a second latent trait, language. Failing to model this second trait using multidimensional IRT would result in a violation of unidimensionality. Researchers can assess unidimensionality by applying exploratory non-linear factor analysis and identifying the number of underlying factors (Hattie, 1985). Other methods exist, but are deemed unsatisfactory (Embretson & Reise, 2000; Hattie, 1985).

Like the assumption of unidimensionality, local item independence (LII) assumes that the response to one item is not dependent on the response to other items after accounting for the main latent trait. When $\phi(\cdot)$ is the density function for the standard normal distribution, this is written as:

$$f(y_i) = \int \prod_{j=1}^{m} f_j(y_{ij}|\theta)\phi(\theta)d\theta. \quad (4)$$

After integrating out the latent trait θ, the probability of the response pattern y_i for person i is the product of all the individual responses from j to m items.

After the latent trait has been accounted for, the item residuals should not correlate (Embretson & Reise, 2000; Lord & Novick, 1968); if they do correlate, this is known as underlying local dependence (ULD) The distinction between multidimensionality and ULD is purely conceptual, dealing with the substantive interpretation of the secondary factor (Chen & Thissen, 1997; Houts & Edwards, 2015). Chen and Thissen (1997) defined the distinction between underlying LD and surface LD. Underlying LD (similar to multidimensionality) occurs if a subset of items is related due to an underlying latent variable apart from the main latent variable of interest (Houts & Edwards, 2013). ULD is caused by underfactoring – that is, extracting too few latent variables to adequately capture the responses for a set of items (Ip, 2010). Surface LD, on the other handconsists of a subset of items that are related for other reasons , such as sharing a characteristic that causes participants to answer them similarly (Chen & Thissen, 1997).

Local dependence (LD) exists when something drives a person to answer two items similarly regardless of their own latent ability, whether due to a surface characteristic or a secondary underlying latent trait. Local dependence is broken down into two categories: underlying local dependence and surface local dependence. Underlying LD occurs when two or more items probe an underlying ability not measured by the rest of the items. For example, word problems in a math exam probe a student's ability to understand language, not just mathematics, and hence would exhibit underlying LD. Surface local dependence is caused by a probability linking two or more items to the same response unconditional on latent ability level (e.g., a responder who runs out of time and fails to answer the last two items).

Violations of Local Independence

In educational testing, LD can occur if the answers to a series of items are linked (e.g., a reading passage followed by questions about that passage), if a student fails to answer the last few questions (e.g., speeded tests or long tests), or if items probe multiple constructs (e.g., word problems in a math test; Yen, 1993).

Yen (1993) describes multiple situations in educational testing that can cause LD, and explains how these situations manifest in the residual correlation. In her paper, she says that proper test administration, following adequate testing procedures, and evaluating items as they are being developed can minimize some of these situations. She also identifies some procedures that can cause LD: external assistance, interference, and speededness. External assistance causes LD when teachers or test administrators deliver the answer or hints about the answer to the test takers. Interference can cause LD if it interrupts the test procedure; for example, a fire alarm may accidentally go off, ruining the students' concentration, or test material may be printed incorrectly, resulting in test administrators having to re-issue tests. Mishandling the proctoring of exams can also result in LD if students do not receive the full amount of time required for the test.

Causes of LD during test development include fatigue, item response formats, passage dependence, item chaining, and explanations from previous answers. Fatigue can arise due to long tests or lengthy passages included near the end of tests (Brennan, 1992; Yen, 1993). If tests are long, the closer to the end of the test items are placed, the larger their slope parameters. In that case, the slope parameter can no longer be interpreted as the strength of the relationship between the item and the underlying latent variable; it is confounded with test fatigue (Masters, 1988). Item formats can also be culprits in violations of LD. Formats that are new to students may confuse students into answering items incorrectly. If students are familiar with a specific format (e.g., multiple choice), they may know that only one of the answers is correct (Brennan, 1992;

Ferrara, Huynh, & Michaels, 1999; Thissen et al., 1989). On the other hand, if students receive multiple-choice items with multiple correct responses, they may mistake these items for multiple-choice questions with only a single correct choice. In cases like these, the probability of endorsing an item is clearly not only due to the latent variable, but also to confusion about the new item type. More commonly, item type is known to cause LD in passage formats, where small sets of items are linked to specific passages (Lee, 2004; Thissen et al., 1989; Wainer & Wang, 2001; Yen, 1984). In these situations, students' interest can also come into play. A student who is interested in the passage may pay closer attention to the questions than an uninterested student.

Item chaining and answers obtained from previous questions may also be causes of LD (Liu & Thissen, 2014). Item chaining is a specific test format where students are asked a series of questions, each relying on the correct response to an earlier item. For example, a student might be asked to compute the mean and standard deviation of a dataset, then asked to calculate a *z*-score based on that mean and standard deviation. If the student incorrectly computes the mean, their *z*-score will also be incorrect. If the answer to an item can be obtained from a previous item, this can have the same effect as item chaining (Brennan, 1992). Imagine the student is, again, asked to calculate the mean and standard deviation of a dataset; the next question asks the student to identify the correct formula for the mean. If the student recognizes the correct formula for the mean, they can answer the previous question with information that was unavailable before.

Instruments used in psychological tests are much different from those used in education. One difference is that psychological tests are intended to measure latent traits that express an opinion or a degree of endorsement, rather than proficiency. Participants are rarely encouraged to study for anxiety questionnaires. Because of this, the causes of LD in personality and psychological tests tend to differ from those in education. Shrout and Fiske (2014) highlight

three unique causes of LD in psychological testing: context effects, serial order effects, and similar or redundant questions.

Context effects can occur if people develop stronger, more salient definitions of the words used in a questionnaire as they progress through the items (Brennan, 1992; Ferrara et al., 1999). For example, if someone is asked whether they feel secure at the beginning of a test, they may interpret this as regarding a sense of financial security, or a sense of being safe from harm. However, as they encounter more items using the word "secure," they may come to understand that security in this context means freedom from care, freedom from anxiety, or emotional security. Such context effects are associated with seriation or ordering effects (Thissen et al., 1989; Zenisky, Hambleton, & Sireci, 2001). The farther along the item is in the test, the stronger the relationship between it and the latent trait, because the item is now interpreted in the context of the test.

Items asking similar or redundant questions may also result in violations of local dependence. Questions with similar wording, regardless of whether the phrasing seems distinctly different to the researcher, may be indistinguishable to the participant (Steinberg & Thissen, 2014). Assume a test includes the items "I feel lonely" and "When I am with friends, I feel alone." Both questions ask about loneliness and may elicit similar responses, although one is about general loneliness and the other about similarity to a friend group. Another possible reason for this type of response may come from the idea of participants' salient memories. Taylor and Fiske (1978) illustrate this concept with an example. If someone is asked "Do coworkers make you anxious?" they may immediately remember a particularly obnoxious coworker they dislike. If, later in the measure, the item "People make me anxious" appears, they may return to the memory of that coworker and, thus, answer this item in a similar manner. The two items address different questions, but the concept, to the test taker, is the same.

These examples highlight only a small portion of current knowledge about sources of LD in education and psychology. Only proper test development and administration can drastically reduce many unintended causes of LD. However, even after common causes of LD are addressed during test development, multiple LD indices exist that can identify any LD that is still present. Identifying any remaining LD is important in high-stakes testing, where even a slight error in calculating a score may mean the difference between passing and failing (e.g., certification exams).

Current LD Indices

Yen's (1984) Q_3 statistic involves correlating the residuals (i.e., the observed response minus the expected response) for two items. Let $r_{ij} = X_{ij} - E_{ij}$ be the residuals, where X_{ij} is the observed response for the *i*th item and *j*th observation and E_{ij} is the expected response pattern for the *i*th item and *j*th observation. The values $\bar{r}_i$ and $\bar{r}_k$ represent the means of r_{ij} and r_{ik} across subjects. The formula is denoted as

$$Q_{3\ \mathrm{ik}} = \frac{\sum_{j=1}^{N}(r_{ij}-\bar{r}_j)(r_{ik}-\bar{r}_k)}{\sqrt{\sum_{j=1}^{N}(r_{ij}-\bar{r}_i)^2}\sqrt{\sum_{j=1}^{N}(r_{jk}-\bar{r}_k)^2}}\,. \tag{5}$$

In other words, the Q_3 statistic is a correlation coefficient between item residuals, which can be tested using a *t* distribution with *n* - 2 degrees of freedom. In practice, this method is expected to produce large Type I errors, because the Q_3 index has a small negative bias.

Fisher's *r*-to-*z* transformed Q_3, proposed by Yen (1993), has a sampling distribution that better approximates the normal distribution with a known variance. The equation for this transformation is

$$z = 0.5\ ln\frac{1+Q_3}{1-Q_3}, \tag{6}$$

and the variance is approximately 1/(n - 3), where n is the number of subjects.

Chen and Thissen (1997) compared the detection rate and power of several statistics and found that the r-to-z transformation of the Q_3 values does not produce Type I error rates that are close to the nominal alpha values of .05 or .10. In fact, Zenisky, Hambleton, & Sireci (2003) found that Q_3 values tend to have a negative correlation; this is due to part-whole contamination, stemming from using the estimated examinee ability level θ to calculate both the expected score and the observed score.

In general, the detection rate of Q_3 values tends to be liberal. Chen & Thissen (1997) suggested that it is possible to estimate the five percent cutoffs by simulating IRT data with specific item parameters and computing the Q_3 values under a null model (with no LD present). Using this distribution of Q_3 values, we can see the specific values at the top and bottom five percent of the distribution. The distribution of Q_3 values tends to be heavy tailed and slightly asymmetrical. As an alternative, Chen and Thissen (1997) suggested using a uniform distribution with a cutoff of .20 as a critical value. However, they found that the use of .20 resulted in very low power to detect LD (50% power). Although a .20 cutoff value produces low powered statistics, liberal detection rates may be preferable to a large Type I error rate.

According to Lee (2004), one of the benefits of Q_3 statistics is that they can be treated as raw correlation coefficients and interpreted accordingly; therefore, we can apply hypothesis testing using the standard error $1/\sqrt{(1-r^2)/(n-2)}$. Although the transformed Q_3 values resemble a Gaussian distribution, the transformed r-to-z values, like the original Q_3 values, are also negatively biased. Therefore, they also do not produce empirical Type I error rates that are close to the expected rates from a standard normal distribution.

A more recent study by Christensen et al. (2016) examined the properties of Yen's Q_3 statistic via a series of simulations. They found that the distribution of Q_3 values is highly dependent on the data itself, and using a clear cutoff value may be inappropriate for some testing conditions (e.g., a small number of items, a small sample size). To remedy this, they propose a relative approach to evaluating Q_3, which they call Q_3*. Q_3* is the difference between the maximum observed Q_3 value and the individual Q_3 values, or $Q_{3jk}^* = Q_{3max} - Q_{3jk}$. The authors propose that, although no clear cutoff exists for this statistic, a cutoff of .20 for Q_3* may be appropriate.

Pearson's χ^2 test was originally meant to evaluate whether differences in a cross tabulation of data might have arisen by chance. Chen and Thissen (1997) proposed using this statistic with pairs of items to assess whether the items are statistically similar. The equation is

$$X_{jk}^2 = \sum_{y_i=0}^{1} \sum_{y_k=0}^{1} \frac{\left(p_{y_j y_k} - \hat{\pi}_{y_j y_k}\right)^2}{\hat{\pi}_{y_j y_k}}, \tag{7}$$

such that $\hat{\pi}_{y_j y_k} = \Pr(Y_j = y_j, Y_k = y_k)$ expresses the bivariate cell probability under the model, and $p_{y_j y_k}$ denotes its sample proportion. Chen and Thissen (1997) suggested evaluating this statistic using a χ^2 distribution with one degree of freedom.

Chen and Thissen (1997) used the χ^2 statistic to detect local dependence in both null and non-null conditions. They found that, under the null condition, χ^2 was slightly conservative (i.e., the Type I error rate was below .05 for all conditions). More specifically, when this method was applied to the 2PL model with no LD, 1.6% - 2.9% of replications were significant, rather than 5%. The Type I error rate was closer to alpha when χ^2 was applied to the 3PL model, this time ranging from 2.6% to 3.6%. The rejection rate increased as the number of items increased.

Liu and Thissen (2012) found that the reference distribution of the X^2 statistic was stochastically smaller than the actual chi-square distribution, resulting in a conservative rejection rate (with Type I error rates lower than alpha). These results were corroborated later (Liu & Maydeu-Olivares, 2012); in addition to under-rejecting, the X^2 statistic rejected too frequently (was too liberal) in the presence of substantial model misspecification (i.e., a bifactor model and threshold shift).

The likelihood-ratio statistic for local item independence (G^2) compares the observed item responses to the expected responses for all item pairs. Chen and Thissen (1997) introduced this statistic as a means of assessing LD. G^2 is calculated as

$$G^2 = -2\sum_{i=1}^{p}\sum_{j=1}^{q} O_{ij} ln\left(\frac{E_{ij}}{O_{ij}}\right), \tag{8}$$

where p and q denote the number of response categories for items i and j (i.e., two for binary items). The observed number of respondents in category i of the first item and category j of the second item is denoted by O_{ij}. The expected number of respondents in the same contingency table cell is denoted by E_{ij}. This statistic also approximately follows a χ^2 distribution, with one degree of freedom if no LD is present.

In 1997, Chen and Thissen used the G^2 statistic to detect local dependence in null and non-null conditions. The performance of G^2 was nearly identical to the performance of Pearson's χ^2. Under the null condition, the Type I error rate for G^2 was near the nominal alpha rate of .05 for all conditions. Specifically, the false rejection rate was between 1.8% and 3.1%, using a χ^2 distribution with one degree of freedom. With the 3PL model, the rate was closer to alpha, ranging from 2.6% (a 10 item test) to 3.6% (an 80 item test). The rejection rate approached nominal levels as the number of items on the test increased.

The G^2 statistic detected local item dependence only about 40% of the time for the 2PL and 3PL models when strong underlying local dependence (ULD) was present. With weaker ULD, the G^2 statistic detected about 37% of incidents for the 2PL model and 31% for the 3PL model (Chen & Thissen, 1997). Kim et al. (2011) investigated the performance of the G^2 statistic in comparison to other directional and non-directional LD statistics. They found that G^2 had nominal detection rates under non-LD conditions and adequate power (defined as power over 65%) under moderate LD conditions.

The threshold shift score test (known as S_t) was introduced by Glas and Falcon (2003) and Liu and Thissen (2012; 2014). The threshold shift statistic evaluates the level of dependence between two items by adding another parameter to the item response model. The model is

$$T_q\left(1\middle|\theta; x_p\right) = \frac{1}{1+e^{-1(a_q\theta_i - c_q - \delta_{pq}x_p)}}, \tag{9}$$

where q and p are two items, and δ_{pq} is the threshold shift for the second item when the response x_p is 1. The δ_{pq} parameter can be evaluated using a score statistic (Glas & Falcon, 2003; Liu & Thissen, 2012), estimating the model as δ_{pq}= 0 and evaluating the derivative with respect to δ_{pq} to determine whether freeing δ_{pq} provides better model fit. This statistic is also approximately distributed χ^2 with one degree of freedom. Unlike the previous methods, the S_t statistic is an *a priori* approach.

Researchers investigated the threshold shift statistic under null and non-null conditions. Glas and Falcon (2003) found that its Type I errors rates were nominal (about .05) in the null conditions. When surface LD was introduced, its power ranged between 16% and 99%. Low sample sizes (n < 1,000) generally resulted in lower power than large sample sizes (n > 4,000). Liu and Maydeu-Olivares (2012) obtained similar results. Liu and Thissen (2014) tested multiple forms of the threshold shift statistic (e.g., linear shift, uniform shift, and conditional shift) and

found that these variations hold adequate Type I error rates under sufficiently large sample size and test lengths.

One major problem with all these traditional item by item testing approaches (e.g., Q_3, G^2 and $LD\text{-}X^2$) is that they rely on multiple pairwise comparisons. No methods currently exist that can avoid the multiple testing problem by focusing on specific item groups. Even when an LD statistic could be used on a specific set of items, current software (i.e., IRTPRO, Cai, Thissen, & du Toit, 2011) does not allow users to select which items they want to test, instead forcing them to make every possible pairwise comparison. Making multiple comparisons is especially problematic for educational measures with a large number of items, since the number of such comparisons increases exponentially (e.g., 15 items yield 105 pairwise comparisons; 20 items yield 190). These existing methods of identifying LD produce an excessive number of comparisons, and they do not allow for testing specific hypotheses about the location of local dependence.

Assessing Local Dependence (Model-Building Approach)

I introduce the model-building approach as a new method for analyzing Q_3 values to detect LD. Traditionally, observed Q_3 values are compared to the expected Q_3 values under a null distribution, but the model-building approach treats the analysis of Q_3 values as a special case of meta-analysis. We implement a random-effects meta-analytic model, with the *r*-to-*z* transformed Q_3 values serving as the effect sizes. We are then able to model suspected local dependence by way of a dummy-coded moderator, with a value of one indicating items suspected of exhibiting LD and zero denoting non-LD items. The variance component in this model accommodates the heterogeneity produced by local dependence in the Q_3 values. The pseudo-likelihood equation of the model-building approach is

$$L(\boldsymbol{\beta}, \tau^2 | \boldsymbol{X}, \boldsymbol{r}, N) \propto -\prod_{i=1}^{k(k-1)/2} \left[\left(\frac{1}{\sqrt{1/(N-3) + \tau^2}} \right) \exp\left(-\frac{1}{2} \frac{\left(Z(r_i) - \boldsymbol{X}'\boldsymbol{\beta}\right)^2}{1/(N-3) + \tau^2} \right) \right]. \quad (10)$$

In this model, we estimate a vector of regression parameters $\underline{\beta}$ and the between-item variance τ^2, conditional on the design matrix $\boldsymbol{X}$, the Q_3 values $\boldsymbol{r}$, and sample size N for k items.

The moderator variable is evaluated using the Q-between statistic (Lipsey & Wilson, 2001), computed as

$$Q_B = \sum_{i=1}^{k} w_{i*} (\bar{T}_{i*} - \bar{T}_{**})^2. \quad (11)$$

This is equal to the sum of squares of the weighted grand mean $(\bar{T}_{**})$ minus the mean of the effect sizes for the group of interest $(\bar{T}_{i*})$. The weights are expressed as $1/(V_i + VC)$, where V_i is the conditional variance and VC is a between-items variance component. The resulting Q-between value is evaluated as a χ^2 statistic with $(k - m)$ degrees of freedom, where k is the number of categories defined by the dummy-coded moderator and m is the number of moderators. Additionally, the individual regression parameters can be tested via a standard z-test. The random-effects model requires only one moderator (Table 1.1), because the variance component accounts for any excess heterogeneity. A fixed-effect model can also be used for this analysis but requires $(m - 1)$ moderators, where m is the number of items in the IRT model. The $(m - 1)$ set of moderators, designated as the "complex structure," reflects the involvement of specific items in calculating correlations. For example, the correlation between residuals for item 1 and item k would be coded '1' for the dummy variable $m1$, indicating the involvement of item 1 and all other items,1 through 5 (Table 1.2). This variation inherently produced by Q_3 is of little interest and adds needless complexity to the model. In practice, simply allowing the variance component of a random-effects model to describe that variability is much more efficient. Finally, the moderator

md reflects the location of suspected local dependence among items (e.g., items 1 and 2 are not suspect and are coded ‘0’; items 3 and 4 *are* suspect and are therefore coded ‘1’).

