Title
Essays in psychometrics and behavioral statistics

Author
Gochyyev, Perman

Publication Date
2015

Peer reviewed|Thesis/dissertation

Essays in psychometrics and behavioral statistics

by

Perman Gochyyev

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Education
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Associate Professor Alan E Hubbard

Fall 2015

Essays in psychometrics and behavioral statistics

Perman Gochyyev

Abstract

Essays in psychometrics and behavioral statistics

by

Perman Gochyyev

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This dissertation consists of three chapters. The main focus of the first chapter is on Lord's paradox. Lord's paradox arises from the conflicting inferences obtained from two alternative approaches that are typically used in evaluating the treatment effect using a pre-post test design. The chapter is designed as a guide to researchers who are using this research design. As an example, I investigate whether the treatment—a new mathematics curriculum—had an effect on student-level outcomes using both approaches. I demonstrate that Lord's paradox can occur even when the two approaches are accounting for the measurement error in variables.

Ordinal response data obtained from surveys and tests are often modeled using cumulative, adjacent-category, or continuation-ratio logit link functions. Instead of using one of these specifically designed procedures for each of these formulations of logits, we can modify the structure of the data in such a way that methods designed for dichotomous outcomes (i.e., binary logistic regression) allow us to achieve the targeted polytomous contrasting (cumulative, adjacent-category, or continuation-ratio). Thus, one can implement procedures designed for dichotomous outcomes on appropriately expanded data. The techniques presented in the second chapter, which I refer to as *data expansion techniques*, represent this approach.

The third chapter aims to contribute to the estimation and interpretation of multidimensional item response theory (MIRT) models within the field of psychometrics and latent variable modeling. The main goal of the chapter is to advance the use of the second-order Rasch model. A second-order Rasch model assumes an overall dimension as a second order factor that explains the covariance between the first-order (component) dimensions. The main contribution of the chapter is to suggest ways of using the model by still preserving the advantages of the Rasch model. Historically, the main challenge in the use of such models were (1) computationally intensive estimation and (2) availability

of software. In addition, it is difficult to obtain reliable and meaningful estimates in cases when a variance of one of the dimensions is low relative to other dimensions. In such cases, one first needs to re-assess if the multidimensional structure is appropriate. One, then, can use alternative parameterization of the model to avoid difficulties in the estimation, and guidelines in this chapter provide recommendations on how to achieve such parameterizations with the Rasch model.

To my parents and teachers.

Contents

List of Figures

List of Tables

Acknowledgements

I will mention a small subset of many wonderful people who have walked alongside.

First and foremost, enormous gratitude is due to Mark Wilson for his mentorship during the last six years. I am deeply grateful for his insightful feedback and for providing me the necessary intellectual trigger. I thank Sophia Rabe-Hesketh for her counsel, careful attention to detail, valuable feedback on my work, and for teaching me how to teach. I thank Alan Hubbard for his instruction and critical comments on my work. I thank Anders Skrondal for his help to understand my work in a broader statistical framework. I thank Kathleen Scalise for her support during the last six years.

Many colleagues and friends have contributed their knowledge, thought, and time to my studies, and I would like to thank David Torres Irribarra, Ronli Diakow, Andrew Galpern, Seth Corrigan, In-Hee Choi, Sira Park, and Alasdair Cohen.

This dissertation would have been impossible without funding provided through the Berkeley Evaluation and Assessment Research Center and opportunities to work on a variety of projects.

I thank my mother for instilling in me a love of learning, and my father for raising me to be strong and independent. I thank my mother-in-law for believing in me, and my father-in-law for advices on how to survive a graduate school.

Most importantly, I thank my wife Ayna and my children Jeren, Omar, and Osman, for their love, encouragement, patience, and understanding (I promise more time with you now).

Chapter 1

Lord's paradox

The main research question of this chapter is to compare methods to determine whether a treatment—say, a new mathematics curriculum—had an effect on student-level outcomes. We will assume the data were collected from cluster-randomized trials, wherein schools from districts were assigned to treatment or control at random within each district and that pretests and posttests were administered to students before and after the treatment. Two main approaches for analyzing such data are: (1) to regress the change from pretest to posttest on the treatment indicator; (2) to regress posttest on treatment indicator and pretest. Yet, these two approaches can yield conflicting results. Lord (1967) warned of this problem decades ago and started a debate that continues to the present day. In this chapter, I elaborate on both of these approaches and examine the appropriateness of each for analyzing the treatment effect and discuss these two approaches from the latent variable and multilevel modeling perspectives, after first discussing the apparent paradox at the heart of this issue.

1.1.1 Introduction

Frederick Lord wrote a two-page note (Lord, 1967)[1], in which he described two hypothetical statisticians who used different but seemingly equally valid methods to analyze the same data on treatment effects but arrived at contradictory conclusions. The problem he postulated, dubbed as "Lord's paradox", has yet to be resolved (see for instance Pearl, 2014; van Breukelen, 2013; London & Wright, 2012). In Lord's own words (Lord, 1967) the context is:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded. (p.304)

Two statisticians independently analyze the data at the end of the year by dividing the students according to gender. Statistician One examines gains in student body weight between girls and boys and finds no significant changes between the beginning and end of the year. Statistician Two adjusts for students' initial weights, finding that the regression coefficient of the initial weight is identical in both genders, and then finding a significant difference in the intercepts between boys' and girls' end-of-year weights. The question is, which of the two statisticians should the university administration listen to when making a decision?