Table 1.1
Simple Structure

	Design Matrix	
	m0	md
$Q_{3\ 1,2}$	1	0
$Q_{3\ 1,3}$	1	0
$Q_{3\ 1,4}$	1	0
$Q_{3\ 1,5}$	1	0
$Q_{3\ 2,3}$	1	1
$Q_{3\ 2,4}$	1	1
$Q_{3\ 2,5}$	1	0
$Q_{3\ 3,4}$	1	1
$Q_{3\ 3,5}$	1	0
$Q_{3\ 4,5}$	1	0

Table 1.2
Complex Structure

	Design Matrix					
	m0	m1	m2	m3	m4	md
$Q_{3\ 1,2}$	1	1	1	0	0	0
$Q_{3\ 1,3}$	1	1	0	1	0	0
$Q_{3\ 1,4}$	1	1	0	0	1	0
$Q_{3\ 1,5}$	1	1	0	0	0	0
$Q_{3\ 2,3}$	1	0	1	1	0	1
$Q_{3\ 2,4}$	1	0	1	0	1	1
$Q_{3\ 2,5}$	1	0	1	0	0	0
$Q_{3\ 3,4}$	1	0	0	0	1	1
$Q_{3\ 3,5}$	1	0	0	1	0	0
$Q_{3\ 4,5}$	1	0	0	0	1	0

Meta-Analysis

Understanding the model-building approach requires a basic understanding of meta-analysis. Traditionally, meta-analysis is a statistical approach to synthesizing the results from multiple studies that address a specific research question (e.g., "What is the impact of the medication fluoxetine on generalized anxiety?"). Combining the results from multiple studies increases researchers' power to detect effects that may be difficult to identify due to low sample sizes or weak measures.

Once the variables of interest are identified, the researcher must decide between a fixed-effect meta-analysis and a random-effects meta-analysis. Fixed-effect (or FE) meta-analysis postulates that there is one correct, or true, effect size. Any variation in the observed effect sizes is either due to sampling error or to systematic heterogeneity (the result of a moderator). With a fixed-effect model, we assume that, using large enough sample sizes, the conditional means of the observed effects is the true effect. Written as a linear model, the fixed-effect meta-analysis is

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e_i, \tag{12}$$

where the observed effect size of each study Y_i is a function of the true mean effect size β_0, a set of $\beta_p \mathrm{X_p}$ moderators, and the sampling variability ε_i, which is regarded as fixed because it is a function of sample size.

In comparison, the random-effects meta-analytic model makes very different assumptions. The random-effects model supposes that the variation between observed studies may be due in part to inherent differences among the populations from which the effects were drawn, not associated with any specific moderator. Essentially, the random-effects (or RE) model assumes that the effect sizes are a random sample from a population of studies of interest. The RE portion of the model can be written as a linear mixed function

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + u_i + e_i, \quad (13)$$

where the observed effect size of each study Y_i is a function of the mean effect size β_0, a vector of β_p of coefficients associated with p fixed-effect moderators, the sampling variability ε_i, and the random effect u_i. The random-effects model can be conceptualized as a variation of the fixed-effect model, including a random intercept.

Simply put, researchers should use the fixed-effect model when they are interested in the effects of only one specific set of studies, and the random-effects model when they wish to estimate the variance associated with sampling true effects from a hyperpopulation of true effect sizes.

The most important distinction between the FE model and the RE model involves the universe to which one wishes to generalize. The FE model addresses what would happen if new primary sampling units (in psychology, these are typically people) were sampled into *the same set* of studies. The RE model, on the other hand, addresses what would happen if both new primary units and *new studies themselves* were sampled, and allows researchers to generalize their conclusions to the entire universe of possible studies, rather than only to observed studies. When we apply that distinction to the analysis of Q_3 statistics in the model-building approach, the FE model addresses the question "Do these items in *this particular test* have local dependence?" However, if researchers are more interested in generalizing beyond the particular test at hand and wish to ask something like "Does using multiple choice items in a predominantly true/false test cause local dependence?" the RE model is the appropriate choice.

In the analysis of LD, researchers are often concerned with a particular test, so they may consider the FE model. However, as described in Hedges and Vevea (1998), if we want to deal with heterogeneity, rather than our results' implications, we can use either the FE or RE to

address our questions. Therefore, both the FE and RE models work with the model-building approach – but the FE model is needlessly complicated due to the uninteresting systematic heterogeneity inherent in generating Q_3 values. To conduct an FE analysis with the model-building approach, in addition to the moderator modeling local dependence, we must generate a moderator for each pair of Q_3 values that deals with the same item. For example, a 15-item measure will produce 105 unique Q_3 values. From these 105, the first 14 effect sizes were generated from correlating item 1 with item 2, item 1 with item 3, and so on. The second 13 effect sizes correlated item 2 with item 3, item 2 with item 4… . If these relationships are not included in the FE analysis, the LD moderator might be flagged as statistically significant even if there is no LD, because it will attempt to account for some of the inherent heterogeneity in the data. On the other hand, if we conduct an RE analysis, this complex moderator structure provides no useful information and can be ignored because it will be captured by the random-effects portion of the model (τ^2). I conducted a simulation comparing the results of an FE model with a complex moderator structure and an RE model; these models were equivalent, and both reached the correct conclusions about the location of LD. For this reason, and because of the simplicity that the RE model provides, this dissertation will focus on the RE model exclusively.

CHAPTER 2

Fixed-Effect vs Random-Effects

Purpose

This study aims to investigate the performance of the model-building approach using both fixed-effect (FE) and random-effects (RE) models; when working with item pairs, I also include Chen and Thissen's (1997) $LD\text{-}X^2$ and $LD\text{-}G^2$ statistics for comparison. Hedges and Vevea (1998) mention that the choice between the FE and RE models is based on the assumptions made about the true effect size(s). With a fixed-effect (FE) model, the researcher assumes that there is a single (but unknown) true effect; with a random-effects (RE) model, they assume there is a distribution of true effects with an unknown mean and variance. In the FE model, the only source of variability is the sampling process. The RE model incorporates both sampling uncertainty and uncertainty due to the variance of the population of true effects. The choice between using an FE and RE model in meta-analysis, and hence in the model-building approach, depends on the desired generalizability of the results. The question of "Do these characteristics cause LD?" requires the RE model, whereas "Do these characteristics cause LD in *this* test?" requires the FE model. Educational test development typically focuses on the second question – the issue of detecting LD in one particular test.

Cochran's Q-test is a simple method of assessing effect sizes for heterogeneity is (Cochran, 1954; Hoaglin, 2016). In the application of the model-building approach, this Q-test is nearly always significant even in the absence of LD, due to the partial correlations between items' residuals used to compute the Q_3 values (as detailed in Chapter 1). To address this issue, we can use a set of moderators that capture the relationships between items used to compute the Q_3 values; this reduces the heterogeneity, which in turn reduces the Q-test value. In this simulation, I

propose a set of cells to identify which method captures the true effect more often under various conditions. I test the FE and RE methods using both simple-moderator and complex-moderator structures. To identify the heterogeneity accounted for by the moderators, I use a test described by Hedges and Olkin (1985). Essentially, the *Q*-statistic is split into within- and between-study variability. Viechtbauer (2010) uses the notation *QM* to denote *Q*-between, notation that I will continue to use throughout this dissertation. The individual regression coefficients can be evaluated with the test statistic $\hat{\beta}$/SE, using a *z*-distribution with a cutoff value of 1.96 for α = .05. In the case of a single moderator, the *p*-value for *QM* will equal the *z*-test *p*-value.

Finally, note that the likelihood function, as described in chapter 1, is not the true likelihood of the parameters given the data, because the function assumes independence. As such, this simulation treats the model as a method that either supports or rejects the approximation.

Methods

In this first simulation, I investigate the performance of the FE and RE meta-analytic models with both simple and complex moderator structures. I generate dichotomous data using the base *R* package (Team, 2014). To simulate underlying local item dependence (ULD), I randomly generate ability values from a bivariate normal distribution with varying degrees of correlation $r(\theta_1, \theta_2)$ that indicate different levels of LD strength (e.g., ρ =1, .20, .50). A correlation of 1 indicates that there is no LD present (the relationship between the two latent variables is exactly the same). A correlation of .50 indicates a moderate degree of LD (the two latent variables have a correlation of .50). Finally, a correlation of .20 indicates a strong degree of LD; items generated from θ_1, are correlated only .20 with items generated from θ_2.

Three factors are varied across cells: number of binary items, number of observations, and LD strength. The levels for the first factor, the number of binary items, are 15, 25, 30, and 50. I obtained these values by examining the number of items used in other IRT LD simulation studies (Christensen et al., 2016; Jansen, 2007; Liu & Thissen, 2012) and applied research (Arthur & Day, 1994; Bacon, Scheltema, & Robinson, 2001; Cooper & Petrides, 2010; Jorm, 1994; Le Moine, Fiestas-Navarrete, Katumba, & Launois, 2016; Milian et al., 2015; Pianta, 1992). The levels for the second factor, number of observations, are 1,000 and 4,000. Some typical levels observed in simulation studies are 200, 500, and 1,000 (Christensen et al., 2016; Glas & Falcon, 2003; Houts & Edwards, 2015; Liu & Thissen, 2014). However, in keeping with a simulation by Glas and Falcon (2003), I added 4,000 as an upper limit because such a larger number provides better estimates of IRT parameters and represents the typical scaling samples for commercial test development. I dropped the 200 and 500 conditions due to restrictions of computing resources and time. The slope parameters for this simulation were obtained from a log normal distribution with a mean of 0 and a standard deviation of 0.50 (Chen & Thissen, 1997; Liu & Thissen, 2012). The difficulty parameter was drawn from a standard normal distribution, and the guessing parameter was sampled from a logit normal distribution with a mean of -1.1 and standard deviation of 0.5 (Chen & Thissen, 1997). The first and second items of the test will exhibit LD in the non-null conditions (LD placement will not affect simulation performance). The correlation $\rho = 1$ reflects the null condition, where there is only one underlying latent trait (i.e., no LD). I use the MIRT *R* package to estimate the item parameters for a 3-parameter logistic (3PL) model and to generate the Q_3 values (Chalmers, 2012).

Two factors vary within cells, the type of meta-analytic model and the type of predictor (simple vs. complex). To estimate the meta-analytic models, I use the R package *metafor* (Viechtbauer, 2010). Levels of this factor include both a random-effects meta-analytic model and a fixed-effect meta-analytic model. As described in Chapter 1, a fixed-effect (FE) model makes

theoretical sense because we are not trying to generalize the results of our data to all possible tests with LD; we are only interested in this particular test. However, the Q_3 statistic is more variable at smaller sample sizes, even using Fisher's *r*-to-*z* transformation, and using a random-effects meta-analytic model will compensate for the additional variability. Each model is specified twice, once with a simple structure (the moderator models only LD) and once with a complex structure (the moderators model not only the participation of particular items in calculating Q_3 statistics but also LD).

After each replication, I store the mean, the standard deviation, and the *p*-values for the $LD\text{-}X^2$ and $LD\text{-}G^2$ statistic. For our proposed meta-analytic LD index, I collect the mean, the standard error, the *QM* value, and the *p*-value of each moderator. Convergence was an issue in estimating the 3PL model due to the difficulty in estimating the guessing parameter (Table 2.1; Table 2.2). Convergence rates varied round .01 for the models with strong model misspecification to .25 for models with no misspecification. Appendix 2A has an overview of these simulation studies, including the studies' goals and a summary of some important technical aspects.

Table 2.1

Proportion of Convergence for the 3PL Model

	N	Items			
		15	**25**	**30**	**50**
LD Clusters					
Null					
	1,000	0.193	0.211	0.225	0.248
	4,000	0.202	0.115	0.077	0.015
Mild					
	1,000	0.120	0.182	0.178	0.179
	4,000	0.132	0.041	0.037	0.011
Strong					
	1,000	0.100	0.109	0.118	0.097
	4,000	0.074	0.034	0.014	0.008
LD Pairs					
Null					
	1,000	0.201	0.212	0.235	0.233
	4,000	0.215	0.110	0.088	0.017
Mild					
	1,000	0.146	0.175	0.191	0.244
	4,000	0.150	0.077	0.058	0.110
Strong					
	1,000	0.099	0.170	0.191	0.190
	4,000	0.110	0.063	0.041	0.020

Simple vs. Complex Moderator Structure

Due to the inherent dependence in generating Q_3 values (e.g., the $Q_{31,2}$ value depends on both item 1 and item 2, regardless of any local dependence), we must consider the moderator structure when applying the model-building approach. We can apply either a "simple" moderator structure or a "complex" moderator structure. With a simple structure, we are only interested in modeling the LD aspect of the assessment; if we suspect items 1, 2, and 3 are locally dependent, we can code all Q_3 values sharing these items as 1. On a five-item test, this moderator structure

looks like Table 2.2. With a complex moderator structure, in addition to the moderator capturing suspected LD, we also model the dependencies among all items (e.g., all Q_3 values that share item 1, all Q_3 values that share item 2…, etc.). For a five-item test, a complex moderator structure matches Table 2.3.

Table 2.2
Simple Structure

Q_3 Value	LD
$Q_{3(1,2)}$	1
$Q_{3(1,3)}$	1
$Q_{3(1,4)}$	0
$Q_{3(1,5)}$	0
$Q_{3(2,3)}$	1
$Q_{3(2,4)}$	0
$Q_{3(2,5)}$	0
$Q_{3(3,4)}$	0
$Q_{3(3,5)}$	0
$Q_{3(4,5)}$	0

Note. LD is the moderator that indicates our suspected local dependence between item 1 and 2.

Table 2.3
Complex Moderator Structure

Q_3 Value	LD	isi1	isi2	isi3	isi4
$Q_{3(1,2)}$	1	1	1	0	0
$Q_{3(1,3)}$	1	1	0	1	0
$Q_{3(1,4)}$	0	1	0	0	1
$Q_{3(1,5)}$	0	1	0	0	0
$Q_{3(2,3)}$	1	0	1	1	0
$Q_{3(2,4)}$	0	0	1	0	1
$Q_{3(2,5)}$	0	0	1	0	0
$Q_{3(3,4)}$	0	0	0	1	1
$Q_{3(3,5)}$	0	0	0	1	0
$Q_{3(4,5)}$	0	0	0	0	1

Note. isi1 indicates the Q_3 values for the relationships between all items and item 1. *isi2* indicates the Q_3 values for the relationships between all items and item 2, and so on until *isi4*, which is one minus the total number of items. LD represents the moderator that indicates our suspected local dependence between item 1 and item 2.

Results

QM Null Condition

Under the null condition (ρ = 1.00), the RE model with a simple moderator structure produces results that are closest to the nominal alpha level (of 0.05). Figure 2.1 shows a comparison between the RE and FE models for item pairs and item clusters, using both a simple and a complex moderator structure. Overall, the RE models result in a Type I error rate that is closest to nominal. The number of items featuring LD also influences the Type I error rate, such that item clusters have inflated Type I error rates relative to item pairs. Despite this inflation, the Type I error rate for the RE model with item clusters is closer to nominal than that for the corresponding FE model. The model that performs worst is the FE model with a simple structure

applied to item clusters; its errors range between 0.10 and 0.18. This is because, as described previously, the Q_3 values generated have inherent heterogeneity that should be captured using the complex moderator structure (0.10 - 0.17). Even using the complex structure, though, the Type I error rate for the FE cluster model is liberal, ranging from 0.09 to 0.10.

In the null condition with a sample size of $N = 4{,}000$, the differences between the models are clarified. The Type I error for the FE cluster model with a simple structure is inflated as high as 0.37 if the test length is short (15 items) and gradually decreases as the number of items increases, although still at 0.20 (much higher than nominal). Such a false positive rate is unacceptable for any statistical test. The error rate decreases if the FE model includes a complex structure, but remains too large. The FE models with item pair moderators, regardless of structure, also exhibit inflated error rates. As for the item cluster analyses, including a complex structure does decrease the error rate from 0.10 - 0.25 to 0.09 - 0.20. This reduction is a step in the right direction, but the result is still much larger than the target of .05.

The RE models achieve results that are closest to a nominal false rejection rate, although the RE model for clusters with a simple structure has a higher error rate (0.10) than its counterpart for single item pairs (0.05). Of the two models, using a complex structure yields a higher false positive rate than using a simple moderator structure.

In Figures 2.3 and 2.4, I compare the Type I error rate of the most successful condition (RE for an item pair with a simple structure) against that of the more established $LD\text{-}X^2$ and $LD\text{-}G^2$ statistics. The $LD\text{-}X^2$ and $LD\text{-}G^2$ statistics have a drastically conservative Type I error rate, close to 0.01, while the *QM* error rate hovers at a steady 0.05. These results remain consistent across levels of both test length and sample size.

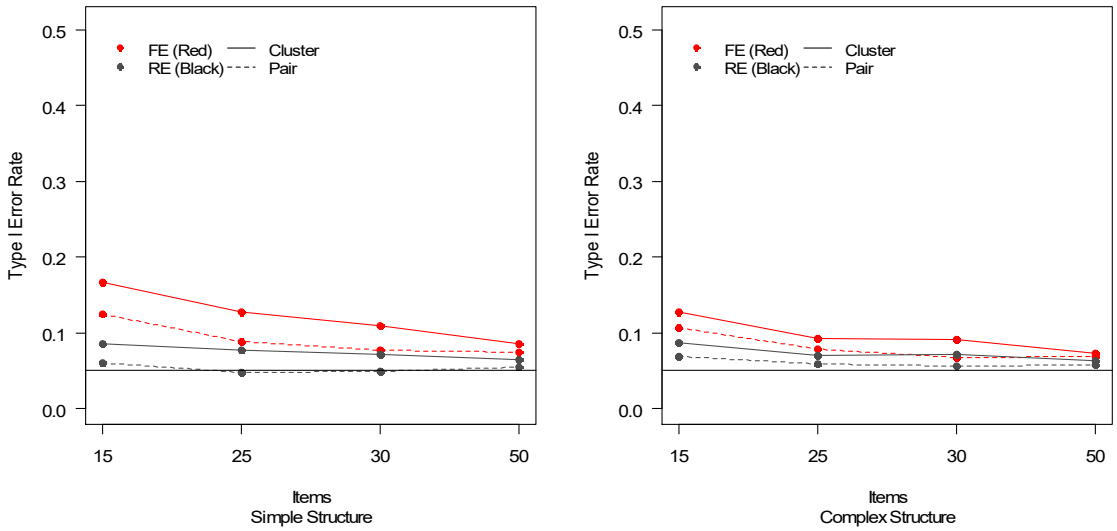


Figure 2.1. Type I error rate for LD detection under the null condition for clusters of items and item pairs; $N = 1{,}000$.

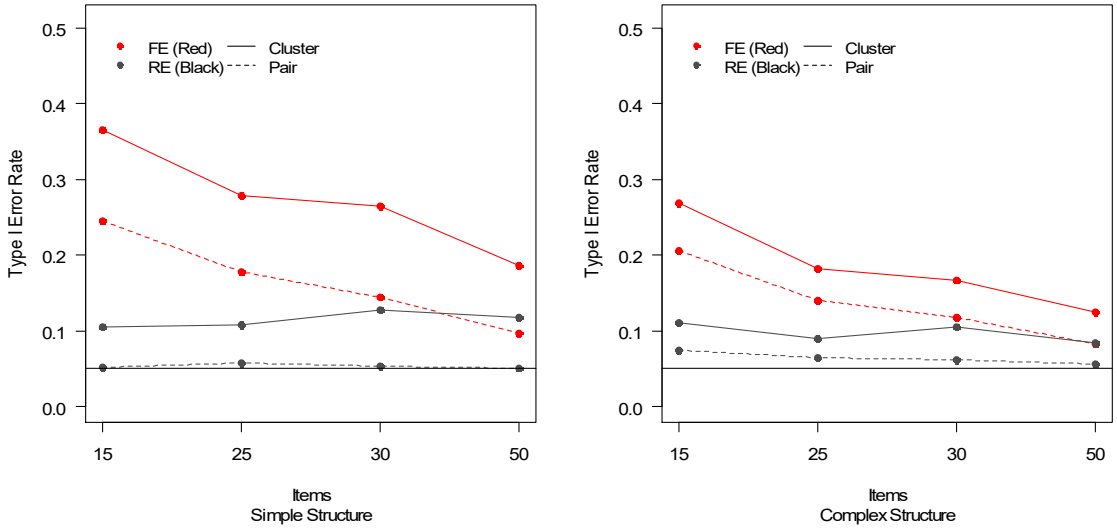


Figure 2.2. Type I error rate for LD detection under the null condition for clusters of items and item pairs; $N = 4{,}000$.