Allison (1990) provides a real data example of Lord's paradox. The treatment group in the quasi-experimental study consisted of 18 children who received plastic

[1] Lord (1967) is cited in all papers discussing the paradox. However Lord (1963) was his first work in which comparison of the two approaches was presented formally. Feldt (1958) also presented a similar comparison but with a different focus.

surgery for craniofacial abnormalities. The control group consisted of 30 children of approximately the same age range. None of the children in the control group had or needed any surgery. The frequency of negative social encounters, which was the dependent variable, was measured shortly before the treatment (pretest) and 18 months later (posttest). Means for treatment and control group on pretest and posttest are shown in Table 1.1.

Table 1.1. Example from Allison, 1990.

	Frequency of negative social encounters	
	Pretest	Posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

The approach of Statistician One (regressing change from pretest to posttest on treatment indicator) yielded a coefficient of treatment indicator very close to zero—not a surprising conclusion when we look at Table 1.1. The approach of Statistician Two (regressing posttest on treatment indicator and pretest) yielded an estimated coefficient that was positive and significant at the $\alpha = 0.03$ level, indicating that the treatment had negative effect on children who had the surgery.

A quick look at Table 1.1 reveals that the means of the treatment group in pretest and posttest hardly changed, and the same for the control group. Why then does the approach taken by Statistician Two—the method that currently dominates social science methodology (Luecken & Tanaka, 2012, p. 264)—give an unintuitive and misleading result?[2]

Lord correctly noted in his 1967 note that "there are as many different explanations as there are explainers", and concluded: " ... there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups."[3](p. 305)

Lord postulated similar question in Lord (1969) and mentioned that the substance of this paradox is being downplayed by some, who argue that simple gains analysis should be the appropriate approach for such a problem. In response, he provided an identical example but with different scales of pre and post measures (GPA as pre and IQ scores as post), to demonstrate that "gains" approach should not come as a default approach and thus to emphasize the validity of the ANCOVA approach in such problems. Lord (1973) also posed a similar problem with a different example[4].

Debates over the paradox postulated by Lord attracted a great deal of attention (and controversy[5]) in behavioral statistics, biostatistics and other related fields dealing

[2] Also, the method used by Statistician Two may lead to the conclusion that there is no treatment effect when means of group at two different timepoints indicate otherwise.

[3] Lord's earlier work focuses on measurement of growth and gains, and propensity of false conclusions in such analyses—see for instance Lord (1956, 1958). He was mostly concerned with issues related scale in measuring growth, a topic I address later in the chapter.

[4] All three of Lord's examples are thoroughly discussed in Holland & Rubin (1983).

[5] Citing Lord (1967), Senn (2008, p.106) wrote: "In a disturbing paper in the Psychological Bulletin in 1967, Lord considered a case where two statisticians analysing a data set come to radically different conclusions.

with methodological issues. Debates spread mainly into three related, yet different, directions. One such direction, driven by psychometricians, argued over the reliability of change from pretest to posttest and issues related to the regression to the mean (Cronbach & Furby, 1970; Linn & Slinde, 1977, Rogosa & Willett, 1983; Zimmerman & Williams, 1982; Willlett, 1988; Collins, 1996; Willett, 1997; Mellenbergh & van den Brink, 1998[6]).

Another direction is related to the issue of ill-defined research questions and contrasts among competing causal frameworks equipped differently to resolve the paradox[7] (Holland & Rubin, 1982; Rubin, 1974; 1977, Pearl, 2014; Wainer & Brown, 2007).

A third direction and corresponding work, which represents the majority in the social science methodology and which attempts to provide suggestions to researchers in the field, resulted in the answer: "it depends" (Kenny,1975; Bryk & Weisberg, 1977; Maris, 1998; Porter & Raudenbush, 1987; Rausch, Maxwell, & Kelley, 2003; Reichardt, 1979; Senn, 2006; van Breukelen, 2013; Wainer, 1991; Weisberg, 1979; Arah, 2008; Wright, 2006).

Among these, the work that belongs to the third direction and that stands out as the most successful resolution of the paradox is, in my opinion, Allison (1990). However, the arguments and derivations in Allison (1990) would benefit from further articulation and elucidation, which is one of the aims of this chapter.

The literature and central arguments related to all of these approaches to resolving the paradox will be presented in this chapter. It is worthwhile to mention that, in many of the papers comparing approaches of the two statisticians (including Lord's (1967)), the approach of Statistician One (regressing change on treatment indicator), explicitly or implicitly, is the underdog. It is important to clarify the approaches of two statisticians, which will be the aim of the next section.

1.1.2 Lord's paradox (or not)

The approach taken by Statistician Two is generally known as "ANCOVA" in the experimental design literature. In structural equation and graphical modeling frameworks, the estimate based on this approach is also known as a "direct effect mediated by the pretest" (pretest is the mediator). In econometrics, the approach of Statistician Two is referred to as the "lagged dependent variable" approach, an approach mainly used in dynamic models.

In contrast, the estimate obtained by Statistician One is known as "Total Effect" in the graphical modeling literature. This approach is referred to as the "ANOVA" (or "RANOVA" [repeated measures ANOVA]) approach in experimental design literature and the "first-differencing" (FD) estimator in the econometrics literature[8]. In statistical

---

[6] Mellenbergh & van den Brink (1998) investigated the gain approach using CTT and binomial error model for single-subject change.

[7] There have been, and still are, disagreements on the research question the two hypothetical statisticians in Lord (1967) were trying to answer.

[8] The question of whether using differencing or lagged dependent variables in econometrics literature comes up as an issue similar to the one I will discuss in Section 1.1.5. In particular, it is considered necessary to use a lagged-dependent variable when the omitted variable bias might arise from time-varying variables. For instance, a subsidized training program might be correlated with the past income of

modeling language, the approach of Statistician One represents an "unconditional" model whereas the approach of Statistician Two represents a "conditional" model[9]. In order to provide further clarity then, throughout my paper, I will refer to the approach taken by the Statistician One as the Change Score (CS) [10] approach and the approach taken by Statistician Two as the Regressor Variable (RV) approach.

The RV approach has strong support in the classical statistical literature provided by Fisher (1951, chapter 9) and Cox (1958, section 4.4). Temporal order is the basis for using this model for inference. Most agree that the CS approach, however, is simpler in its interpretation. This approach was first formally introduced in John Snow's cholera study (Snow, 1855).

Snow showed that cholera was a water-borne disease by investigating the difference in changes of death rate between two districts serviced by two different water companies: Southwark & Vauxhall Company and the Lambeth Company. The Lambeth company changed its water supply—moved to a cleaner location (i.e., the "intervention")—and as a result death rates were reduced in districts serviced by Lambeth (the treatment group). Had the Southwark & Vauxhall company moved their water supply together with the Lambeth company, Snow concluded, approximately 1000 lives would have been saved.[11]

Figure 1.1 illustrates Lord's paradox as it arises from comparison of the two groups. Groups A (treatment) and B (control) are different at the pretest (x-axis). The same difference is retained in the posttest (y-axis). The bold 45° line represents the CS approach, leading to the interpretation that there is no difference between the groups, hence the single bold line. The dashed line (with a slope of 0.5, the coefficient of the pretest) represents the RV approach. The difference between the two dashed lines is the estimate of the size of the regression coefficient for the group indicator obtained from the RV approach leading to the interpretation that there is a difference between the two groups. When the two groups differ on the pretest, the CS and the RV approaches will generally give different results.

---

participants—for example, those whose income decreased perhaps want to increase their labor market options and enroll in the treatment program.

[9] Bock (1975, p. 490), categorizes these two approaches as "unconditional" vs. "conditional" inferences.

[10] Unlike many authors, I prefer using change score—using "gain" implies positive change while change can also be negative.

[11] See (Snow 1965 [1855]) and Freedman (2010) for more details.

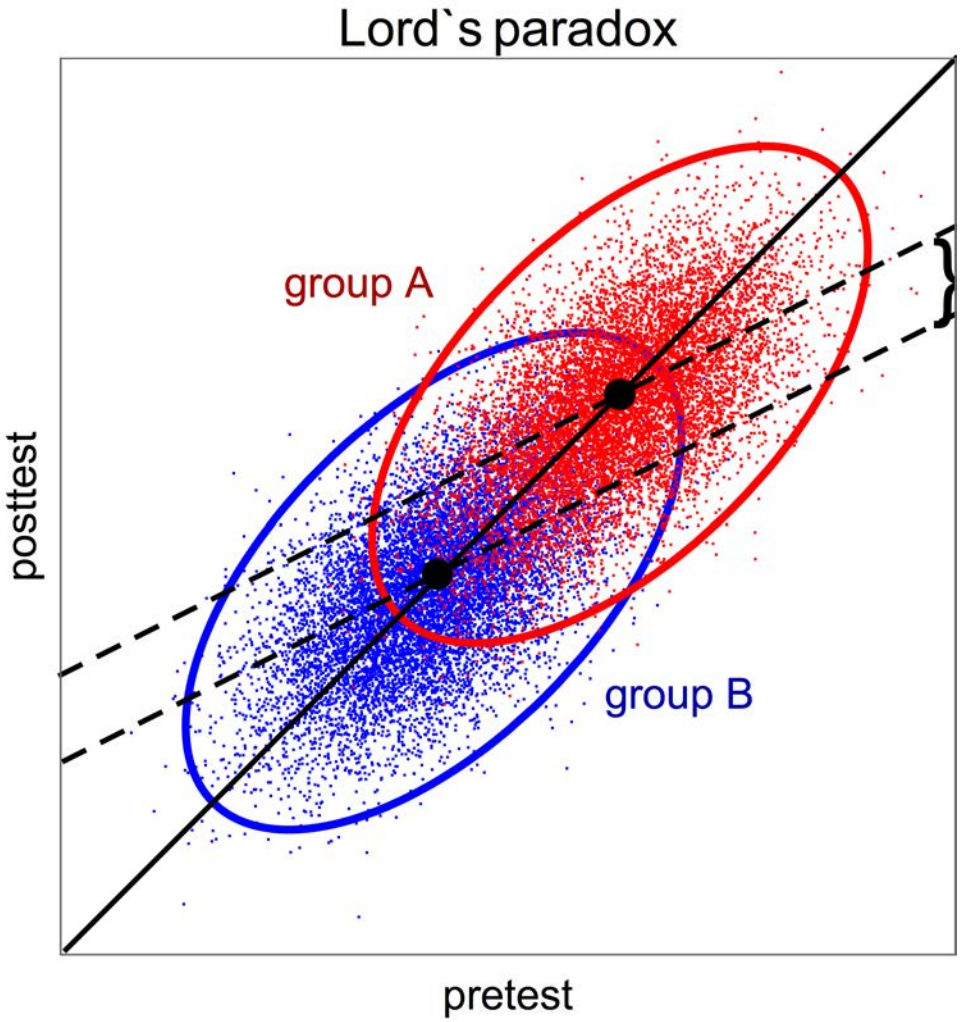


Figure 1.1. Lord's paradox: Difference between the group means in the pretest is same as the difference in posttest. The RV approach (dashed lines) and the CS approach (bold line).