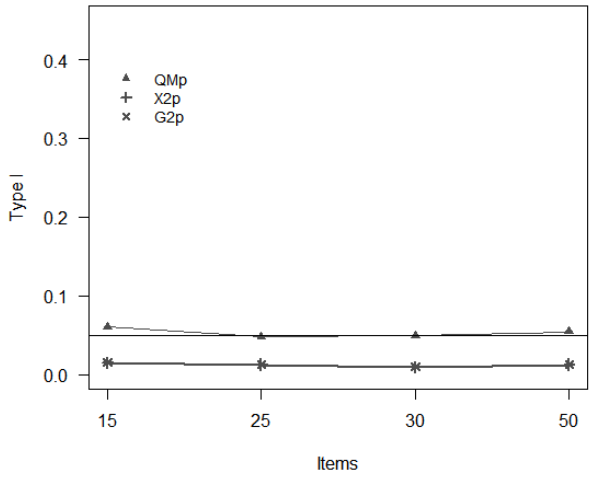


Figure 2.3. Type I error rate comparison between *QM*, X^2, and G^2 for the random-effects simple structure LD pairs; null condition; $N = 1{,}000$.

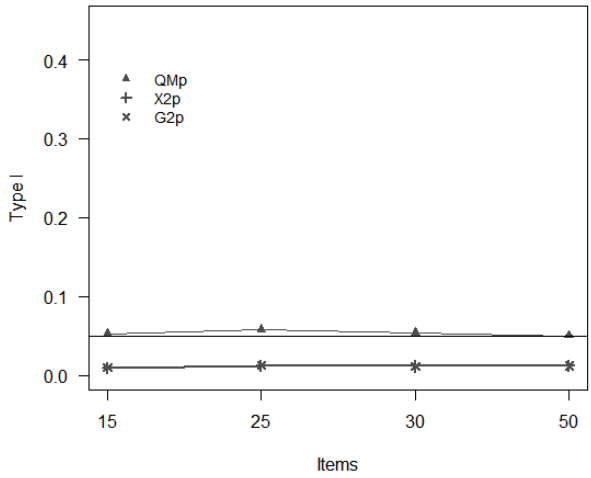


Figure 2.4. Type I error rate comparison between *QM*, *LD-*X^2, and *LD-*G^2 for the random-effects simple structure LD pairs; null condition; $N = 4{,}000$. "*" is a product of the overlap between the "+" and "x" symbols used to indicate *LD-*X^2 and *LD-*G^2, respectively.

Mild LD Condition

The FE cluster models, regardless of moderator structure, have the most power to detect mild local dependence (ρ = 0.50), ranging from 0.85 to 0.95. The RE simple structure model for item clusters comes in second, with a detection rate of 0.90 across the number of items for the *N* = 1,000 condition (Figure 2.5). Of the remaining models, the FE model with a simple structure for item pairs has a higher detection rate, ranging between 0.70 and 0.80. Its RE counterpart has the lowest power to detect mild local dependence, hovering at 0.60 regardless of number of items.

For the conditions with a large sample size (N = 4,000), a similar pattern emerges (Figure 2.6). The FE cluster models have higher power than the RE item pair models. However, given that these are the same models that have a much higher Type I error rate, these results are expected and are not particularly relevant.

Figures 2.7 and 2.8 compare the LD detection rates for the model with Type I error rates closest to nominal levels (the RE model for item pairs with a simple structure) to the detection rates for the $LD\text{-}X^2$ and $LD\text{-}G^2$ statistics. With smaller sample sizes, the model-building approach has a consistently higher detection rate (0.60); as sample size increases, so does detection rate. For the model-building approach, increasing the test length has little effect on detection rate.

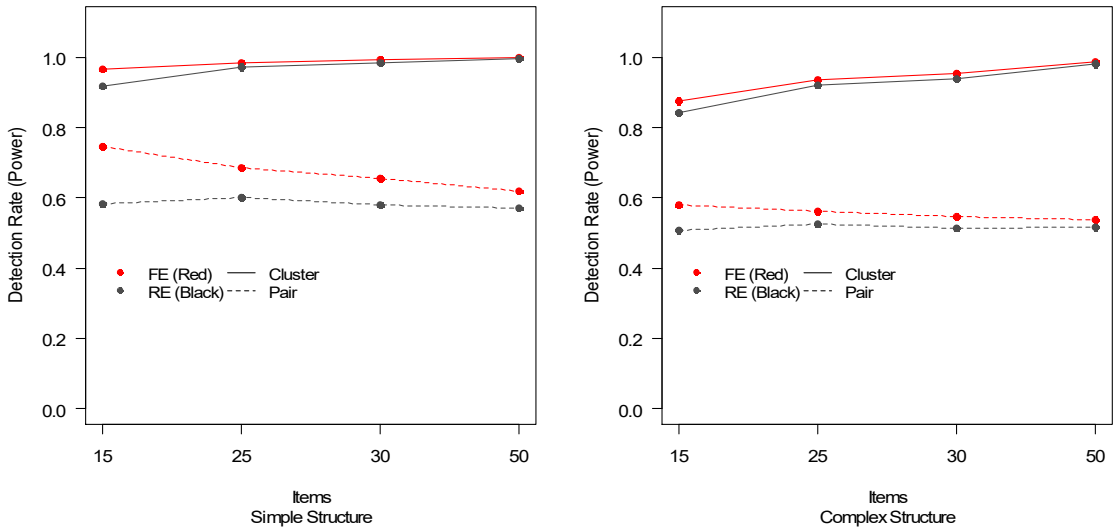


Figure 2.5. Power of LD detection under the mild condition for clusters of items and single pairs; $N = 1{,}000$.

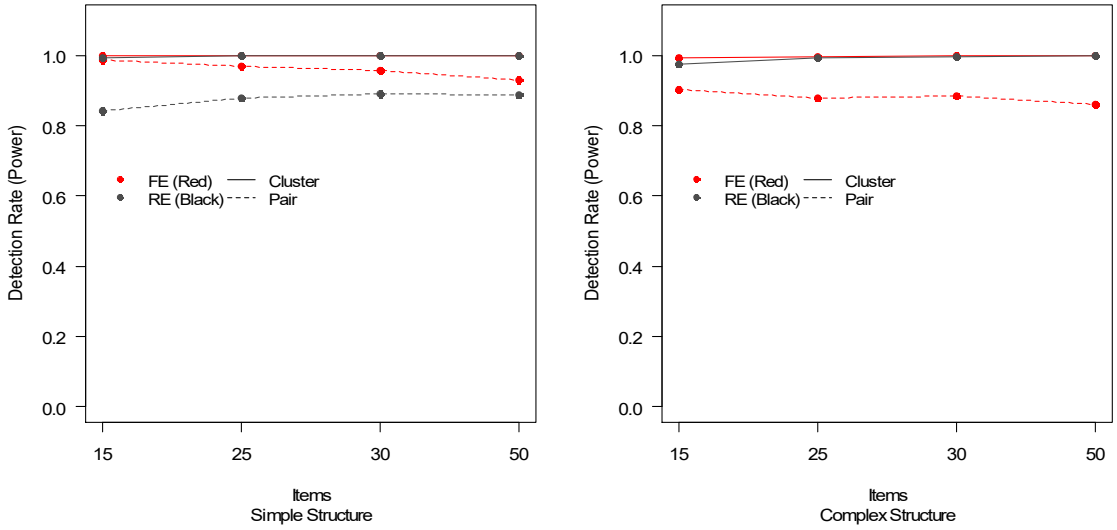


Figure 2.6. Power of LD detection rate under the mild condition for clusters of items and single pairs; $N = 4{,}000$.

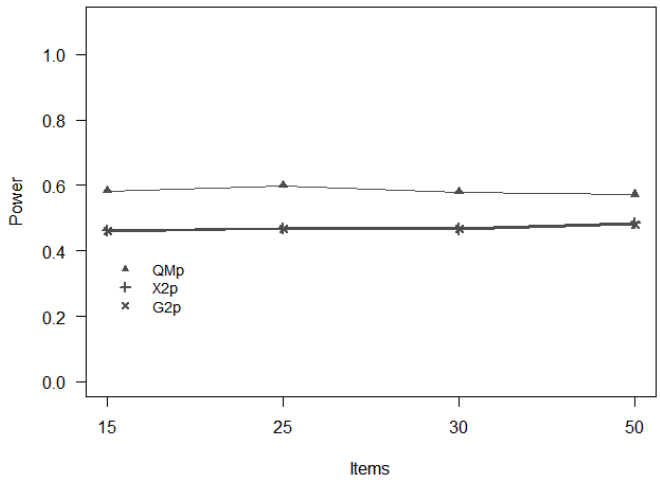


Figure 2.7. Power comparison between *QM*, X^2, and G^2 for the random-effects simple structure LD pair; mild condition; $N = 1{,}000$.

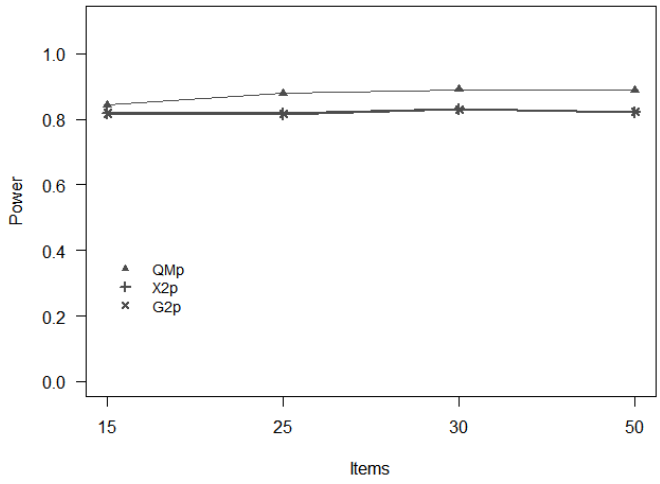


Figure 2.8. Power comparison between *QM*, X^2, and G^2 for the random-effects simple structure LD pair; mild condition; $N = 4{,}000$.

Strong LD Condition

Like the mild condition, in the strong condition (ρ = .2) the complex moderator structure continues to work better than the simple structure (Figure 2.9). Among the simple structure models, the RE cluster model outperforms all others, including the single pair and FE cluster models. The RE cluster model always detects LD (1.00), while the FE pair approach detects LD between 0.65 and 0.85 of the time, depending on test length. Interestingly, for the FE pair simple structure model, power to detect LD decreases as test length increases. That pattern is not evident for its RE counterpart, perhaps due to the way that the model accounts for excess heterogeneity. When N = 4,000 (Figure 2.10), this reduction in power no longer appears. Additionally, almost all models detect LD in every instance (1.00), except for the RE simple structure pair model, which hovers around 0.85 to 0.90.

Figures 2.11 and 2.12 compare the RE pair simple structure model to the X^2 and G^2 statistics. Again, the RE method has a higher detection rate, consistent at 0.70, than either X^2 or G^2 (at 0.60).

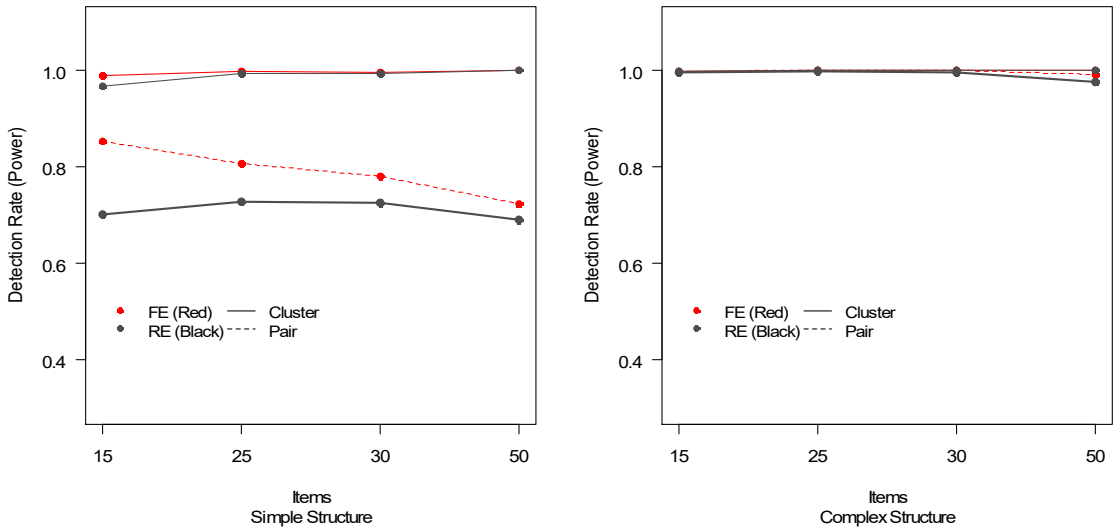


Figure 2.9. Power of LD detection rate under the strong condition for clusters of items and single pairs; $N = 1{,}000$.

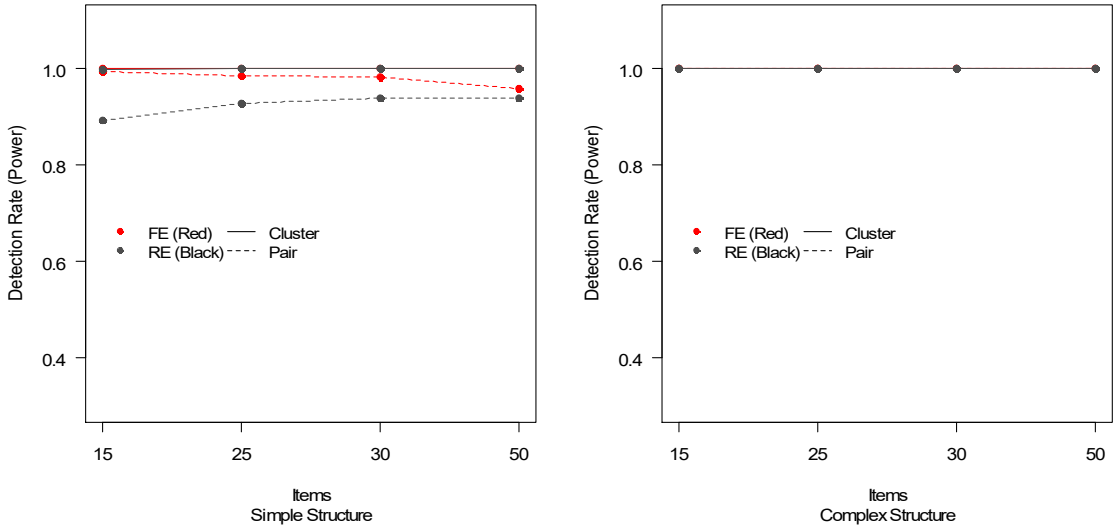


Figure 2.10. Power of LD detection rate under the strong condition for clusters of items and single pairs; $N = 4{,}000$.

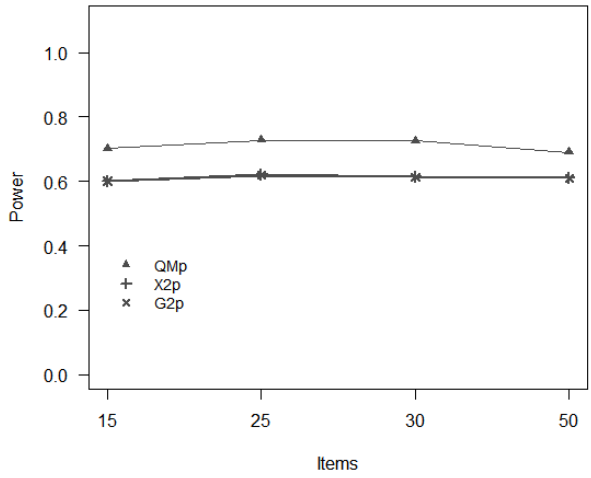


Figure 2.11. Power comparison between *QM*, X^2, and G^2 for the random-effects simple structure LD pair; strong condition; $N = 1{,}000$.

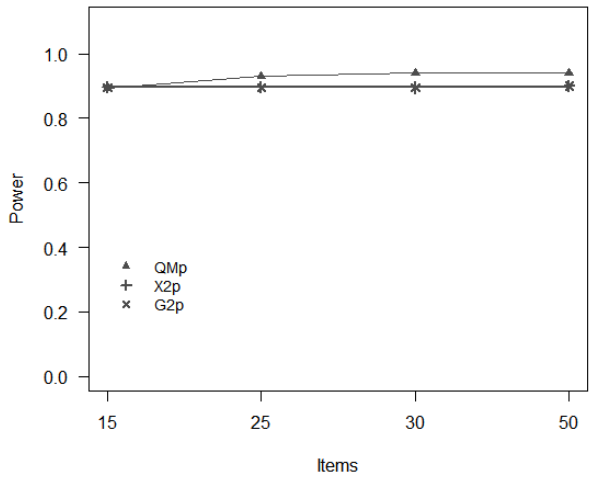


Figure 2.12. Power comparison between *QM*, X^2, and G^2 for the random-effects simple structure LD pair; strong condition; $N = 4{,}000$.

Discussion

As described above, the *QM* statistic for the RE single pair and item cluster models with a simple moderator structure has desirable properties under these conditions. These results hold across levels of sample size, number of items, and degree of LD. Under the null condition, these models yield better Type I error rates than the commonly used $LD\text{-}X^2$ and $LD\text{-}G^2$ statistics, and they possess equal or greater power to detect local dependence, regardless of its magnitude.

The RE models show more promise than the FE models as statistical tools for applied researchers. The RE models minimize the chance of falsely identifying LD when none exists, while maintaining reasonable power when LD is present. To choose between a simple structure or a complex structure, the results suggest that an RE model with a simple moderator structure maintains a nominal alpha rate and maintains an adequate amount of power to detect both mild and strong levels of LD. Using a simple structure also minimizes the complexity of the actual analysis.

In terms of power, I consider only the models that have either nominal or close to nominal false rejection rates (in other words, the RE simple structure models). The RE simple structure model can detect LD from item clusters more often than for single item pairs. When testing item pairs, I compared the model-building approach to the $LD\text{-}X^2$ and $LD\text{-}G^2$ statistics. The model-building approach (or the RE simple structure model) outperforms both the $LD\text{-}X^2$ and $LD\text{-}G^2$, not only in achieving a nominal Type I error rate but also in having increased power. Note, however, that although the model-building approach outperforms the established statistics, the model-building approach by nature is confirmatory, as opposed to the traditional exploratory approaches of identifying LD.

The model-building approach works best for the RE model with a simple moderator structure. Under null conditions, its detection rates are very close to alpha, and under non-null conditions it has strong power to detect local dependence. Although the simple structure RE model is useful with single item pairs, the true advantage of the model-building approach is in modeling item clusters. This allows researchers to identify test characteristics that impact multiple items and to reduce the opportunity for an inflated Type I error rate that would otherwise emerge from the pairwise comparisons traditional approaches employ.

CHAPTER 3

Polytomous Underlying Local Dependence Using the Model-Building Approach

Purpose

This study aims to investigate the performance of the model-building approach for tests that have either all binary items or all polytomous items. This chapter extends the research in Chapter 2 by assessing the performance of the model-building approach in cases of polytomous response type data. Additionally, this project uses only the random-effects meta-analytic model, as this model has demonstrated better statistical properties than the fixed-effect model regardless of moderator structure.

Multiple-choice items are a popular assessment method in educational testing. These items are typically scored using a single correct option and reduced to a binary dataset. The binary data can then be modeled using 1-, 2-, or 3-parameter logistic (1PL, 2PL, 3PL) item response theory models (IRT) models. In psychology, the Likert-type scale is a more popular format in which participants select a response from an ordered categorical set of options. This type of data is modeled using an extension of the 2PL model that Samejima (1969) developed and popularized, the graded response model (GRM) is defined as

$$P\left(Y_{ijk}\middle|\theta_i\right) = \frac{1}{1+\exp(-(a_j\theta_i - c_{jk})} - \frac{1}{1+\exp(-(a_j\theta_i - c_{jk+1})},$$

where the probability of response Y for person i on item j in category k is the difference between the probability of a response for category k and the response for the next higher category, k+1. As

in the 2PL model, the discrimination parameter, *a,* determines the degree to which an item is related to the latent variable θ. The *c* parameter is the intercept for item *j* and category *k*. Because it is an extension of the 2PL model, this model requires that the same assumptions be met – those of unidimensionality, monotonicity, and local item independence (LII) (Samejima, 1969). This chapter focuses on detecting violations of LII. Repetitive items, commonly known as item doubles, are a common example of such a violation (Steinberg & Thissen, 1995). For example, content experts might know that two items with similar wording like "I feel sad" and "I feel unhappy" distinguish between two different constructs, but these might be indistinguishable to test takers. Ignoring violations of LD impacts both model fit and parameter estimation (Chen & Thissen, 1997; Liu & Maydeu-Olivares, 2012; Yen, 1984; Zenisky et al., 2001). Although there is no particular reason that the Q_3 values should differ between the 2PL model and the GRM, confirming this through simulation is important.

Data Simulation

This study investigates the performance of the model-building approach for detecting underlying local dependence (ULD) for item tests that are either all binary or all polytomous. ULD is generated by sampling from two correlated latent traits, θ_1 and θ_2, with several different levels of correlation strength (i.e., 0.20, 0.50). The simulation investigates the effects of varying four factors: item type (binary or polytomous), number of items, number of observations, and LD (correlation) strength.

For the number of binary items, the levels are 25, 30, and 50. I obtained these values by evaluating the number of items used in other IRT LD simulation studies (Houts & Edwards, 2013; Liu & Thissen, 2012, 2014). Although the referenced simulation studies included a 10-item condition, I dropped this cell due to time restrictions. For the number of observations, the levels

are 1,000 and 4,000 (Christensen et al., 2016; Glas & Falcon, 2003; Liu & Thissen, 2012, 2014). Sample sizes in the cited studies ranged between 200 and 4,000; I restricted the levels to 1,000 and 4,000 due to time constraints, but future studies will include a larger range. The slope parameters for the binary items are sampled from a normal distribution with a mean of 1.70 and standard deviation of 0.30. The ULD structure is relatively simple; the first and second items exhibit LD in the non-null conditions. I generated two correlated latent traits (e.g., math and science). A correlation of 1.00 reflects the null condition, where there is only a single latent trait (the latent traits are perfectly correlated). When LD is present, correlations of 0.50 and 0.20 represent the mild and strong conditions, respectively. Finally, I modeled responses in the all binary condition using a two-parameter logistic (2PL) model.

For the polytomous item conditions, the slope parameters are sampled from a normal distribution with a mean of 1.70 and a standard deviation of 0.30. The first item threshold is drawn from a normal distribution with a mean of -1.50 and a standard deviation of 0.50. The second threshold was computed as the sum of the first threshold and a second value sampled from a normal distribution with a mean of 1.00 and a standard deviation of 2.00. Again, I designated the first and second items to tap two correlated latent traits (e.g., math and science) in the non-null conditions, which created underlying local dependence. A correlation of 1.00 reflects the null condition, where there is only a single latent trait. When LD exists, correlations of 0.50 and 0.20 represent the mild and strong conditions. Finally, I modeled the polytomous items with a graded response (GRM) model.

During each replication, I record the mean, the standard deviation, and the *p*-values for all of the specified LD indices. For the proposed meta-analytic LD index, I collect the *QM* value, the *p*-value of each moderator, and the estimates of the regression coefficients and τ^2.

Analysis

To assess Type I error rates, I calculate the number of replications that detect local dependence when none exists (in the null conditions) and divide by the total number of replications. This produces a value ranging from 0 - 100 indicating the percentage of times that local dependence is detected. This same process is applied to the non-null conditions, in which cases it represents power rather than Type I error rate. Additionally, I compute the mean value for each parameter across each cell and present them in table format (Table 3.1).

I assess simulation convergence by plotting a running mean, where the number of replications is plotted against the current average of each parameter value (β_0, MQ, etc.). I consider the model to be fully converged when the running mean remains steady at any particular value over a period of multiple replications.

Results

Figures 3.1 and 3.2 summarize the results. Table 4.1 presents data for the all binary conditions, and Table 4.2 presents the all polytomous condition. In the all binary null condition, the model erroneously detects an effect nearly 5% of the time for both sample sizes investigated (N = 1,000 and N = 4,000). This nominal detection rate was consistent across test lengths (P = 25, 30, 50). Figure 3.2 plots the average estimated effect size, which was -0.02, consistent with previous research. For items with suspected LD (under the null condition), the change in effect size was negligible (close to zero). When LD existed, power to detect an effect was near 100%. This detection rate was higher than expected, possibly due to the conditions chosen.

The results for the polytomous condition modeled with the GRM are in Figure 3.2 and Figure 3.3. These results mirrored the findings of the binary condition, indicating that the model-

building approach performs similarly with both the 2PL model and the GRM. With no local dependence present and an arbitrary moderator, the false positive rate was nearly 5%. This was consistent across sample size and test length. Closer inspection (Figure 3.3) indicates that the average effect size for items not suspect of LD (in both the LD and null conditions) was consistently close to -0.03. With no LD present (the null condition), the average effect size remained near 0 (no change), although the average change (β_1) was close to 0.34 in the mild condition and to 0.40 in the strong condition.

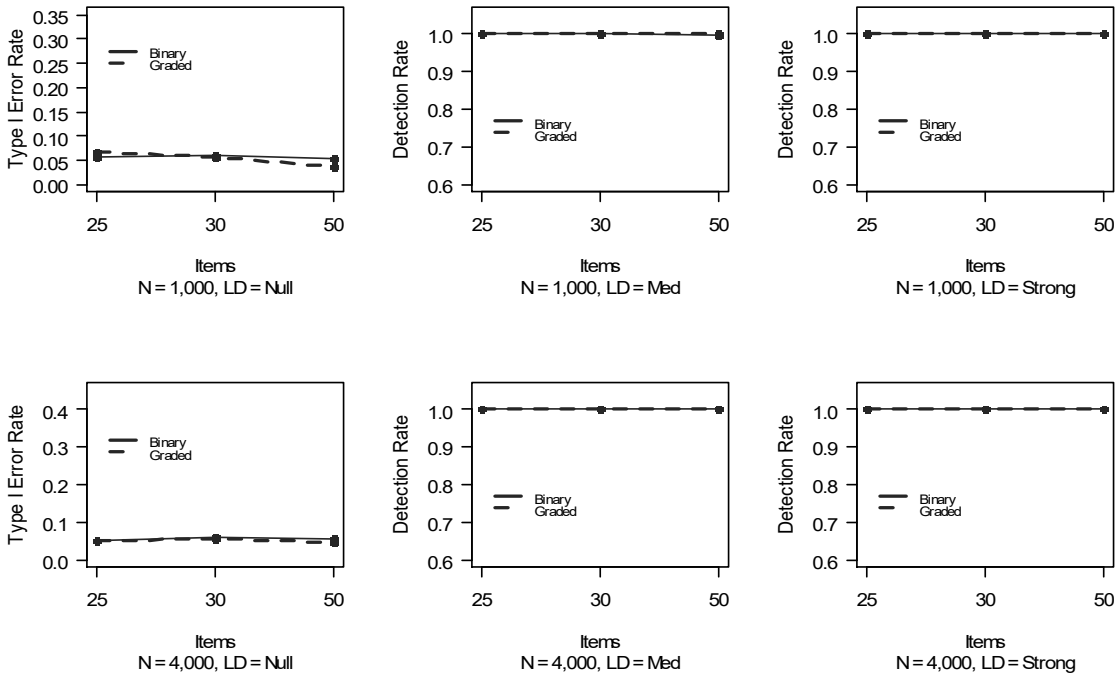


Figure 3.1. Type I error rate and power for all binary items and all polytomous items.

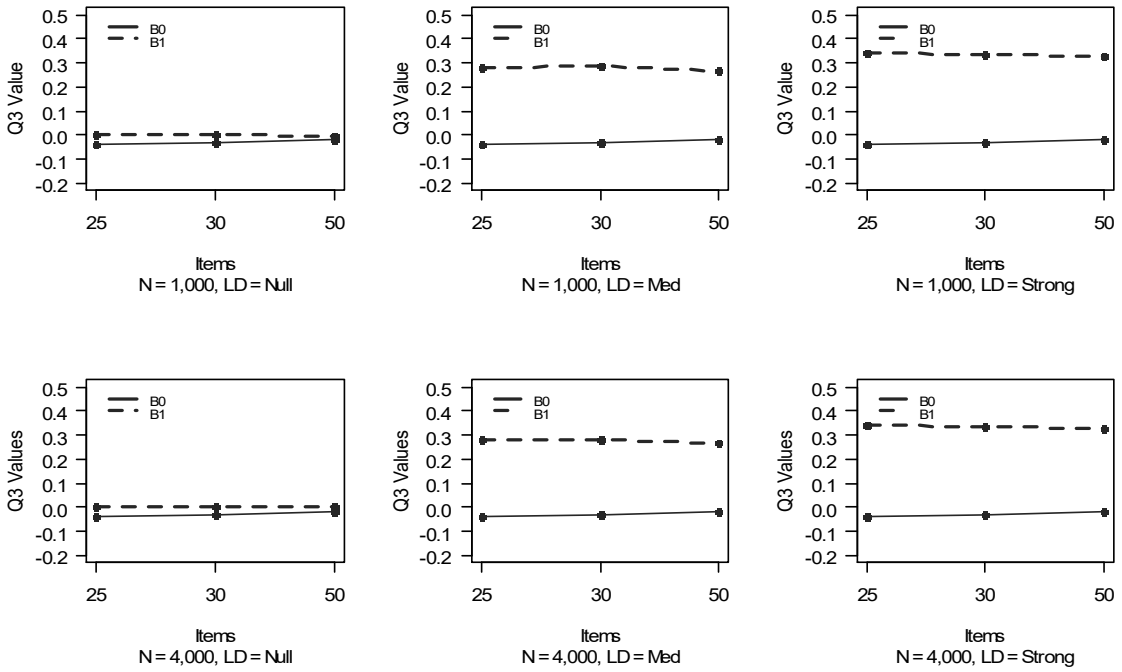


Figure 3.2. β_0 and β_1 values for all binary items.

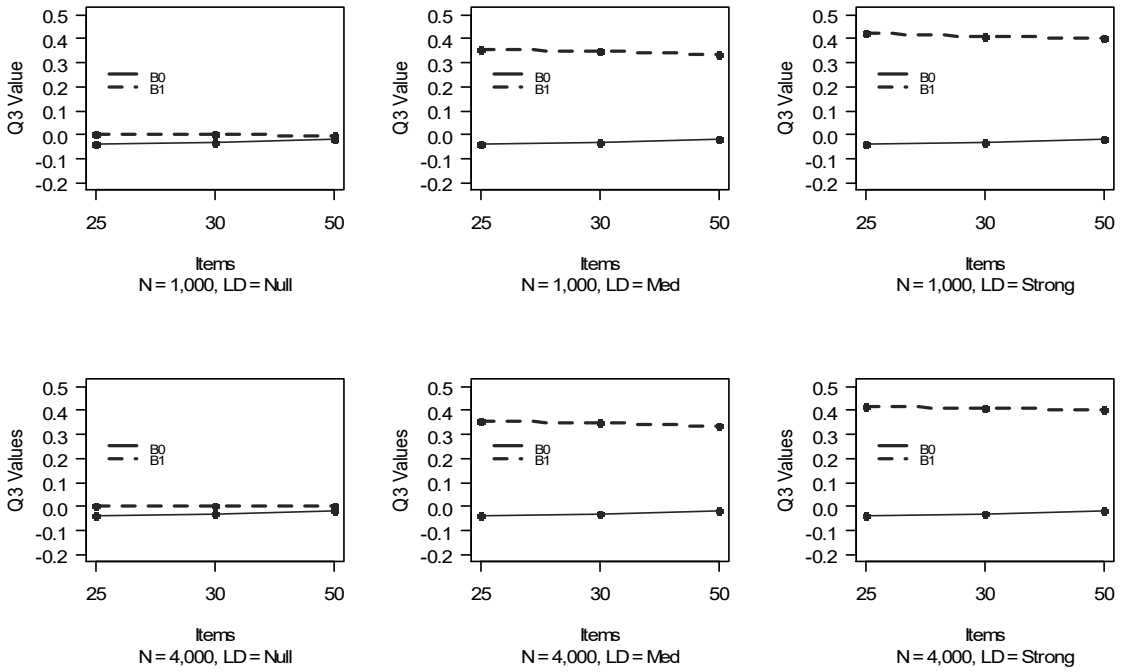


Figure 3.3. β_0 and β_1 values for all polytomous items.

Discussion

These results indicate that the model-building approach is appropriate for both binary and polytomous data. When there is no sign of local dependence, the model rarely identifies an effect, suggesting that users can employ this approach and maintain confidence in the results. The model also consistently detects an effect when one exists, which also supports the approach. The model not only provides a statistical test to detect LD but also produces an estimate of the effect (the Q_3 values) for both the average non-LD item and the average LD item (those coded as locally dependent). These results are consistent across sample sizes and test length.

The parameters in this simulation study were carefully selected based on prior applied research and simulation studies. However, due to time constraints, only a few levels are included

for each factor. In future studies, I intend to include a larger number of sample size levels (e.g., 150, 200, 500) and a larger number of test lengths (e.g., 10, 15, 20, 25) to assess the performance of the approach in more extreme conditions. Furthermore, I plan to include a wider range of correlations (degrees of LD), as the ones selected for this condition are optimal. Suboptimal conditions must also be examined.

Table 3.1

All Binary Items

Cell	Items	N	ρ/θ	β_0	B_0P%	β_1	β_1 P%	*QE*	QEP%	*QM*	*QM*P%	*Tau2*
1	25	1,000	1	-0.0391	1.0000	-0.0005	0.0563	370.4162	0.8412	1.1107	0.0563	0.0002
2	30	1,000	1	-0.0325	1.0000	-0.0007	0.0601	522.2249	0.8659	1.0273	0.0601	0.0002
3	50	1,000	1	-0.0195	1.0000	-0.0018	0.0520	1406.3022	0.9430	1.0255	0.0520	0.0001
4	25	4,000	1	-0.0391	1.0000	0.0007	0.0510	468.9319	0.9980	1.0515	0.0510	0.0001
5	30	4,000	1	-0.0326	1.0000	-0.0001	0.0590	620.2724	0.9990	1.1126	0.0590	0.0001
6	50	4,000	1	-0.0195	1.0000	0.0001	0.0570	1497.2380	0.9990	1.0391	0.0570	0.0001
7	25	1,000	0.2	-0.0372	1.0000	0.3420	1.0000	408.3304	0.9910	89.7756	1.0000	0.0004
8	30	1,000	0.2	-0.0312	1.0000	0.3345	1.0000	550.8696	0.9790	92.0155	1.0000	0.0003
9	50	1,000	0.2	-0.0190	1.0000	0.3276	1.0000	1416.8023	0.9760	96.8161	1.0000	0.0002
10	25	4,000	0.2	-0.0372	1.0000	0.3423	1.0000	636.1061	1.0000	230.2781	1.0000	0.0003
11	30	4,000	0.2	-0.0313	1.0000	0.3354	1.0000	758.8779	1.0000	268.4951	1.0000	0.0002
12	50	4,000	0.2	-0.0191	1.0000	0.3246	1.0000	1580.4896	1.0000	339.0281	1.0000	0.0001
13	25	1,000	0.5	-0.0386	1.0000	0.2808	0.9990	380.0598	0.8960	65.6250	0.9990	0.0003
14	30	1,000	0.5	-0.0322	1.0000	0.2840	1.0000	527.3136	0.8920	69.9662	1.0000	0.0002
15	50	1,000	0.5	-0.0194	1.0000	0.2693	0.9970	1402.3234	0.9570	67.0042	0.9970	0.0001
16	25	4,000	0.5	-0.0386	1.0000	0.2812	1.0000	523.3487	1.0000	192.4553	1.0000	0.0002
17	30	4,000	0.5	-0.0322	1.0000	0.2799	1.0000	662.5934	1.0000	216.6617	1.0000	0.0001
18	50	4,000	0.5	-0.0194	1.0000	0.2675	1.0000	1522.2691	1.0000	241.1971	1.0000	0.0001

Note. ρ/θ is the correlation strength between latent traits used to generate LD. Headings with P% indicate the proportion $\alpha<.05$.

Table 3.2
All Polytomous Items

Cell	Items	N	ρ/θ	β_0	B_0P%	β_1	β_1 P%	*QE*	QEP%	*QM*	*QM*P%	*Tau2*
19	25	1,000	1	-0.0382	1.0000	0.0002	0.0659	366.8546	0.8136	1.1261	0.0659	0.0002
20	30	1,000	1	-0.0317	1.0000	0.0006	0.0553	522.8948	0.8592	1.0279	0.0553	0.0002
21	50	1,000	1	-0.0189	1.0000	-0.0018	0.0350	1443.0358	0.9910	0.9303	0.0350	0.0002
22	25	4,000	1	-0.0382	1.0000	0.0011	0.0530	416.1634	0.9940	1.0114	0.0530	0.0001
23	30	4,000	1	-0.0317	1.0000	0.0006	0.0550	570.0351	0.9910	1.0457	0.0550	0.0001
24	50	4,000	1	-0.0189	1.0000	0.0000	0.0480	1487.1330	1.0000	0.9742	0.0480	0.0001
25	25	1,000	0.2	-0.0364	1.0000	0.4205	1.0000	399.4704	0.9730	135.7060	1.0000	0.0003
26	30	1,000	0.2	-0.0305	1.0000	0.4118	1.0000	548.5579	0.9680	137.4629	1.0000	0.0003
27	50	1,000	0.2	-0.0185	1.0000	0.4000	1.0000	1447.1109	0.9900	138.3292	1.0000	0.0002
28	25	4,000	0.2	-0.0364	1.0000	0.4173	1.0000	573.5830	1.0000	373.1136	1.0000	0.0002
29	30	4,000	0.2	-0.0305	1.0000	0.4108	1.0000	702.5648	1.0000	427.3960	1.0000	0.0002
30	50	4,000	0.2	-0.0185	1.0000	0.4011	1.0000	1557.7330	1.0000	516.3887	1.0000	0.0001
31	25	1,000	0.5	-0.0378	1.0000	0.3531	1.0000	374.0991	0.8760	102.6790	1.0000	0.0002
32	30	1,000	0.5	-0.0315	1.0000	0.3465	1.0000	529.4983	0.9000	101.3437	1.0000	0.0002
33	50	1,000	0.5	-0.0188	1.0000	0.3365	1.0000	1439.0527	0.9870	99.0668	1.0000	0.0002
34	25	4,000	0.5	-0.0379	1.0000	0.3525	1.0000	461.6585	1.0000	333.3387	1.0000	0.0001
35	30	4,000	0.5	-0.0315	1.0000	0.3472	1.0000	612.4929	1.0000	350.5030	1.0000	0.0001
36	50	4,000	0.5	-0.0188	1.0000	0.3375	1.0000	1505.9416	1.0000	379.0398	1.0000	0.0001

Note. ρ/θ is the correlation strength between latent traits used to generate LD. $\alpha<.05$.

CHAPTER 4

Using the Model-Building Approach to Detect Violations of Local Item Independence in Tests with Mixed Items

Purpose

This study has two goals. First, I aim to demonstrate that mixing item types can be a cause for concern. Researchers who construct tests should be aware of the pitfalls that may arise from item mixing and should use models capable of accounting for violations of local item independence (LII), or multidimensionality. Second, I want to evaluate the usefulness of the model-building approach in detecting such violations.