In contrast, Figure 1.2 below shows the scenario we would see when the treatment is assigned at random (i.e., no difference at pretest). In this case, there is no difference between the two groups on pretest and both approaches (CS and RV) give the similar answers (no paradox).

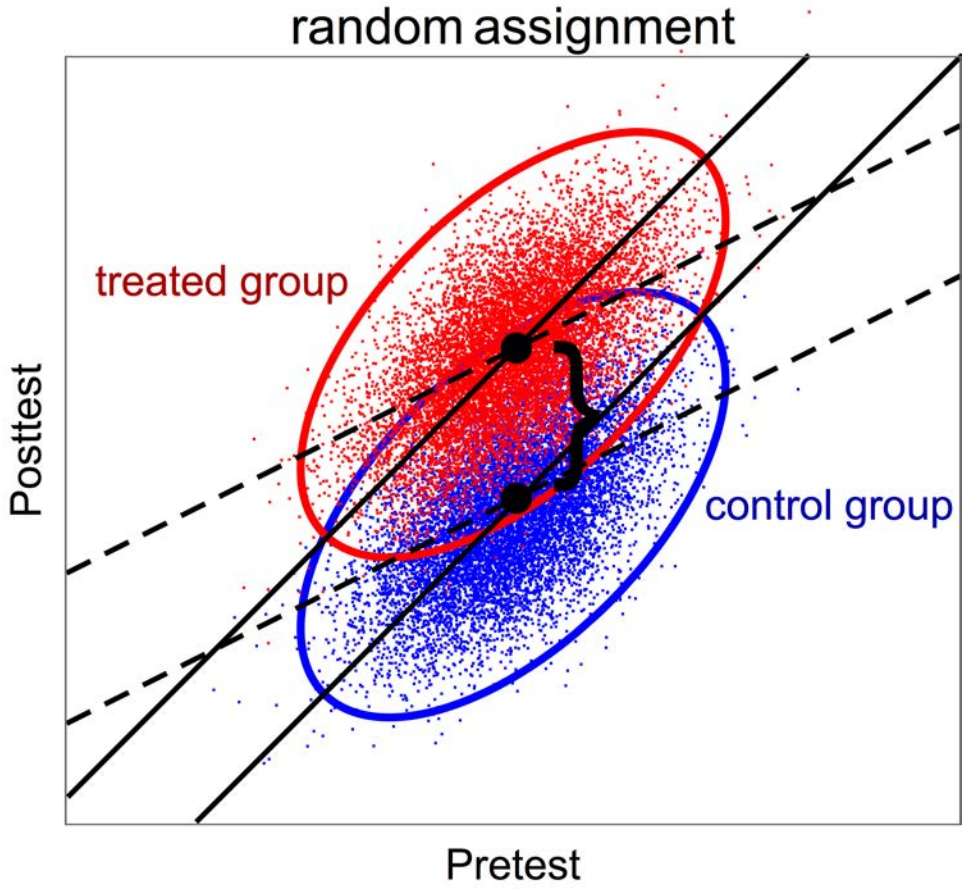


Figure 1.2. Bivariate distributions for two groups when treatment is assigned at random: CS and RV approaches yield the same results in the absence of difference on pretest

1.1.3 RV approach

Assume a two-wave design with only two snapshots of the construct: pretest and posttest. Throughout the paper, I will refer to pretest as Y_{1j} and to the posttest as Y_{2j} for the person j. Also assume that W_j is a binary variable taking value 1 if person j is in the treatment group and 0 if in the control group. The RV approach is written as:

$$Y_{2j} = \beta_1 + \beta_2 W_j + \beta_3 Y_{1j} + \epsilon_j \quad \textbf{(1)}$$

in which β_1 is the intercept, β_2 is the difference between treatment and control groups controlling for the pretest, β_3 is the effect of pretest conditional on the group indicator, and ϵ_j residual term. To obtain unbiased estimates of coefficients using Ordinary Least Squares (OLS) we assume that residuals ϵ_j are uncorrelated with Y_{1j} and W_j. In other words, the RV approach assumes that factors that are not in the model and thus absorbed by the residual term, influence the posttest but not the pretest. Assuming regularity conditions, OLS is consistent and unbiased. Also note that—even though it is not explicitly laid out as a model assumption—the implicit assumption is that Y_{1j} and Y_{2j} are measured without error. I consider this assumption closely in Section 1.1.8.

Equation 1 can also be rewritten to resemble the CS approach (presented below in Equation 6). In particular, one can rewrite the RV approach as regressing the change from

pretest to posttest on the treatment indicator and pretest to obtain:

$$Y_{2j} - Y_{1j} = \beta_1 + \beta_2 W_j + (\beta_3 - 1)Y_{1j} + \epsilon_j. \quad \textbf{(2)}$$

Estimates from the Equation 1 and Equation 2 are identical since these two are algebraically identical.