In educational and psychological testing, item mixing involves combining binary and polytomous items on the same exam, a practice referred to as mixed-format tests. A survey by Lane (2005) found that 63% of state assessments use both multiple-choice items and constructed-response items. If we assume that all these items are loading onto the same underlying construct, their dimensionality should be of little concern. However, item mixing can violate unidimensionality, which can cause problems when estimating item parameters and generating test scores (Thissen, Steinberg, & Mooney, 1989).

Researchers, test developers, and educators often implement mixed-format tests for various reasons. Researchers may find that some items are more easy to implement in a multiple choice (MC) format (e.g., binary questions), while other items may attempt to understand students' process of answering questions (e.g., constructed response items) (Livingston, 2009; Wainer & Thissen, 2009). Other items may use polytomous responses to deal with issues of independence; one such example involves collapsing multiple questions following a reading

passage into a single testlet (Thissen et al., 1989). Regardless of the reason, researchers should understand the consequences of their actions, and should be aware of methods for mitigating the resulting impact on item response theory (IRT) parameter estimation. This study explores the impact of mixing items on local item independence (LII) as assessed by the model-building approach.

Constructed Response Items

The multiple choice (MC) item is the most popular choice in test design. Typically, these are items for which examinees must select the correct response from a list of choices; their answer is then dichotomized as either correct or incorrect (Thissen & Steinberg, 1984). Constructed response (CR) items, on the other hand, are items with open-ended questions that are scored on the response's degree of correctness (Livingston, 2009). Although they are time-consuming and expensive (often multiple human raters must inspect each student response to each item), CR items have gained popularity in educational testing; they were adopted by the Educational Testing Service and the National Assessment of Educational Progress (McClellan, 2010; NCES, 2008; Thissen, Wainer, & Wang, 1994). CR items are more feasible because some testing companies implement automatic scoring (e.g., ETS).

Dependence and dimensionality are topics of interest that arise from combining item types. That is, does combining different item types result in multidimensionality, or violations of LII? Studies on this subject have reached mixed conclusions. Ercikan et al. (1998) attempted to answer this question by combining and calibrating different ratios of CR to MC items on a set of tests (involving reading, language, mathematics and science). They concluded that mixing MC and CR items did not lead to local item dependence, because the correlated item residuals were small. However, an investigation by Thissen, Wainer, and Wang (1994) found that Advanced

Placement Computer Science items loaded onto a general factor while the CR items loaded onto a specific factor, raising questions about dimensionality.

Consequences of Violating LII

The assumption of independence is necessary because IRT models rely on maximum likelihood estimation. Maximum likelihood estimation of IRT models depends on calculating the joint probability of a set of items. After integrating out the ability parameter, the likelihood of the response function is the product of the individual item probabilities:

$$f(y_i) = \int \prod_{j=1}^{m} f_j(y_{ij}|\theta_i)\phi(\theta_i)d\theta_i.$$

In this equation, y_i is the likelihood of the response pattern for person i, m represents the total number of items in the dataset, θ represents each person's latent ability and $\phi(\cdot)$ is the density function of the standard normal distribution. After ability is accounted for, the individual item response probabilities should be independent of one another (Mcdonald, 1982).

The consequences of violating independence are well documented; such violations can produce biased parameter estimates (Chen & Thissen, 1997; Reese, 1995; Steinberg & Thissen, 1996). Reese found that under high LD conditions, histograms of the observed and expected score distributions were dramatically different and no longer normally distributed. Scores also spread out, which underestimated the scores of low ability students and overestimated the scores of high ability students. Yuan Li's dissertation found that it is more difficult to link multiple forms of tests with violations of LII, and Zenisky, Hambleton, and Sireci (2003) found that score distributions diverged from the truth by as much as one standard deviation. The goal of the present study is to simulate data using a single latent variable (i.e, a situation with no local dependence) and to examine the degree to which item mixing can produce local dependence.

Additionally, this simulation study will demonstrate the effectiveness of the model-building approach in identifying test characteristics (binary vs. polytomous items) that cause surface-level local dependence.

The Model-Building Approach to Detecting Local Dependence

This study uses the model-building approach to detect local dependence (LD). The moderator is a dummy variable that differentiates two groups of items, here defined by a characteristic of the item type. For example, if we have a five-item test that combines two constructed response items and three multiple choice items, the Q_3 matrix will look like Table 4.1. Each Q_3 entry represents the correlated residual of an item pair after accounting for the latent trait. In this measure, item one and item two are both constructed response questions; we assign the moderator value 1 to the Q_3 value for this item pair. Table 4.2 presents the final moderator structure for this mixed-format measure.

Table 4.1
Q_3 Matrix

		Constructed Response (CR)		Multiple Choice (MC)		
		Item 1	Item 2	Item 3	Item 4	Item 5
CR	Item 1	1	$Q_{3\,(1,2)}$	$Q_{3\,(1,3)}$	$Q_{3\,(1,4)}$	$Q_{3\,(1,5)}$
	Item 2	-	1	$Q_{3\,(2,3)}$	$Q_{3\,(2,4)}$	$Q_{3\,(2,5)}$
MC	Item 3	-	-	1	$Q_{3\,(3,4)}$	$Q_{3\,(3,5)}$
	Item 4	-	-	-	1	$Q_{3\,(4,5)}$
	Item 5	-	-	-	-	1

Table 4.2
Simple Structure

Q_3 Value	LD
$Q_{3\,(1,2)}$	1
$Q_{3\,(1,3)}$	0
$Q_{3\,(1,4)}$	0
$Q_{3\,(1,5)}$	0
$Q_{3\,(2,3)}$	0
$Q_{3\,(2,4)}$	0
$Q_{3\,(2,5)}$	0
$Q_{3\,(3,4)}$	0
$Q_{3\,(3,5)}$	0
$Q_{3\,(4,5)}$	0

Methods

Data Simulation

This simulation evaluates the performance of the model-building approach when identifying violations of local LII in mixed-format tests (e.g., a combination of binary and polytomous items). I generated data using the *sim()* command from the base R statistical software. I used the 2PL model in conjunction with the *mirt R* package to generate binary data and the graded response model (GRM) to generate polytomous data (Chalmers, 2012; Team, 2014). The item characteristic curve (ICC) for the 2PL model can be defined as

$$P(Y_{ij}|\theta_i) = \frac{1}{1 + \exp(-(a_j\theta_i - c_j))},$$

where a_j is the discrimination parameter and c_j is the intercept parameter.

Data for polytomous items were generated with a GRM, which is defined by the ICC

$$P(Y_{ijk}|\theta_i) = \frac{1}{1+\exp(-(a_j\theta_i - c_{jk})} - \frac{1}{1+\exp(-(a_j\theta_i - c_{jk+1})}.$$

In other words, the probability of a response *Y* for person *i* on item *j* in category *k* is the difference between the probability of a response for category *k* and the response for the next higher category, *k* + 1. As with the 2PL model, the *a* parameter is the discrimination parameter, which determines the strength of the relationship between the item and the latent variable *θ*. Finally, *c* is the intercept parameter for item *j* and category *k* (Chalmers, 2012; R Core Team, 2016).

The simulation varies three factors between cells: the number of binary items, the number of observations (test takers), and the item mixture type (all binary, all polytomous, or mixed). The levels for the number of binary items are 10, 30, and 50. These values came from examining the number of items used in other IRT LD simulation studies. They approximate some common test lengths in psychology, particularly for short forms of measures (Cooper & Petrides, 2010; Henry & Crawford, 2005; Le Moine et al., 2016; Milian et al., 2015; Thompson, 2007). The levels for the number of observations are 1,000 and 4,000. The typical values in the simulation literature are 200, 500, and 1,000 (Christensen et al., 2016; Glas & Falcon, 2003; Houts & Edwards, 2015; Liu & Thissen, 2014), but I include 4,000 because a larger participant pool better estimates the IRT parameters and represents scaling samples for commercial test development. For the binary data, the slope parameters are randomly generated from a normal distribution with a mean of 1.70 and standard deviation of 0.30. These values were based on Hill's (2004) distribution of 15 psychological scales and have been used in previous simulation studies (Houts & Edwards, 2013, 2015; Woods, 2009). The difficulty parameters are sampled from a normal distribution with a mean of 0 and a standard deviation of 1.50, and then are converted to intercept parameters with the same prior justification. For the polytomous data, the first threshold is sampled from a normal

distribution with a mean of -1.50 and a standard deviation of 0.50, with each subsequent threshold sampled from a normal distribution with a mean of 1 and standard deviation of 0.20. This results in a total of three thresholds and four categories, much like the parameters used by Liu and Thissen (2014).

I create some mixed tests by combining the polytomous and binary items into one test. For the 10-item condition, three items are generated using the GRM, and the remaining seven are 2PL items. For the 30-item test condition, nine items are generated via GRM and 21 by 2PL. For the 50-item condition, 15 are GRM and 35 are 2PL. This results in a general test makeup of 30% polytomous items. There is no clear guideline for the *proper* ratio of CR to MC items. Examining the literature resulted in mixture ranges between 8% and 30% (Ercikan et al., 1998; Lissitz, Hou, & Slater, 2012; Thissen et al., 1994). I simulated a 30% mixing condition to maximize detecting possible issues from LD. In future studies, I plan on examining a more diverse range of mixtures.

I evaluate each replication using the random-effects (RE) meta-analytic model. The RE model avoids the complication involved in modeling item dependencies that is inherent in the generation of Q_3 values. Additionally, the Q_3 statistic (which serves as our effect size) is more variable at smaller sample sizes, even with Fisher's *r*-to-*z* transformation. For this reason, using the random-effects meta-analytic model makes sense. Chapter 2 in this dissertation demonstrates that the RE model has greater power and a lower Type I error rate than the FE model when identifying LD, which further justifies using only the RE model. Each cell includes 6,000 replications.

Regarding the moderator structure, a simple structure (without dummy variables representing all the dependencies) is the most efficient form of the model-building approach, and is effective with binary items (see Study 2); hence that is the structure used here. In future, the model-building R package will automatically generate a complex moderator structure to reduce

the Type I error rate. For the mixed conditions in this study, the simple moderator structure indicates a relationship between the polytomous items. For the all binary and all polytomous test conditions, the moderator structure is arbitrary; a random cluster of items is selected and the moderator highlights those items.

Within each replication, I collect the *QE*, *QM*, β_0, and β_1 values, along with all of their associated *p*-values. This study focuses on the *p*-value of the *QM* statistic, since this indicates whether or not the RE model found evidence of LD.

Results

Results indicate that mixing binary and polytomous data results in larger than nominal detection rates by the model-building approach. The factors influencing the false detection rate are the number of items and the sample size. (Those are factors that will influence the power of any test statistic). The effect detected is relatively small for the $N = 1{,}000$ condition and relatively large for the $N = 4{,}000$ condition. The detection rate in the smaller sample condition ranges between 20% and 25% (see Figure 4.1). For the larger sample size condition, the detection rate ranges between 36% and 73% (Figure 4.2). This indicates that item mixing does have a substantial impact on Q_3 values, and that this effect is moderated by test length. Further research should identify the level of bias resulting from these inflated Q_3 values.

In the all binary and all polytomous conditions, where the moderator is arbitrary, results indicate that the model-building approach rejects the null hypothesis at rates close to the nominal alpha level (0.05). That is, about 5% of replications are statistically significant, which corresponds to our acceptable Type I error rate. Depending on the condition, Type I error rates range between 0.04 and 0.11. Type I error rates are higher if the number of items is small (10

items), but decrease to the nominal alpha as the number of items increases (30 and 50 items) (Figure 4.1, Figure 4.2, and Table 4.3). Similarly, the all polytomous condition has a Type I error rate of 7% for cells with 10 items and 1,000 observations. As the number of items and observations increases, the Type I error rate reduces to about 4%. The number of items has more impact on Type I error than the number of observations. The condition with the most inflated Type I error rate (11%) has a small number of items (10) and a large sample size (4000).

Discussion

Type I errors are closest to nominal in conditions with the largest numbers of items. Ideally, measures should contain more than 10 items before applying the model-building approach. Conditions where the number of items is either 30 or 50 almost always have a Type I error rate close to the nominal level (for both all binary and all polytomous tests). Based on these results, I cannot recommend using the model-building approach for mixed-item tests as no LD was actually generated in the conditions examined.

As for power, the condition with the largest sample size and the smallest number of items had the highest detection rate. This could be due to the small number of items, which resulted in a large variance. In the condition with 10 items, only the first 3 were polytomous, while the other 7 were binary. We can conclude that mixing items on a small test can create issues with local dependence, although these issues may be small.

Based on these results, I cannot recommend item mixing in high-impact tests (i.e., tests where the stakes are high) unless the correlated errors are dealt with through the use of bifactor or hierarchical models. For tests with more items (e.g., at least 30 items), item mixing may produce local dependence.

The factors manipulated in this simulation are limited, with few levels. It would be advisable to increase the number of levels in both of these conditions to better identify both optimal and suboptimal conditions for the model-building approach. Additionally, the mixture of 30% polytomous items with 70% binary items is in the higher range of applied testing. After examining the literature on personality testing, constructed response items make up 5% to 20% of total test items, with only a few cases of 30% or greater (Ercikan et al., 1998; Lissitz et al., 2012; Thissen et al., 1994). Including a wider range of item mixtures would provide a better impression of the impact that common testing practices have on LD.

It would also be helpful to identify the impact of item mixing on other aspects of measurement, like model fit and score accuracy.

Figures & Tables

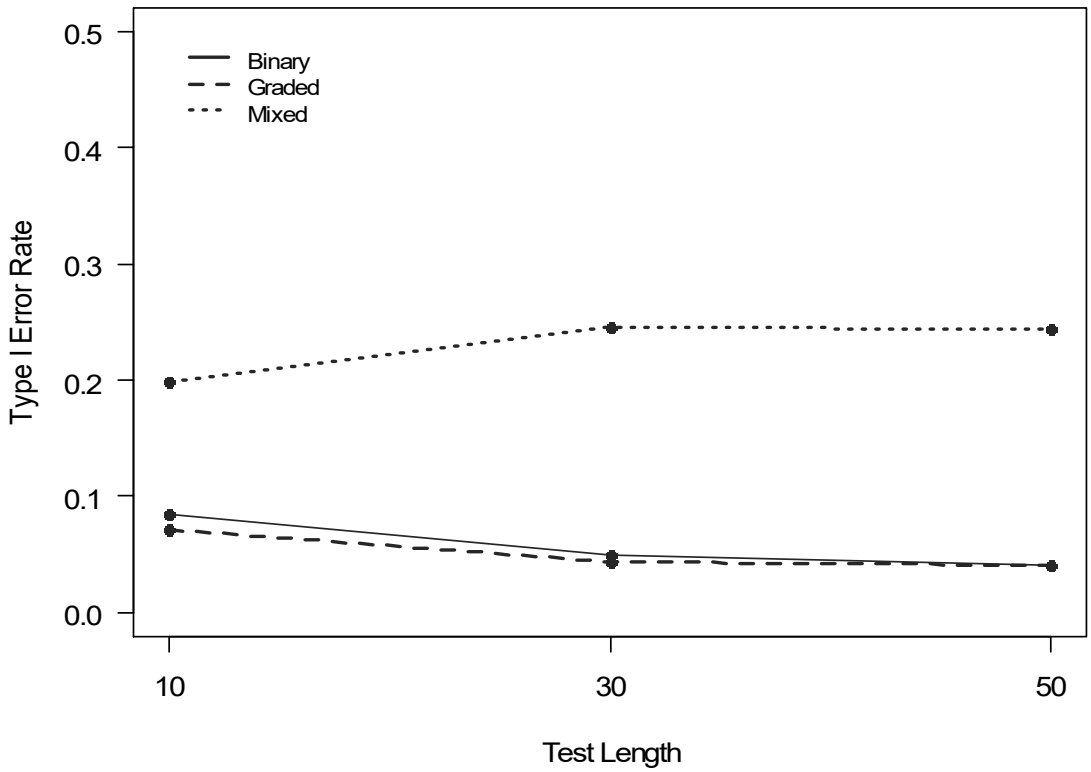


Figure 4.1. Detection rate of the *QM* statistic for binary, graded, and mixed models under the N = 1,000 condition. We expect detection rates to be close to 0.05 for the all binary and all polytomous cases, and to be higher for the mixed cases because item mixing may cause inflated Q_3 values.

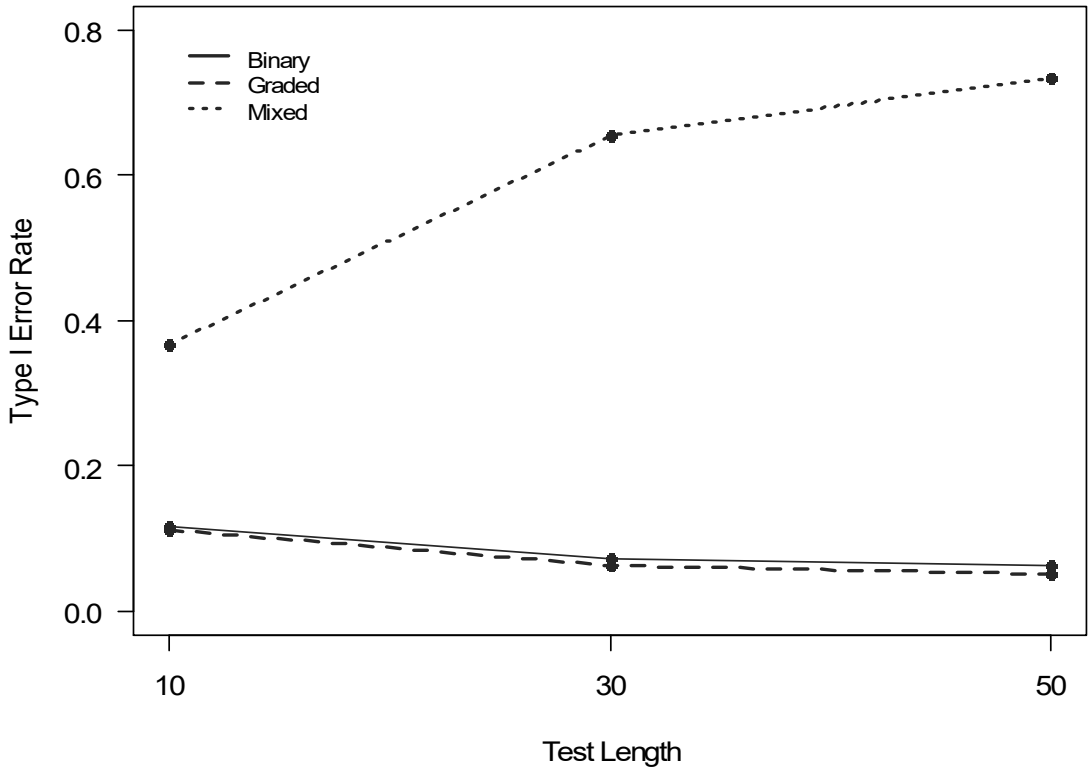


Figure 5.2. False detection rate of the *QM* statistic for binary, graded, and mixed models under the N = 4,000 condition. We expect detection rates to be close to 0.05 for the all binary and all polytomous cases, and to be higher for the mixed cases because item mixing may cause inflated Q_3 values.