1.1.4 CS approach

For ease of interpretation and derivation that follow later, I will present the CS approach from two alternative perspectives. Both perspectives result in the identical model but emphasize different assumptions. First, I present derivations following Allison (1990), which are also related to arguments presented in Kenny (1975), Chamberlain (1984), and Heckman & Robb (1985). The second formulation is from a traditional approach to panel data usually found in econometrics textbooks (see for instance Wooldridge, 2002) or books on longitudinal modeling (see for instance Rabe-Hesketh & Skrondal, 2012).

Assume G_j is a binary variable indicating the treatment status: 1 if person j ends up in the treatment group and 0 otherwise. The pretest score (Y_{1j}) can be expressed as:

$$Y_{1j} = \beta_1 + \delta G_j + \epsilon_{1j}. \quad \textbf{(3)}$$

Coefficient δ represents group differences between treatment and control group that are stable ("preexisting" differences). When treatment is assigned at random, we would expect no preexisting differences between two groups (see Figure 1.2). The posttest score (Y_{2j}) is expressed as

$$Y_{2j} = \beta_1 + \tau + \delta G_j + \beta_2 W_j + \epsilon_{2j}, \quad \textbf{(4)}$$

in which W_j is the treatment indicator. Note that for Equation 4, $G_j = W_j$ and thus these are perfectly collinear and a solution cannot be estimated in the current form since the rank conditions of OLS are not satisfied. τ represents the change that is occurring in both groups (e.g., gained knowledge during school-year). By subtracting Equation 3 from Equation 4 we obtain:

$$Y_{2j} - Y_{1j} = (\beta_1 - \beta_1) + \tau + (\delta G_j - \delta G_j) + \beta_2 W_j + (\epsilon_{2j} - \epsilon_{1j}). \quad \textbf{(5)}$$

Equation 5 can be expressed as:

$$\Delta Y_j = Y_{2j} - Y_{1j} = \tau + \beta_2 W_j + \epsilon_j^{\Delta}, \quad \textbf{(6)}$$

in which ϵ_j^{Δ} is $\epsilon_{2j} - \epsilon_{1j}$. Assuming ϵ_j^{Δ} is not correlated with W_j, OLS is consistent and hence the estimates are unbiased.

Thus, the only difference between the CS approach and the RV approach is that the latter is conditional on the pretest. In most applications, the correlation between the pretest and the posttest is between 0 and 1, and the correlation between the pretest and the gain is between -1 and 0[12].

[12] The coefficient of pretest in the RV approach should be below unity for dynamic stability, otherwise the process becomes a so called explosive time series (e.g., Y_{ij} is nonstationary).

It is important to emphasize that by differencing (Equation 5), we eliminated all unobservable and observable time-invariant factors (including cluster effects such as the effects of school membership on the units of analysis [students]). Thus, there is no need for assumptions concerning the mean and variance of these effects (as in random-effects models) and these unobserved variables are allowed to have arbitrary correlation with covariates in the model since they will be eliminated from the model after differencing. Thus, as long as there is no endogeneity arising from time-varying factors, we are controlling for all stable unobserved variables.

Since our inference is based on intra-individual variation, the CS approach uses each individual as its own control. Even though this might not be an efficient approach (compared to random-effects approaches that use between-individual variability), this approach reduces bias (Allison, 2009). If we have time-varying covariates that are observed and need to be included in the specified model, we can easily control for these by including them directly in the equation.

We can rewrite Equation 6 by moving Y_{1j} to the left-hand side and restricting its coefficient to unity to obtain:

$$Y_{2j} = \tau\ + \beta_2 W_j + (1)Y_{1j} + \epsilon_j^{\Delta}. \quad \textbf{(7)}$$

Equations 6 and 7 are identical. We might be tempted to conclude that the RV approach shown in Equation 1 then shares the nice properties of the CS approach since Equation 7 seems merely a special case of Equation 1. In particular, if we free the coefficient of the pretest in Equation 7 and rewrite it as:

$$Y_{2j} = \tau\ + \beta_2 W_j + \beta_3 Y_{1j} + \epsilon_j^{\Delta}, \quad \textbf{(8)}$$

we obtain an equation similar to that of Equation 1. Thus, Equation 8 seems like a special case of Equation 7 with the only difference being that the regression coefficient of the pretest in Equation 7 is constrained to unity. In fact, this is what has been argued by many (see for instance Gelman & Hill, 2007, p. 177; Hedeker & Gibbons, 2006, p. 8; van Breukelen, 2013, p. 903) and it has been argued that the CS approach unnecessarily constrains this coefficient to unity (Gelman & Hill, 2007).

However, this is not an appropriate way to compare the two models: note that in Equation 8, the pretest Y_{1j} is negatively correlated with ϵ_j^{Δ} (by construction), which can be seen from Equation 6. Due to this negative correlation, Equation 8 cannot be estimated consistently by OLS, and an OLS estimate would be biased.

A further derivation[13] of the bias is presented in Section 1.1.9, but it is important to stress upfront that the CS approach is not a special case of the RV approach—rather, the two approaches represent two completely different models! Overlooking this crucial distinction has been the most common error in past comparisons of the two approaches, and any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading.

In Section 1.1.8, I show that the assumption that ΔY_j is a measurement error-free measure of change is not necessary to obtain unbiased estimates from OLS. However,

[13] I will show that, if the CS model is the correct model, then OLS using the RV approach will always give biased estimates of the coefficients. This derivation, first presented in Allison (1990) is the most important piece of the resolution of the paradox and deserves elaborate discussion.

this will not alter the fundamental point made above—it will only influence the variance of the ϵ_j^{Δ} term and hence reduce power.