Table 5.3

Monte Carlo Simulation Results for Binary, Graded, and Mixed Tests

type	N	P	β_0	D %	β_1	D %	QM	D %	τ^2
Binary									
	1,000	10	-0.0978	100%	0.000	8.44%	1.314	8.44%	0.001
	1,000	30	-0.0325	100%	0.000	5.00%	0.993	5.00%	0.000
	1,000	50	-0.0195	100%	0.000	4.04%	0.922	4.04%	0.000
	4,000	10	-0.0979	100%	0.000	11.67%	1.579	11.67%	0.001
	4,000	30	-0.0326	100%	0.000	7.32%	1.209	7.32%	0.000
	4,000	50	-0.0195	100%	0.000	6.38%	1.070	6.38%	0.000
Graded									
	1,000	10	-0.0979	100%	0.000	7.17%	1.178	7.17%	0.001
	1,000	30	-0.0317	100%	0.000	4.40%	0.949	4.40%	0.000
	1,000	50	-0.0189	100%	0.000	4.13%	0.948	4.13%	0.000
	4,000	10	-0.0980	100%	0.000	11.14%	1.538	11.14%	0.000
	4,000	30	-0.0317	100%	0.000	6.22%	1.109	6.22%	0.000
	4,000	50	-0.0189	100%	0.000	5.05%	1.007	5.05%	0.000
Mixed									
	1,000	10	-0.0958	100%	-0.026	19.82%	2.296	19.82%	0.001
	1,000	30	-0.0314	100%	-0.008	24.57%	2.676	24.57%	0.000
	1,000	50	-0.0188	100%	-0.005	24.40%	2.644	24.40%	0.000
	4,000	10	-0.0957	100%	-0.028	36.70%	4.076	36.70%	0.001
	4,000	30	-0.0314	100%	-0.008	65.50%	7.677	65.50%	0.000
	4,000	50	-0.0188	100%	-0.005	73.50%	8.316	73.50%	0.000

Note. D % (detection percent) is the percent of times the model found an effect with alpha = 0.05, N = sample size, P = test length, β_0 = intercept, β_1 = slope, τ^2 = variance component, binary = all binary item condition, graded = all polytomous item condition, mixed = mixed binary and polytomous item condition.

CHAPTER 5

Illustrated Examples

Purpose

This chapter aims to demonstrate the utility of the model-building approach through two empirical applications. The first application uses the IQ1 measure, a test which is similar to the Raven's Progressive Matrices test (Raven, 2003); the second application uses the Narcissistic Personality Inventory (Raskin & Hall, 1979). Each of these examples demonstrates the utility and flexibility of the model-building approach for detecting violations of local item independence (LII) across different scenarios.

Local Dependence as a Function of Study Characteristics: Progressive Matrices Test

To illustrate the model-building approach on assessing local dependence (LD) caused by qualitative item characteristics, I implement the approach using the 25-item IQ1 test. The IQ1 is a nonverbal intelligence test that presents subjects with a series of 3 x 3 matrices, which contain shapes that follow a different pattern for each item. The dataset, with $N = 400$ participants, is available from a public online repository ("Personality Tests," 2015; personality-testing.info). The repository collects data anonymously from users, who agree to allow their responses to be used for education and research purposes.

The IQ1 test is similar to the Raven's Progressive Matrices test (Raven, 2003). Each pattern appears in a 3 x 3 matrix, where the bottom right block is missing. Participants must

observe the pattern across and down the columns to determine the rules that compose it (Carpenter, Just, & Shell, 1990). Finally, participants attempt to select the most reasonable missing block from a set of possible answers (see appendix A for sample items). As participants progress through the test, the patterns featured become increasingly complex.

Observing the pattern for each item, we can see there are two main rules that must be followed to answer each item correctly. The first rule, evident in question one and question five, consists of a simple pattern with a single figure (e.g., spades, clubs, hearts, or diamonds) across all cells. The second rule can be observed in questions eight and eleven, where participants must understand that there are now two patterns in play – one across rows, and one across columns. These two rules govern the majority of the 25-item test. However, a third rule applies to a subset of items. Items four, twelve, and thirteen employ a central block as a key, which informs the item's solution. These three items share this characteristic, which the rest of the items lack. Due to this shared characteristic, we should inspect these items for LD.

To examine these items for local dependence, we start by recoding the data such that participants receive a one for answering an item correctly and a zero for answering an item incorrectly (Appendix B contains the *R* code to implement this analysis). Next, we apply a 2PL model to the data using the function *mirt(data,1,'2PL')*, where the "1" indicates a unidimensional model and the "2PL" indicates that we are estimating the difficulty and discrimination parameters. It would be more appropriate to apply a 3PL model, because we can expect that participants will guess when they do not know the correct answer. However, since we have only $N = 400$ participants, we could not obtain reliable parameter estimates for the 3PL model (Cook & Eignor, 1991; Hambleton & Jones, 1993). After estimating the model, we can use the function *residuals(object, type="Q3")* to extract the Q_3 values from our data. These values are formatted as a 25 x 25 matrix; we have to extract the upper diagonal and convert it to a vector. This new

vector will have (25*(25 - 1))/2 = 300 values. We apply a Fisher's *r-to-z* transformation to the Q_3 values and compute their conditional variances (defined as 1/397, because $N = 400$).

To set up the moderator, we first must create a vector populated with zeroes. The items that we suspect of possessing LD are 4, 12, and 13, so we must find the location of the Q_3 values generated using those items. The locations are 59, 70, and 78. We can now substitute ones for those elements of the vector. Finally, we can obtain our results using the *summary()* function from base *R*.

The results indicate that the moderator is statistically significant, $QM(1) = 15.70$, $p < .001$ (Table 5.1). The average Q_3 value of the data is -0.046 for the non-suspect items, and this increases by 0.15 for items flagged as locally dependent, supporting our prediction that the shared characteristic creates issues of dependence. This illustrates how the model-building approach can be used to locate item dependence based on common characteristics.

To demonstrate what happens when we misspecify the model, I randomly selected three items (five, seven, and eleven) for which we have no reason to suspect LD. Results indicate that the moderator for these randomly selected items is not statistically significant, $QM(1) = 0.03$, $p = 0.862$ (Table 5.2). There is no evidence of LD for items that were arbitrarily selected. The average Q_3 value for the data is -0.049, and the moderator estimate is -0.0068, which is not significantly different from zero.

LD as a Function of Multiple Factors: The Narcissistic Personality Inventory

The Narcissistic Personality Inventory (NPI) is designed to measure subclinical narcissism (Raskin & Hall, 1979; Raskin & Terry, 1988). The test consists of 40 forced-choice items. Each item presents a pair of statements, one that is narcissistic and one that is non-narcissistic, and test takers must endorse one of the options. For example, participants are asked to select between the statements "I have a natural talent for influencing people" and "I am not good at influencing people." Selecting the first statement over the second indicates a higher level of narcissism. Rasking and Terry (1988) found that the NPI consists of the following seven factors: (a) Authority; (b) Self-Sufficiency; (c) Superiority; (d) Exhibitionism; (e) Exploitativeness; (f) Vanity; and (g) Entitlement (Table 5.3). To assess whether such a structure leads to LD, I apply the model-building approach to a set of NPI responses from an online personality data repository ("Personality Tests," 2015; personality-testing.info). The dataset contains $N = 11{,}243$ anonymous responses collected from an online sample.

Appendix C contains the *R* code to implement this analysis. We first recode the data according to the direction of the response statement (e.g., narcissistic statements receive a one and non-narcissistic statements a zero). Next, we apply a 2PL model to the data and extract Q_3 values from our data. These values are formatted as a 40 x 40 matrix, from which we extract the upper diagonal and convert it to a vector. The new vector has (40*(40 - 1))/2 = 780 values. We apply a Fisher's *r*-to-*z* transformation to the Q_3 values and compute their conditional variances ($N = 11{,}243$, so $v = 1/11{,}240$).

To specify moderators to model the multivariate structure of the data, we first must create a set of 6 moderators (one for each factor minus one, to avoid collinearity) that contain only zeroes. Finally, we specify which transformed Q_3 values involve the items thought to load on each specific factor. For example, Factor 4 (Exhibitionism) is composed of items 15, 19, and 29.

The Q_3 values that belong to these items are located at positions 168, 393, and 397, so we can replace the zeroes with ones at these locations. We repeat this process for all 6 moderators. Finally, we obtain our results by using the function *summary(rma(q3~mods,v))*, with *summary()* from the base functions for *R* and rma() from the *metafor* package (Team, 2014; Viechtbauer, 2010).

I fit the model with six moderators (one for each latent variable minus one), and the results indicate that the moderator structure specified is statistically significant, $QM(6) = 312.89$, $p < .001$ (Table 5.4). Notice that every moderator increased the transformed Q_3 values by at least 0.07, and one by as much as 0.40. This suggests a strong degree of local dependence, or multidimensionality. Therefore, these data must be modeled using techniques that can account for multidimensionality, such as a bifactor model or a multi-dimensional item response model. This example illustrates how the model-building approach can be used to assess the suspected presence of multidimensionality.

Discussion

This chapter highlights the utility and flexibility of the model-building approach to assess LD based on surface-level characteristics of items. In the first example, I model LD based on a key characteristic of three items that are distinctively different from the rest of the items, since they rely on a mechanism that guides the participant to correctly answer the item. Failing to detect an effect could provide evidence supporting the use of a unidimensional IRT model. Detecting an effect, on the other hand, would provide motivation to further inspect the items for LD. In the case of the IQ1, a researcher should apply a different method to account for the dependence between the three items, such as the bifactor model or a testlet IRT model (Thissen et

al., 1989). In the case of the NPI, the model detected an overall effect for each moderator, indicating that the test represents a multidimensional trait.

The *R* code that is included to carry out the model-building approach can be adapted and modified for any type of IRT model supported by the *mirt R* package (Chalmers, 2012). If the Q_3 matrix is obtained using a different software package, the model-building approach can be conducted using other popular commercial meta-analysis software, such as HLM (Raudenbush, Bryk, & Congdon, 2004), SAS (SAS Institute Inc., 2003), Stata (StatCorp, 2013), or SPSS (IBM Corp., 2013).

Tables and Figures

Table 5.1

Results of the Model-Building Approach-IQ1 Progressive Matrices Test

Moderator	B	SE	z-val	*p*-val
Intercept	-0.0465	0.0038	-12.1334	<0.0001
Moderator	0.1518	0.0383	3.9628	<0.0001

Note. *QE* (298) = 515.49, $p < .001$; *QM* (1) =15.70,p<0.001; τ^2=.0018.

Table 5.2

Results of the Model-Building Approach-IQ1 Progressive Matrices Test – Misspecified Model

Moderator	B	SE	*z*-val	*p*-val
Intercept	-0.0449	0.0039	-11.4228	<0.0001
Moderator	-0.0068	0.0393	-0.1737	0.8620

Note. *QE* (298) = 515.49, $p < .001$; *QM* (1) = 0.03,p=0.862; τ^2 = .0021.

Table 5.3

Factor Structure of the Narcissistic Personality Inventory

Authority	Entitlement	Superiority	Vanity	Self-Sufficiency	Entitlement	Exploitative
1	5	4	15	17	2	6
8	14	9	19	21	3	13
10	18	26	29	22	7	16
11	24	37		31	20	20
12	25	40		34	28	23
32	27			39	30	35
33					38	
36						

Table 5.4

Results of the Model-Building Approach-NPI

Moderator	β	SE	z-val	p-val
Exploitative	-0.0346	0.0023	-15.0976	<0.0001
Authority	0.0883	0.0116	7.6178	<0.0001
Entitlement	0.0872	0.0157	5.5516	<0.0001
Superiority	0.0854	0.0192	4.4550	<0.0001
Vanity	0.4120	0.0348	11.8382	<0.0001
Self-Sufficiency	0.0764	0.0157	4.8641	<0.0001
Entitlement	0.1069	0.0133	8.0232	<0.0001

Note. QE (773) = 31438.51, $p < .001$; QM (6) = 312.89, $p < .001$; $\tau^2 = .0035$.

CHAPTER 6

Concluding Remarks

The assumption of local item independence (LII) is central to item response theory (IRT) models. The probability of responding to each item should be independent after accounting for the latent construct, resulting in uncorrelated item residuals (McDonald, 1981; Steinberg & Thissen, 1996). Violating this assumption leads to many problems, from generating biased parameter estimates to artificially shrinking the standard errors, inflating test reliability estimates, and impacting model fit (Chen & Thissen, 1997; Liu & Maydeu-Olivares, 2012; Yen, 1984; Zenisky et al., 2001). Preventative steps can be taken to minimize the chance of locally dependent (LD) items (Steinberg & Thissen, 1996; Yen, 1993). Examining items during the test development phase for item chaining, explanations provided in previous items, and similar wording can avoid or minimize LD. During item scaling, multiple statistics exist to identify LD. Yen's (1993) Q_3, Chen and Thiene's (1997) $LD\text{-}X^2$ and $LD\text{-}G^2$ are among the most popular exploratory approaches. There are also some confirmatory approaches, like the score test (S_b) for bifactor and threshold shift (S_t) models (Liu & Thissen, 2014).

The new model introduced in this dissertation is the model-building approach to assessing Q_3; it is an additional tool that researchers can use to identify and model LD. This approach combines techniques from the meta-analytic literature with the diagnostic assessment of IRT models. By treating Q_3 values as effect sizes, a meta-analytic model can estimate and test the degree of LD for item groups based on observed characteristics, such as shared content or item

structure. A series of simulation studies, along with applications on real data, assess the effectiveness of the model-building approach.

Simulations indicate that the model can adequately capture violations of local item dependence when such dependence exists. It can capture and model the degree to which item pairs and item clusters are locally dependent within a dataset, based on surface level characteristics. A summary of the simulations can be found in Table 6.1. This model possesses good statistical properties, with a nominal detection rate under null conditions; that is, this method will rarely flag an item pair or item cluster when no dependence exists. Additionally, when there is an effect, the method has strong power even with as few as 15 items and moderate sample sizes.

Table 6.1
Dissertation Simulation Overview

	Sim 1	Sim 2	Sim 3	Sim 4
Chapter	2	2	3	4
Models	3PL	3PL	2PL & GRM	2PL & GRM
LD	Pair	Cluster	Pair	Cluster
Pars	a ~log$N$(0,0.5^2) b~N(0,1.5^2) g~P(N(-1.1,.5^2))	a ~log$N$(0,0.5^2) b~N(0,1.5^2)	a~N(1.7,0.3^2) b1~N(-1,.5^2) b1+bk~N(1,0.2^2) b~N(0,1.5^2)	a~N(1.7,0.3^2) b1~N(-1,0.5^2) b1+bk~N(1,0.2^2) b~N(0,1.5^2)
Factors	ρ=1,0.5,0.2 P = 15, 25, 30, 50 N = 1000, 4000	ρ=1,0.5,0.2 P =15, 25, 30, 50 N = 1000, 4000	ρ=1,0.5,0.2 P = 25, 30, 50 N = 1000, 4000	ρ=1,0.5,0.2 P = 10, 30, 50 N = 1000, 4000 Mixed or not
Effects	FE & RE How well does this model compare to X^2, G_2, and Q_3?	FE & RE How well does this model handle item clusters?	RE How does this model perform with polytomous data?	RE How well does this model work with mixed-format tests?

I demonstrated the utility of the model-building approach using two applied examples. The approach successfully identified local item dependence based on suspect characteristics from the set of items used in the measures. In the first example, progressive matrices were introduced to a series of students. The matrices were governed by several rules in order to reach a correct answer. One rule in particular was apparent because only a few (3) items shared it. Once this feature became clear, the test taker could answer the more difficult questions with ease. The second example, the Narcissistic Personality Inventory (NPI), showed that the model-building approach can be used when researchers suspect multidimensionality. The NPI produces total scores of narcissism, but Raskin and Terry (1988) identified several factors that highlighted a possible multidimensional aspect. Here, the test characteristics were the factors to which each item potentially belonged. The model-building approach successfully identified each of these factors as significant, possessing local dependence; the moderators used to identify the item-factor relationships were significant. Finally, I randomly generated a moderator to demonstrate that in the absence of multidimensionality, the moderator was not significant. This does not *prove* that the model is correct, but does lend support and confidence to the previous results.

These results demonstrate the utility of the model-building approach and its appealing statistical properties. However, the empirical simulation studies in this dissertation are far from exhaustive; for instance, participant sample size was restricted to three levels across most studies included. The levels were selected with prior research and potential application in mind, but examining a greater number of sample sizes could be useful. This also extends to selecting the levels of local dependence and test length. In particular, a fine-grain approach should be taken to generating and investigating local dependence, in order to fully capture the precision of the model-building approach. This includes adding more levels to the factors of sample size, test length, and degree of local dependence.

In conclusion, the model building approach has demonstrated favorable statistical properties that give researchers a new tool to assess the assumption of local dependence in IRT models.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), 7–16. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Arthur, W., & Day, D. V. (1994). Development of short form for the Raven advanced progressive matrices test. *Educational and Psychological Measurement*, *54*(2), 394–403.

Bacon, J. G., Scheltema, K. E., & Robinson, B. E. (2001). Fat phobia scale revisited: the short form. *International Journal of Obesity and Related Metabolic Disorders : Journal of the International Association for the Study of Obesity*, *25*(2), 252–257. https://doi.org/10.1038/sj.ijo.0801537

Beier, M. E., & Oswald, F. L. (2012). Is cognitive ability a liability? A critique and future research agenda on skilled performance. *Journal of Experimental Psychology: Applied*, *18*(4), 331–345. https://doi.org/10.1037/a0030869

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*(3), 225–264. https://doi.org/10.1207/s15324818ame0503_4

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*, 4279–4292.

https://doi.org/10.1002/sim.2673

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows. Lincolnwood, IL: Scientific Software International, Inc.

Campbell, D. T. (1988). Administrative experimentation, institutional records, and nonreactive measures. In *Methodology and Epistemology for Social Sciences: Selected Papers* (pp. 243–260). University of Chicago Press.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Chalmers, P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Christensen, K. B., Makransky, G., & Horton, M. (2016). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, 1–17. https://doi.org/10.1177/0146621616677520

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129.

Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature*, *505*(7485), 612–613. https://doi.org/10.1038/505612a

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45.

Cooper, A., & Petrides, K. V. (2010). A psychometric analysis of the Trait Emotional Intelligence Questionnaire-Short Form (TEIQue-SF) using item response theory. *Journal of Personality Assessment*, *92*(5), 449–457. https://doi.org/10.1080/00223891.2010.497426

Corp IBM. (2013). IBM SPSS for Windows. Armonk, NY: IBM Corp.

Davidian, M. (2005). Simulation studies in statistics. *Simulation*. Retrieved from http://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186–219. https://doi.org/10.1037/a0031609

Dixon, P. M. (2002). Bootstrap resampling. *Encyclopedia of Environmetrics*, *19*(1), 9. https://doi.org/10.1002/9780470057339.vab028

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Taylor & Francis.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137–154. https://doi.org/10.2307/1435236

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusterse in a large scale hands-on science performance assessment. *CBT/McGraw-Hill*.

Glas, C. A., & Falcon, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

https://doi.org/10.1177/0146621602250530

Gloster, A. T., Rhoades, H. M., Novy, D., Klotsche, J., Senior, A., Kunik, M., … Stanley, M. A. (2008). Psychometric properties of the Depression Anxiety and Stress Scale-21 in older primary care patients. *Journal of Affective Disorders*, *110*(3), 248–259. https://doi.org/10.1016/j.jad.2008.01.023

Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W., & Oliver, P. H. (2009). A latent curve model of parental motivational practices and developmental decline in math and science academic intrinsic motivation. *Journal of Educational Psychology*, *101*(3), 729–739. https://doi.org/10.1037/a0015084

Hallquist, M. (2016). Package " MplusAutomation ."

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Hambleton, R. K., & Swaminathan, H. (1985). Assumptions of item response theory. In *Item Response Theory* (pp. 15–31). Springer Netherlands.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics*, *17*(4), 297–313.

Harwell, M. R., Rubinstein, E., Hayes, W., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*(4), 315–339. https://doi.org/10.3102/10769986017004297

Harwell, M. R., Stone, C. A., & Hsu, T. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504.

Hedges, L. V, & Olkin, I. (1985). *Statistical methods for meta-analysis* (1st ed.). Academic Press.

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. *The British Journal of Clinical Psychology / The British Psychological Society*, *44*(Pt 2), 227–39. https://doi.org/10.1348/014466505X29657

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model*. Chapel Hill: University of North Carolina.