Another assumption, which is intuitive but perhaps needs to be stated explicitly, is that Y_{1j} and Y_{2j} are assumed to be on a common metric (e.g., such as an interval scale). Cross-time linkage of pretest and posttest measured by different sets of items is established by anchoring item intercepts to the estimated parameters obtained from a data set that contains both set of items. Strong factorial measurement invariance (Millsap, 2011) is achieved by having all item loadings set to unity with the Rasch model. The cross-time linkage method using the partial credit model (Masters, 1982) and the cumulative Rasch model (Agresti & Lang, 1993) is discussed in Section 1.2.3.

Next, I present the CS approach using the derivation commonly used in panel data modeling.

Assume W_{tj} is a binary variable taking value 1 if person j is in the treatment group at time t and 0 otherwise. Assume T is the linear time trend taking the value of t (e.g., a dummy indicator for the posttest). The response of person j at time t can be expressed as:

$$y_{tj} = \beta_1 + \alpha T + \beta_2 W_{tj} + \zeta_j + \epsilon_{tj}, \quad \textbf{(9)}$$

in which ζ_j represents the time-invariant effect of person j. At time point one (pretest), $W_{j1} = 0$ for all j (nobody receives the treatment). At the posttest, $W_{j2} = 1$ only for subjects in the treatment group. First-differencing will result in:

$$y_{2j} - y_{1j} = \alpha + \beta_2(W_{2j} - W_{1j}) + (\zeta_j - \zeta_j) + (\epsilon_{2j} - \epsilon_{1j}), \quad \textbf{(10)}$$

thus

$$\Delta Y_j = \alpha + \beta_2 W_j^{\Delta} + \epsilon_j^{\Delta}, W_j^{\Delta} = W_j = W_{2j} - W_{1j}. \quad \textbf{(11)}$$

Note that β_2 in Equation 11 is the difference-in-difference estimator[14] since

$$\beta_2 = \Delta \bar{y}_{W=1} - \Delta \bar{y}_{W=0} \quad \textbf{(12)}$$

in which $\Delta \bar{y}_{W=1}$ represents a mean change from pretest to posttest in the treatment group and $\Delta \bar{y}_{W=0}$ represents a mean change from pretest to posttest in the control group. We assume:

$$E\left[\epsilon_j^{\Delta} \middle| W_j^{\Delta}\right] = E\left[\epsilon_j^{\Delta} \middle| W_{2j}\right] = E\left[\epsilon_{2j} \middle| W_{2j}\right] - E\left[\epsilon_{1j} \middle| W_{2j}\right] = 0, \quad \textbf{(13)}$$

which implies strict exogeneity. In other words, we assume that treatment indicator at time point two ($W_j^{\Delta} = W_{2j}$) is uncorrelated with residuals at any time point. Assuming serial correlations in Equation 11 is too strong and not necessary for the consistency of OLS.

[14] Often, the difference-in-difference estimator is also used when groups in pretest and posttest are representative random samples (Wooldridge, 2002). Then, coefficient of interest is the interaction between treatment and time indicator. However, in the estimator presented in Equation 11, the two scores are obtained for the same persons. This difference-in-difference estimator is identical to the fixed-effects estimator in the two-wave panel design.

1.1.5 Choosing an appropriate approach

Kenny (2011; 1975), argues that one needs to consider what exactly might influence the change or the post score. In particular, he proposes that the pretest needs to be thought to be composed of three components: (1) the component that is stable—the population characteristic or the permanent, time-invariant characteristic; (2) the component that is the one that changes over time—time-varying, transitory component; and (3) the component that is the result of random variation, random error that also varies over time[15].

It is important to articulate on which of these three components the selection into the treatment groups is determined. Kenny (1975) argues that if it is the stable component, CS approach is justifiable approach: if the selection to treatment is related to the permanent component then the CS approach accounts for it since this stable component will present in both pretest and posttest. If, however, the selection into the treatment depends on the transitory component, then the RV approach is the only remedy. When the treatment is randomized, it does not depend on any of these components and thus both approaches are expected to yield the same result.

Arguments in Allison (1990) are related to the approach taken in Kenny (1975), in which he argues that CS approach is superior to the RV approach when the treatment indicator is uncorrelated with the transient component of the pretest (Allison, 1990). Allison (1990) also argues that when effects of other variables on the outcome variable are invariant from pretest to posttest (e.g., gender or cluster-level random effect in random intercept model), we don't need to control for such variables in the CS approach[16].

In the RV approach, the intention to use the pretest (particularly in the non-equivalent control group design) is to adjust for the prior differences, but it underadjusts for these preexisting differences (Allison, 1990)[17]. In Table 1.1, the fact that pretest difference between two groups is similar to the posttest differences between two groups is not fully accounted by the inclusion of the pretest using the RV approach. As a result, estimates from the RV approach can be biased in cases in which we do not expect the pre-test measures to be similar (e.g., no randomization, self-selection). This is the scenario in which the residual term in Equation 8 is correlated with the independent variable in the model—pretest—to yield biased estimates. Thus, even though the gain approach will have higher variance, it should be preferred since it is unbiased.

Are there cases in which RV approach is unbiased and CS approach is biased?—we will discuss these further below. But it suffices to mention that if the selection to the groups is based, on or correlated with, the pretest measure, then the pretest becomes a confounder and definitely needs to be controlled. Regression discontinuity

---

[15] Crowder and Hand (1990, p. 25) referred to these three components as "immutable constant of the universe", "lasting characteristic of the individual", and "fleeting aberration of the moment".

[16] Unless there is an interaction of these invariant variables with the time, which implies a relationship of the measure at pretest with the transient component.

[17] We don't expect prior differences between groups in the randomized studies and use of posttest only is sufficient (no need to control for the pretest) for the causal statement. However, the pretest measure is still used in such design to increase power, by explaining more variance at posttest and reducing the residual variance (Moerbeek, Van Breukelen, & Berger, 2008).

(Thistlethwaite & Campbell, 1960), for instance, is an obvious example for a such case.

One important point in comparing the two approaches, often overlooked, is what Statisticians One and Two are assuming? Since they are employing different procedures, they are relying on different (generally untested) assumptions. Statistician One (CS approach) assumes that, if the treatment has no effect, two groups would show the similar gain. Statistician Two (RV approach) assumes that the total gain of the groups is the same as the gain within groups (Holland & Rubin, 1983; Allison 1990).

Holland & Rubin (1983) articulated on which assumptions are both statisticians are relying when drawing their conclusions and took an attempt to address the paradox using the potential outcomes framework. They noted that both statisticians' statements are descriptive in nature, and not causal: Statistician One makes an "unconditional descriptive statement" that average gains are equal for males and females. Statistician Two makes the conditional statement (conditional on pretest) that for males and females of equal pre-test score, the males gain more than females.

One critique by Holland & Rubin (1983) and by Rubin, Stuart, & Zanutto (2004) of the example presented in Lord (1967) was that Lord's example was a "poorly formulated causal assessment" (Rubin, et al., 2003) since the potential outcome under the "control" diet is missing. The difficulty in Lord's postulation of the problem, they wrote, is that there is no control group, and researcher investigating "gain" wouldn't know if changes in scores would have occurred with no treatment anyway.

However, in my opinion, attempts by Holland & Rubin (1983) to clarify the paradox move away from the research question at the center of Lord (1967). In other words, the paradox has been somewhat misrepresented with respect to the research question that Lord (1967) had in mind (see for instance Pearl, 2014). The hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet. This is precisely why Lord didn't mention about the control diet condition in his presentation of the paradox, and instead focused on the "differential effect" of the diet. In other words, the gender variable can be thought of as a treatment indicator in Lord (1967).

Under the Neyman-Rubin (a.k.a. potential outcomes) causal framework —the framework of Holland & Rubin (1983)—however, the effect of gender cannot be a causal research question. Gender, an immutable characteristic of the research unit, and hence cannot be a "cause"[18]. Wainer & Brown (2007) tried to resolve the paradox under the potential outcomes framework by replacing the gender variable by "dining tables serving the two genders". "Dining table", then, can be a causal effect under Neyman-Rubin causal framework since the condition is manipulatable (but still definitely confounded by gender).

To further clarify the approaches, below I present the two approaches (RV and CS) using the Directed Acyclic Graph (DAG) framework (Pearl, 2000)—a framework alternative to the Neyman-Rubin framework.

Figure 1.3 below represents both CS (the "total effect" in the DAG framework) and RV (the "direct effect" in the DAG framework) approaches. Y_1 and Y_2 are pretest and posttest respectively (I drop the j subscript for person for the simplicity of the presentation), W is the treatment indicator, and C is the change score.

[18] "No causation without manipulation". Holland (1986, p. 959).

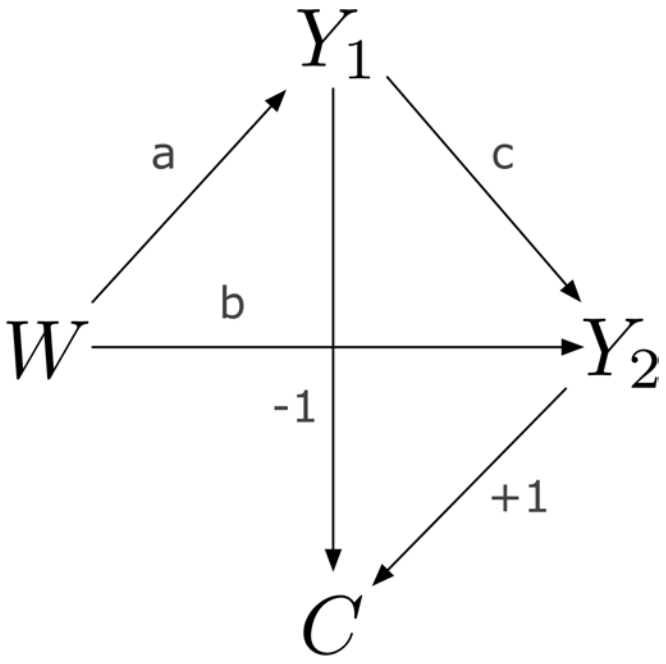


Figure 1.3. CS and RV approaches in the DAG framework (Pearl, 2014).