Hoaglin, D. C. (2016). Misunderstandings about Q and "Cochran"s Q test' in meta-analysis. *Statistics in Medicine*, *35*(4), 485–495. https://doi.org/10.1002/sim.6632

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika*, *46*(1), 79–92. Retrieved from http://hdl.handle.net/10995/23838

Houts, C. R., & Edwards, M. C. (2013). The Performance of Local Dependence Measures With Psychological Data. *Applied Psychological Measurement*, *37*(7), 541–562. https://doi.org/10.1177/0146621613491456

Houts, C. R., & Edwards, M. C. (2015). Comparing surface and underlying local dependence levels via polychoric correlations. *Applied Psychological Measurement*, *39*(4), 293–302. https://doi.org/10.1177/0146621614561492

Inc., S. I. (2003). SAS Software. Canary, NC.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*(1), 109–132. https://doi.org/10.1007/BF02295736

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *The British Journal of Mathematical and Statistical Psychology*, *63*(2), 395–416. https://doi.org/10.1348/000711009X466835

Jansen, M. G. H. (2007). Testing for local dependence in Rasch's multiplicative gamma model for speed tests. *Journal of Educational and Behavioral Statistics*, *32*(1), 24–38. https://doi.org/10.3102/1076998606298032

Jorm, A. F. (1994). A short form of the Information Questionnaire on Cognitive Decline in the Eldery (IQCODE): development and cross-validation. *Psychol Med*, *24*(1), 145–153.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 65–81. https://doi.org/10.1177/01466216000241004

Kean, J., & Reilly, J. (2014). Item response theory. In *Handbook for Clinical Research: Design, Statistics and Implementation* (pp. 195–198). New York, NY: Demos Medical Publishing.

Kim, D., de Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). Assessing relative performance of local item dependence (LID) indexes. *Applied Psychological Measurement*, *35*(6), 447–471.

Koning, E., Sijtsma, K., & Hamers, J. H. M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement*, *26*(3), 302–320. https://doi.org/10.1177/0146621602026003005

Le Moine, J. G., Fiestas-Navarrete, L., Katumba, K., & Launois, R. (2016). Psychometric validation of the 14 items chronic venous insufficiency quality of life questionnaire (CIVIQ-14): confirmatory factor analysis. *European Journal of Vascular and Endovascular Surgery*, *51*(2), 268–274. https://doi.org/10.1016/j.ejvs.2015.08.020

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74–100. https://doi.org/10.1191/0265532204lt260oa

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (illustrate). University of Michigan: Sage Publications. https://doi.org/0761921672

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment. *Journal of Applied Testing Technology*, *13*(3), 1–52.

Liu, Y., & Maydeu-Olivares, A. (2012). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254–274. https://doi.org/10.1177/0013164412453841

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*(8), 670–688. https://doi.org/10.1177/0146621612458174

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *The British Journal of Mathematical and Statistical Psychology*, *67*(3), 496–513. https://doi.org/10.1111/bmsp.12030

Livingston, S. (2009). Constructed-response test questions: Why we use them; How we score them. *R & D Connections*, *11*(11), 1–8. Retrieved from http://144.81.87.152/Media/Research/pdf/RD_Connections11.pdf

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. (A. Birnbaum, Ed.). Oxford, England: Addison-Wesley.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15–29. https://doi.org/10.1111/j.1745-3984.1988.tb00288.x

McClellan, C. A. (2010). Constructed-response scoring — doing it right. *R&D Connections*, (13). Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

Mcdonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6*(4), 379–396. https://doi.org/10.1177/014662168200600402

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100–117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

Milian, M., Kreitschmann-Andermahr, I., Siegel, S., Kleist, B., Führer-Sakel, D., Honegger, J., … Psaras, T. (2015). Validation of the Tuebingen cd-25 inventory as a measure of postoperative health-related quality of life in patients treated for Cushing's disease. *Neuroendocrinology*, *102*, 60–67. https://doi.org/10.1159/000431022

Monahan, J. F. (2009). A guide for simulation studies in statistics.

Muthén, L. K., & Muthén, B. O. (2007). Mplus user's guide. *Journal of the American Geriatrics Society*, *2006*, 676. https://doi.org/10.1111/j.1532-5415.2004.52225.x

Muthén, L. K., & Muthén, B. O. (2009). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904

National Institute of Health. (2016). Reproducibility. Retrieved March 16, 2017, from

https://grants.nih.gov/reproducibility/faqs.htm#4834

NCES. (2008). NAEP technical documentation: Scoring monitoring. Retrieved March 10, 2017, from https://nces.ed.gov/nationsreportcard/tdw/scoring/scoring.asp

Patel, P. C., Messersmith, J. G., & Lepak, D. P. (2013). Walking the tightrope: An assessment of the relationship between high-performance work systems and organizational ambidexterity. *Academy of Management Journal*, *56*(5), 1420–1442. https://doi.org/10.5465/amj.2011.0255

Paxton, P., Curran, P. J., Bollen, K. A., & Kirby, J. (2001). Monte Carlo experiments : Design and implementation. *Structural Equation Modeling*, *8*(2), 278–312. https://doi.org/10.1207/S15328007SEM0802

Personality Tests. (2015). Retrieved October 9, 2015, from http://personality-testing.info/_rawdata/

Pianta, R. C. (1992). Student-teacher relationship scale: Short form. *Unpublished Instrument*, 1. https://doi.org/10.1017/CBO9781107415324.004

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. *Danish Institute for Educational Research*. Copenhagen: Danish Institute for Educational Research.

Raskin, R., & Hall, C. S. (1979). Narcissistic Personality Inventory. *Psychological Reports*, *45*, 590. https://doi.org/10.1037/0022-3514.54.5.890

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*(5), 890–902. https://doi.org/10.1037/0022-3514.54.5.890

Raudenbush, S. W., Bryk, A., & Congdon, R. (2004). HLM 6 for Windows. Skokie, IL: Scientific

Software International, Inc.

Raven, J. (2003). Raven progressive matrices. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 223–237). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-0153-4_11

Reese, L. M. (1995). The impact of local dependencies on some LSAT outcomes. *LSAC Research Report Series*.

Samejima, F. (1969). Estimation of latentability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4).

Shrout, P. E., & Fiske, S. T. (2014). *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. New York, NY: Psychology Press.

StatCorp. (2013). Stata statistical software. College Station, TX: StataCorp LP.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In *Personality Research, Methods and Theory: A Festschrift Honoring Donald W. Fiske* (pp. 161–181).

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*(1), 81–97. https://doi.org/10.1037//1082-989X.1.1.81

Studerus, E., Gamma, A., & Vollenweider, F. X. (2010). Psychometric evaluation of the altered states of consciousness rating scale (OAV). *PLoS ONE*, *5*(8). https://doi.org/10.1371/journal.pone.0012412

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology*, *11*, 249–288. https://doi.org/10.1016/S0065-2601(08)60009-X

Teachman, B., Marker, C. D., & Smith-janik, S. B. (2009). Automatic associations and panic disorder: Trajectories of change over the course of treatment. *Journal of Consulting and Clinical Psychology*, *76*(6), 988–1002. https://doi.org/10.1037/a0013113.Automatic

Team, R. C. (2014). R: A language environment for statistical computing. Vienna, Austria.: R Foundation for Statistical Computing.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, *16*(1), 109–119. https://doi.org/10.1007/slll36-007-9169-5

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501–519. https://doi.org/10.1007/BF02302588

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 148–177). Sage Publications. https://doi.org/http://dx.doi.org/10.4135/9780857020994.n7

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*(1), 77–83. https://doi.org/10.3102/10769986027001077

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets : A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests ? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113–123.

https://doi.org/10.1111/j.1745-3984.1994.tb00437.x

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology*, *38*(2), 227–242. https://doi.org/10.1177/0022022106297301

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(2), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wainer, H., & Thissen, D. (1993). Applied measurement in education combining multiple-choice and constructed-response test scores : Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118. https://doi.org/10.1207/s15324818ame0602

Wainer, H., & Wang, X. (2001). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203–220.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. https://doi.org/10.1080/00273170802620121

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments : Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test, and ability statistics. *MCAT Monograph*.

Appendix 2.A

Monte Carlo Methods and Purpose

This section gives readers a gentle introduction to Monte Carlo (MC) simulation studies and their methods, background, and applications. First, I introduce the concept of MC methods and their applications. Next, I discuss the main features that make MC studies in statistics and psychology useful. Finally, I discuss the important decisions that researchers must make when moving forward with a full Monte Carlo study. This chapter can serve as a primer for applied researchers who are thinking of conducting simulation work and for incoming quantitative psychology graduate students.

Monte Carlo simulation serves an important purpose in the worlds of finance, healthcare, energy, manufacturing, and any other field that requires the use of statistics and projections. The idea behind a Monte Carlo simulation is simple – select a model and a set of parameters, sample data from the implied distribution a large number of times, and estimate parameters and other items of interest (e.g., inferential results or confidence intervals) from each sample. Each sample will differ slightly from the population, but over repeated samples, the average estimates will mirror the population. For example, assume that we want to know the probability of rolling two six-sided dice and having both land on five. The question implies a binomial distribution with two trials, and a probability of success on each trial equal to 1/6. It is easily solved using analytic methods – given independent dice, simply multiply the probability that the first die lands on the number five, 1/6, by the probability that the second die lands on 5 (also 1/6). The product of these probabilities will be $1/6*1/6 = 1/36 \approx .0278$. Suppose that we did not have the background in basic probability, but instead had access to a computer and knowledge of a computing language. We could simulate the condition by generating a series of discrete, integer-valued, uniformly distributed random numbers ranging from 1 - 6 and counting the number of times we observe both rolls resting on five.

```
results<-rep(0,10000)
```

```
for(i in 1:10000){
   rolls<-sample(1:6,2,replace=TRUE)
   if(rolls[1]==5 & rolls[2]==5){results[i]<-1}
   return(results)
}
sum(results)/10000
```

Although this example is trivial, it demonstrates the utility of Monte Carlo methods and their use in solving probability problems. This approach can be employed for multiple reasons, including to address a function that does not have a closed form solution. If a researcher is interested in applying a known statistical model on a population with an unknown distribution, the researcher can apply MC methods to conduct a power analysis before the actual study begins. They can also use MC to approximate the sampling distribution of a statistic and use the middle 95% as a confidence interval, a procedure called parametric bootstrapping or bootstrap resampling (Dixon, 2002).

Monte Carlo Studies

Monte Carlo studies, or Monte Carlo experiments, are particularly important in assessing the properties of new statistical models under typical conditions or established statistical models under atypical conditions such as high variability, an unknown population distribution, or model misspecification (e.g., Depaoli, 2013; Liu & Thissen, 2014). With an increase in computing power in the last 20 years, the relative barrier for entry into increasingly complex simulation studies has diminished, and the popularity of "canned" statistical software like *R* and *MPLUS* makes conducting MC studies easier (Muthén & Muthén, 2007; Muthén & Muthén, 2009; R Core Team, 2014). The increased processing power has allowed even the most applied researchers to implement MC studies for answering

simple questions, like assessing sample size requirements and power (Gloster et al., 2008; Gottfried, Marcoulides, Gottfried, & Oliver, 2009; Patel, Messersmith, & Lepak, 2013; Studerus, Gamma, & Vollenweider, 2010; Teachman, Marker, & Smith-janik, 2009). Additionally, funding agencies like the National Institute of Health have started to implement rigor and transparency rules that encourage the reporting of power analyses before studies are considered for funding, in an effort to enhance reproducibility (Collins & Tabak, 2014; National Institute of Health, 2016). Power analyses for complex SEM models require the use of MC methods. For all of these reasons, MC has seen increased use from both methodologists and applied researchers.

In the context of computer program experiments, Monte Carlo simulation studies are used to evaluate the properties of estimators or of hypothesis tests (Burton, Altman, Royston, & Holder, 2006; Monahan, 2009). Simulation studies evaluate the performance of models across different scenarios that are of interest to substantive researchers. Such scenarios may include the performance of an estimator under conditions of high variability, small sample sizes, or small effect sizes. For example, Depaoli (2013) conducted an MC study investigating the conditions under which a growth mixture model could detect the correct number of mixture classes across degrees of class separation. In addition, she compared frequentist and Bayesian estimation methods. In this study, Depaoli investigated both the performance of two estimation methods (frequentist vs. Bayes) and the detection accuracy under varying conditions, including class separation and sample size. She identified the conditions under which substantive researchers could confidently apply growth mixture modeling methods.

Although MC studies are widely used in quantitative psychology, few sources offer standardized procedures for performing such studies. Paxton et al. (2001) attempt to address this issue by proposing a set of steps to standardize the practice of MC studies. The steps are as follows: (a) propose a research question; (b) generate representative models; (c) select values for the parameter estimates of the model; (d) choose a software package; (e) execute simulations; (f) save the file output; (g) troubleshoot/verify the data; and (h) summarize the results. These steps, though, overlook some important issues, such as: (a)

whether there have been sufficient replications to provide a reliable answer to the simulation questions; (b) what factors should be varied; and (c) what outcomes are of interest (e.g., coverage rates, bias, or root mean squared error).

Pilot Studies. A pilot study is always a good idea in empirical research. MC simulations are a form of empirical research, so a pilot simulation is always a good idea before attempting a full-scale experiment. A pilot study is a practical way of anticipating problems that may arise later on. Pilot studies also provide ways to identify or confirm important features that may not have been highlighted during the literature review (e.g., unique ways in which a statistic or a model might be incorrectly applied). A rule of thumb for selecting parameters for a pilot study is to identify both typical levels and levels that you expect will be problematic (e.g., very low sample size conditions). This can also serve as a way to identify the minimum number of replications that the simulation requires to stabilize. If the most difficult cells require, say, 10,000 replications to arrive at a stable parameter estimate, the simpler cells may require far fewer replications.

Issues Important in MC studies

Factors and Levels. A researcher planning an MC experiment may think of many factors to manipulate, including multiple levels for each factor. This can result in a prohibitively large design. Consider, for example, a simple simulation to investigate a statistic. One may naturally want to vary sample size and parameter magnitude. In this simulation, choosing five levels for the sample size and five levels for the parameter value, the resulting design will have 25 cells. Suppose we add a third factor, distribution If this factor has three levels (e.g., normal, negatively skewed, and positively skewed), the design now has 5*5*3 = 75 cells. Any additional condition, however justifiable, may double (150) or even triple (225) the number of cells. Allowing a design to proliferate in this way can produce results that are impossible to summarize succinctly.

Other limitations to the number of factors and levels are computation time and resources. Depending on the complexity of the simulation, researchers may spend weeks or even months running a single simulation cell. Simulations can be split across multiple computers and processor cores, but even with such resources, reducing the number of factors or levels will always reduce the time of running a simulation. In addition, note that not all factors require a new cell (or set of cells). For example, a researcher interested in comparing five different models can generate a single dataset and apply all five models to a single replication within a cell.

Monahan (2009) puts forward the view that simulation studies can select similar levels to those previously studied. This allows for simple comparisons across multiple simulations. This has to be weighed, however, with the importance of identifying factors and levels that feature in the real world (situations applied researchers might encounter). If not, applied researchers may dismiss the study as irrelevant (e.g., a simulation study that focuses on large sample sizes in an industry like medicine, which encounters modest samples). One benefit of selecting some levels similar to those of previous MC studies is that results can be combined via meta-analysis, and overlapping factors can be used to assess the validity of the study, akin to the way researchers can meta-analyze applied studies (Harwell, 1992; Harwell, Rubinstein, Hayes, & Olds, 1992).

Parameters. In the preceding section, I described the need to vary parameter values as part of the factorial design of the simulation. However, some parameters in a model may be not particularly relevant. The experimenter must carefully consider such nuisance parameters, and, in particular, consider which may be set as constants and which should vary between replications in the simulation. When a certain parameter is not of interest to the researcher (e.g., the item difficulty of a set of items), the researcher must decide whether to fix the values or to sample them from a distribution. Each choice has its strengths and weaknesses. Using a constant parameter reduces the amount of error for each MC replication. One possible drawback is that these results may be less generalizable than those chosen from a distribution of parameters. This decision is similar to the decisions that applied researchers must make

when choosing to conduct either laboratory studies or field experiments – laboratory studies may be more precise, but the results are less generalizable (Campbell, 1988). Ultimately, the decision rests on the field in which the simulation study will be published and on researcher's preference.

Replications. Another decision is that of determining the number of replications (or simulation iterations) to conduct. Some simulations may be large and take an excessive amount of time to conduct a large number of replications (say, more than 100,000). (This is likely true, for example, with Bayesian models, where the estimation requires simulations within each simulation.) To adequately report the results from a simulation study, researchers must ensure that the final answers have reached a point of stability such that further replications would not change the results.

One method of assessing convergence involves estimating and plotting a running mean for each parameter of interest. Plotting this running mean (a summary statistic) against the replication number can be a good approximation of the true sampling properties of the test statistic under the specified conditions (Davidian, 2005). This gives the researcher a clear goal for the number of replications needed to obtain stable parameter estimates.

If the researcher fails to achieve convergence after an excessively large number of replications, there are many possible causes. If the model is complex and each parameter is sampled from a distribution (resulting in added variability), the picture may take some time to stabilize. An unusual sample may result in a model that is not estimable. When that occurs, parameters are likely to asymptote to infinity, producing results that are clearly incorrect. Assessing model convergence in situations for which iterative estimation is necessary can help identify such pathological situations. Examining the results via a histogram can also help. Pilot studies are useful to assess is the likelihood that such issues will arise in the full simulation study. There can also be problems within the code itself if, for example, a statistic is calculated improperly or a seed is set incorrectly. This can be addressed by inspecting the code line-by-line and step-by-step to ensure that each piece does what it should.

Assessment criteria

Bias. Bias is the difference between the true data generating parameter and the estimated parameter based on the sample: $bias = \bar{\hat{\theta}} - \theta$, where $\bar{\theta}$ is the average estimate of the parameter of interest over all replications, θ is the true parameter value and $\hat{\theta}$ is the estimate (Burton et al., 2006). Bias is an important criterion when assessing simulations, in that it quantifies the precision of the instrument in question. Expressing bias as a percentage is helpful in interpreting the degree to which the estimated parameter differs from the population parameter, $\left(\frac{bias}{\theta}\right) * 100$, although relative bias can be unstable if the true parameter value is near zero.

Variance. Variance is another source of estimation error. In the context of MC studies, variance refers to the sampling variability of the parameter estimate, and can be expressed as$(SE(\bar{\hat{\theta}}))^2$. Variability indicates the spread of the estimated parameters. A large spread indicates a greater amount of uncertainty in each replication.

Mean Squared Error. Mean squared error (MSE) is the total parameter estimation error, including that from both bias and variance. The MSE can be expressed as $MSE = bias + (SE(\hat{\theta}))^2$. It is important to assess both sources of variability independently first, before examining the total error. When evaluating the MSE, researchers should consider the bias/variance tradeoff. Due to the two sources of error, the MSE may be the same for an estimation model that has high bias and low variance as it is for a model that has low bias and high variance.

Coverage. Coverage is the proportion of times that an estimated parameter's confidence interval captures the true parameter value. The coverage rate for a confidence interval can be calculated by assessing how many times $\theta_i \pm Z_{1-\alpha/2}\, SE(\hat{\theta}_i)$ captures the true parameter θ and dividing this number by

the total number of replications in the simulation, where $Z_{1-\alpha/2}$ is the critical value. We expect to capture the true parameter $100 * (1 - \alpha)$ percent of the time. It is also worth noting that confidence intervals with varying target coverage rates may be of interest. Typically, researchers evaluate coverage rates for 95% intervals, but it is possible that 95% intervals can work well where, for example, 99% intervals are far from nominal coverage. For that reason, it is also useful to evaluate the distribution of p-values for z-tests of the null hypothesis that $\hat{\theta} = \theta$. A uniform distribution of those p-values will confirm the correct coverage rates for all possible confidence intervals.

Power & Type I Error. The performance of hypotheses a model produces is the focus of many simulations. Type I error is defined as the proportion of times that we can reject the null hypothesis when there truly is no effect (i.e., a null condition in the simulation). Power indicates the number of times that the simulation finds an effect when there is indeed an effect. Of these two, a researcher should first focus on identifying the Type I error rate. A proper model's Type I error rate should be approximately α (typically 0.05 in psychology). Smaller Type I error rates indicate that the model is conservative (Type I error < α). A larger Type I error rate indicates the model is liberal (Type I error > α). Once Type I error rates have been established, the researcher can assess the model's performance under power conditions. A model with a nominal Type I error rate (type I error = α) and high power has excellent statistical properties. A model with a high Type I error rate and high power is flawed and merits a revision.

Discussion

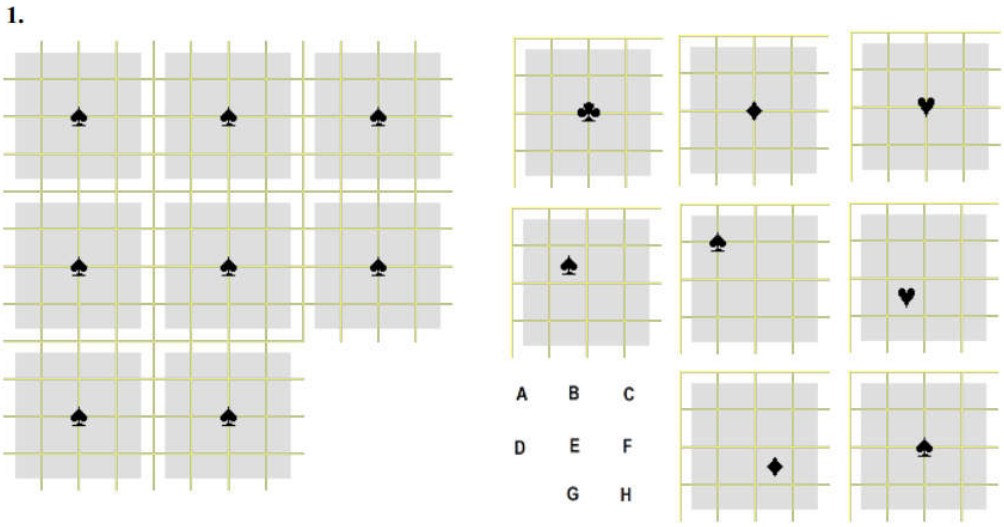
Monte Carlo simulation serves an important purpose in psychology and statistics. The increase in processing power and access to software has allowed researchers to use MC studies in a wide array of settings, including sample size planning. *R* packages like *MplusAutomation*, used in conjunction with programs like *MPLUS*, make setting up simulation studies and planning empirical studies simpler (Hallquist, 2016; Muthén & Muthén, 2007).

To summarize, MC studies should be treated like traditional experimental studies. Researchers should propose a research question, identify all of the parameters that will be monitored and the factors that will be manipulated, and consider how the results can be clearly presented. The researcher should avoid increasing the number of cells to a point where the study becomes unmanageable. It helps to implement a pilot study including both typical and atypical cells to understand the time that the entire simulation will take and how many replications it will require. Once the pilot study is complete, close examination of the results and an inspection of the code are necessary to identify potential problems. It is better to catch problems early, before the full simulation study, in order to save time.

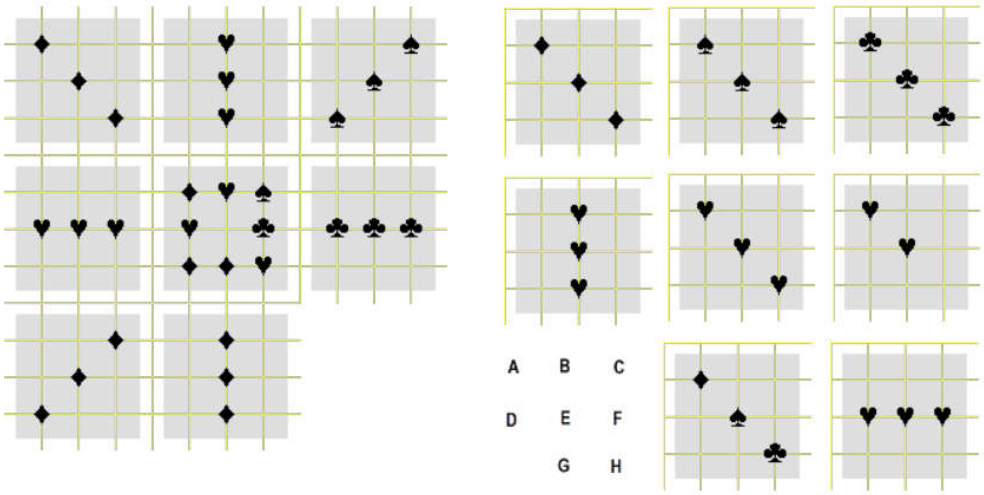
Appendix 6.A

Example Item for the Progressive Matrices Test (IQ1)

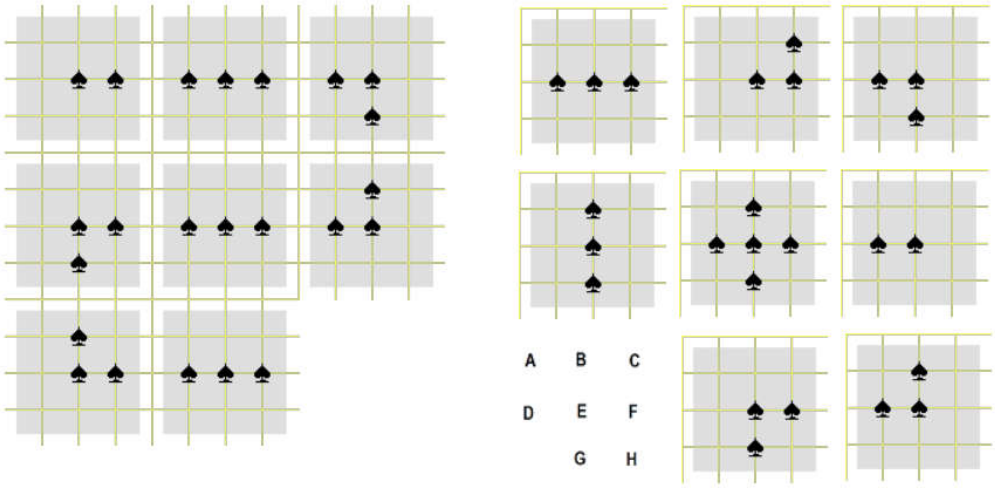
Question 1, 5, 8, and 11 show the first two rules to correctly answering the item. Question 4 demonstrates the radial rule where the central block serves as the key. The left-hand side under each item is the stem, and the right-hand side presents the alternative options.



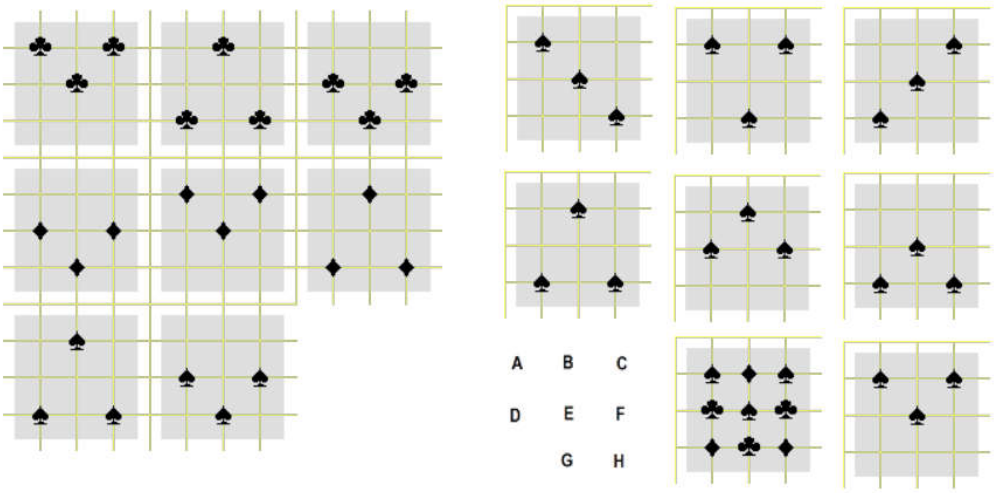
4.



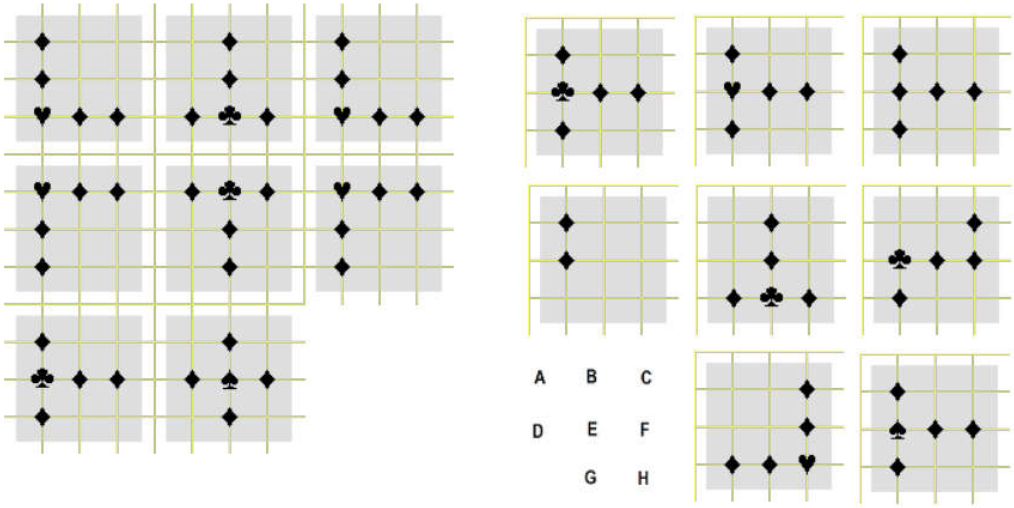
5.



8.



11.



Appendix 6.B

R Code for Implementing the Model-Building Approach-Progressive Matrices

```
setwd("C:/Users/Ruben/Desktop/Data for Example/IQ1/")
data<-read.table("data.csv",header=TRUE,sep=",")

library(mirt)
library(metafor)

data$Q1[data$Q1!=10]<-0
data$Q1[data$Q1==10]<-1

data$Q2[data$Q2!=10]<-0
data$Q2[data$Q2==10]<-1

data$Q3[data$Q3!=10]<-0
data$Q3[data$Q3==10]<-1

data$Q4[data$Q4!=10]<-0
data$Q4[data$Q4==10]<-1

data$Q5[data$Q5!=10]<-0
data$Q5[data$Q5==10]<-1

data$Q6[data$Q6!=10]<-0
data$Q6[data$Q6==10]<-1

data$Q7[data$Q7!=10]<-0
data$Q7[data$Q7==10]<-1

data$Q8[data$Q8!=10]<-0
data$Q8[data$Q8==10]<-1

data$Q9[data$Q9!=10]<-0
data$Q9[data$Q9==10]<-1

data$Q10[data$Q10!=10]<-0
data$Q10[data$Q10==10]<-1

data$Q11[data$Q11!=10]<-0
data$Q11[data$Q11==10]<-1

data$Q12[data$Q12!=10]<-0
data$Q12[data$Q12==10]<-1

data$Q13[data$Q13!=10]<-0
data$Q13[data$Q13==10]<-1

data$Q14[data$Q14!=10]<-0
```

```
data$Q14[data$Q14==10]<-1

data$Q15[data$Q15!=10]<-0
data$Q15[data$Q15==10]<-1

data$Q16[data$Q16!=10]<-0
data$Q16[data$Q16==10]<-1

data$Q17[data$Q17!=10]<-0
data$Q17[data$Q17==10]<-1

data$Q18[data$Q18!=10]<-0
data$Q18[data$Q18==10]<-1

data$Q19[data$Q19!=10]<-0
data$Q19[data$Q19==10]<-1

data$Q20[data$Q20!=10]<-0
data$Q20[data$Q20==10]<-1

data$Q21[data$Q21!=10]<-0
data$Q21[data$Q21==10]<-1

data$Q22[data$Q22!=10]<-0
data$Q22[data$Q22==10]<-1

data$Q23[data$Q23!=10]<-0
data$Q23[data$Q23==10]<-1

data$Q24[data$Q24!=10]<-0
data$Q24[data$Q24==10]<-1

data$Q25[data$Q25!=10]<-0
data$Q25[data$Q25==10]<-1

tdata<-data[,1:25]


#First we need to apply a 2PL model to the data
res<-mirt(tdata,1,'2PL')

#Now we obtain the Q3 values of our data.
#This will give us an mirt object with the Q3 values
resids<-residuals(res,type="Q3")

#Now we convert that object into a matrix format.
#Since we have 25 items we know that it will be a 25*25 matrix with 1's on the diagonal.
q3<-matrix(resids,ncol=25,nrow=25)

#Extract upper diagonal of matrix
newq3 <-matrix(c(q3[upper.tri(q3,diag=FALSE)]),nrow=1)
```

```
#Convert the upper diagonal of Q3 values into a vector for analysis
newq3 <-as.vector(newq3)

#Apply Fisher's r-to-z transformation to the Q3 values.
q3= .5*log((1+newq3)/(1-newq3))

#Compute variances for each effect size
k<-length(q3)
obs<-length(tdata[,1])
v<- rep(1/(obs-3),k)

#Create moderator that groups items 4, 12 and 13.
#LD is suspected in items 4, 12, and 13.
#The Q3 value related to item 4 and 12 is the 59th observation
#For item 4 and 13 it is the 70th observation and for item 12 and 13 it is the 78th observation
#The ES's for 59, 70, 78
mod<-rep(0,300)
mod[c(59,70,78)]<-1;

#Now that we have our r-to-z transformed Q3 values, their respective variances and the moderator that
#models the dependent relationship between our 3 items we can estimate the model using the rma()
function2
ld<-rma(q3~mod,v)
#Insepct results.
summary(ld)

#RANDOMLY SELECTED ITEMS

#11, 7, 5 Items
#ES: 20, 50, 52.
tmod<-rep(0,300)
tmod[c(20,50,52)]<-1
ld<-rma(q3~tmod,v)
summary(ld)
```

Appendix 6.C

R Code for Implementing the Model-Building Approach-NPI

```
setwd("~/Data for Example/NPI/")
data<-read.table("data.csv",header=TRUE,sep=",")
library(mirt)
library(metafor)

data$Q1[data$Q1!=1]<-0
data$Q1[data$Q1==1]<-1

data$Q2[data$Q2!=1]<-0
data$Q2[data$Q2==1]<-1

data$Q3[data$Q3!=1]<-0
data$Q3[data$Q3==1]<-1

data$Q4[data$Q4!=2]<-0
data$Q4[data$Q4==2]<-1

data$Q5[data$Q5!=2]<-0
data$Q5[data$Q5==2]<-1

data$Q6[data$Q6!=1]<-0
data$Q6[data$Q6==1]<-1

data$Q7[data$Q7!=2]<-0
data$Q7[data$Q7==2]<-1

data$Q8[data$Q8!=1]<-0
data$Q8[data$Q8==1]<-1

data$Q9[data$Q9!=2]<-0
data$Q9[data$Q9==2]<-1

data$Q10[data$Q10!=2]<-0
data$Q10[data$Q10==2]<-1

data$Q11[data$Q11!=1]<-0
data$Q11[data$Q11==1]<-1

data$Q12[data$Q12!=1]<-0
data$Q12[data$Q12==1]<-1

data$Q13[data$Q13!=1]<-0
data$Q13[data$Q13==1]<-1

data$Q14[data$Q14!=1]<-0
data$Q14[data$Q14==1]<-1
```

```
data$Q15[data$Q15!=2]<-0
data$Q15[data$Q15==2]<-1

data$Q16[data$Q16!=1]<-0
data$Q16[data$Q16==1]<-1

data$Q17[data$Q17!=2]<-0
data$Q17[data$Q17==2]<-1

data$Q18[data$Q18!=2]<-0
data$Q18[data$Q18==2]<-1

data$Q19[data$Q19!=2]<-0
data$Q19[data$Q19==2]<-1

data$Q20[data$Q20!=2]<-0
data$Q20[data$Q20==2]<-1

data$Q21[data$Q21!=1]<-0
data$Q21[data$Q21==1]<-1

data$Q22[data$Q22!=2]<-0
data$Q22[data$Q22==2]<-1

data$Q23[data$Q23!=2]<-0
data$Q23[data$Q23==2]<-1

data$Q24[data$Q24!=1]<-0
data$Q24[data$Q24==1]<-1

data$Q25[data$Q25!=1]<-0
data$Q25[data$Q25==1]<-1

data$Q26[data$Q26!=2]<-0
data$Q26[data$Q26==2]<-1

data$Q27[data$Q27!=1]<-0
data$Q27[data$Q27==1]<-1

data$Q28[data$Q28!=2]<-0
data$Q28[data$Q28==2]<-1

data$Q29[data$Q29!=1]<-0
data$Q29[data$Q29==1]<-1

data$Q30[data$Q30!=1]<-0
data$Q30[data$Q30==1]<-1

data$Q31[data$Q31!=1]<-0
data$Q31[data$Q31==1]<-1
```

```
data$Q32[data$Q32!=2]<-0
data$Q32[data$Q32==2]<-1

data$Q33[data$Q33!=1]<-0
data$Q33[data$Q33==1]<-1

data$Q34[data$Q34!=1]<-0
data$Q34[data$Q34==1]<-1

data$Q35[data$Q35!=2]<-0
data$Q35[data$Q35==2]<-1

data$Q36[data$Q36!=1]<-0
data$Q36[data$Q36==1]<-1

data$Q37[data$Q37!=1]<-0
data$Q37[data$Q37==1]<-1

data$Q38[data$Q38!=1]<-0
data$Q38[data$Q38==1]<-1

data$Q39[data$Q39!=1]<-0
data$Q39[data$Q39==1]<-1

data$Q40[data$Q40!=2]<-0
data$Q40[data$Q40==2]<-1

tdata<-data[,2:41]

res<-mirt(tdata,1,'2PL')

resids<-residuals(res,type="Q3")
obs<-length(tdata[,1])

q3<-matrix(resids,ncol=40,nrow=40)
#Extract upper diagonal of matrix
newq3 <-matrix(c(q3[upper.tri(q3,diag=FALSE)]),nrow=1)
#Expand obtained Q3 values into vector for analysis
newq3 <-as.vector(newq3)
#Apply Fisher’s r-to-z transformation
q3= .5*log((1+newq3)/(1-newq3))
k<-length(q3)
#Compute variances for each effect size
v<- rep(1/(obs-3),k)

#Create moderator that groups the first 3 items
#LD is suspected in items 4, 12, and 13. To model that we need to flag:
#The ES's for 59, 70, 78
mod1<-rep(0,780)
mod2<-rep(0,780)
```

```
mod3<-rep(0,780)
mod4<-rep(0,780)
mod5<-rep(0,780)
mod6<-rep(0,780)
mod7<-rep(0,780)

mod1[c(11,37,46,56,466,497,596,44,53,63,473,504,
603,55,65,475,506,605,66,476,507,606,477,508,607,527,626,628)]<-1
mod2[c(83,141,258,281,330,150,267,290,339,271,294,343,300,349,350)]<-1
mod3[c(32,304,634,745,309,639,750,656,767,778)]<-1
mod4[c(168,393,397)]<-1
mod5[c(207,227,452,545,720,231,456,549,724,428,550,725,559,734,737)]<-1
mod6[c(3,17,173,353,408,668,18,174,354,409,669,
178,358,413,673,371,426,686,434,694,696)]<-1
mod7[c(72,111,177,237,567,118,184,244,574,187,247,577,251,581,584)]<-1

mods<-cbind(mod1,mod2,mod3,mod4,mod5,mod6)
#Estimate the model
ld<-rma(q3~mods,v)
#Inspect results.
summary(ld)
```