In Figure 1.3 above, Y_1 is the mediator between W and Y_2. Similar to the β_2 in Equation 12, the total effect can be estimated by:

$$TE = E[C|W = 1] - [C|W = 0]. \quad \textbf{(14)}$$

Using arrows in Figure 1.3, the expression in Equation 14 is identical to $(b + ac) - a$. To express the direct effect (RV approach)—β_2 in Equation 1—we obtain:

$$DE = \sum_Y (E[C|W = 1, Y_i = y] - E[C|W = 0, Y_i = y])P(Y_i = y|W = 0), \quad \textbf{(15)}$$

which is averaged across values of pretest and $(E[C|W = 1, Y_i = y] - E[C|W = 0, Y_i = y])$ is known as the "controlled direct effect" (appropriate when it is assumed that Y_i is uniform over the entire population). In the DAG framework, the last part of Equation 15, $P(Y_i = y|W = 0)$, is a weighing function (see for instance Pearl, 2009, p. 131). It sets the pretest for each person to the value it would have obtained before the treatment (when $W = 0$)[19] (i.e., "controlling" for the pretest). Estimate in the Equation 15 is also known as the "natural direct effect" and is shown with the arrow b in Figure 1.3.

Allison (1990) noted that RV approach is only preferable when the pretest score can bias the posttest score or when the treatment assignment and the pretest score are strongly related (e.g., regression discontinuity). In these cases, pretest is a "confounder" and needs to be controlled. In the DAG framework, this can be expressed by reversing one of the arrows (arrow "*a*"), as shown in Figure 1.4 below (Pearl, 2014).

[19] also see Pearl, 2014 and Morgan & Winship, 2007.

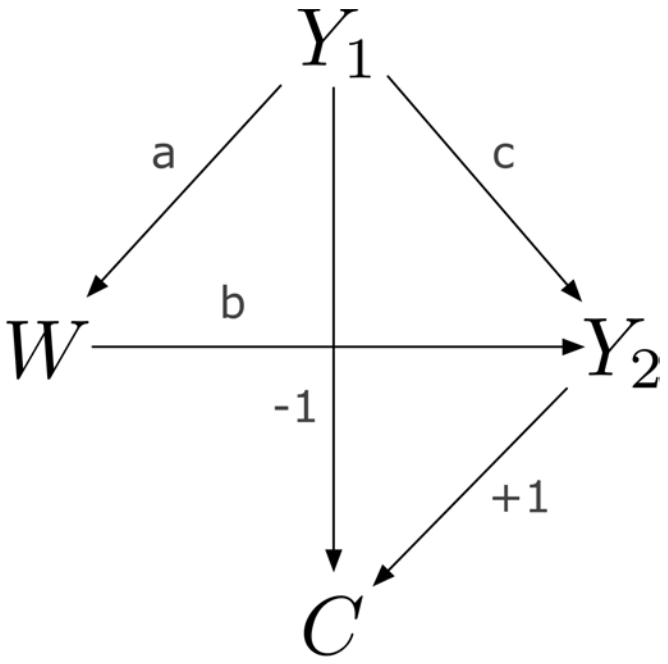


Figure 1.4. CS and RV approaches in the DAG framework when the pretest is a confounder (Pearl, 2014).

Novick (1983) and Lindley & Novick (1981), provide a global ("framework-free") approach and explanation for the paradox. They stress that the inference must be based on a careful specification of the relevant subpopulations involved—and especially assumption of exchangeability[20]. In particular, their general approach is based on De Finetti's exchangeability (de Finetti, 1964, 1972) or Fisher's subpopulation concept (Fisher, 1937): inference is conditional on the subpopulation the subject belongs to, and a careful specification of the relevant subpopulation is necessary.

Consider an extreme example to understand their argument: group A is a sample of mice and group B is a sample of elephants and one of the groups receives the treatment. Both approaches can technically provide conclusions, but the interpretation of the output from the RV approach will require a hypothetical population of mice and elephants with the same baseline (pretest) weights[21]. One needs to query, then, whether the samples in groups A and B (e.g., mice and elephants) are exchangeable. If not, the RV approach is not helpful at all. The CS approach is the only alternative, although this approach relies on the assumption that if the treatment has no effect, two groups would show the similar proportions of gain[22].

We assume that two groups are exchangeable if groups are "equivalent in all relevant respects" (Weisberg, 2010). Equivalence of the means or medians of two groups on relevant covariates is a limited form of exchangeability. As a more general definition,

[20] Novick (1983): "randomization is useful, but exchangeability through modeling, blocking, and covariation is fundamental" (p.47).

[21] One clear advantage of the Neyman-Rubin and DAG frameworks is that such comparison of mice and elephants will be dismissed due to the lack of "common support" (a.k.a. "positivity" in DAG framework) assumption.

[22] In this extreme example, we can think of "gain" as increase in percentages.

two study groups are exchangeable if and only if they have identical distributions of response patterns. This type of exchangeability is what Rosenbaum & Rubin (1983) referred as strong ignorability (conditional on a set of covariates). See McCullagh (2005) and Weisberg (2010) for a careful presentation of the concept of exchangeability.

There are multiple definitions of exchangeability[23], and therefore I am intentionally using the concept of exchangeability here very loosely. By exchangeability of the groups, I mean comparability of the groups. Careful definition of exchangeability is definitely a necessary part of the argument, but it is not the main focus of this chapter. With imbalance on the pretreatment covariates (e.g., significant difference in the proportion of Hispanic students when comparing schools), we have a basis to be suspicious that the comparability of the two groups can be questioned.

Inferences depend on assumptions. Assumptions require careful considerations of their plausibility. Arah (2008) notes that explanations and solutions of the Lord's paradox lie in causal reasoning (and background knowledge) and not on statistical criteria. Bryk & Weisberg (1977) also note that the resolution of such arguments requires careful analysis of processes generating the data. They caution against blind application of RV and similar methods ("conditional methods")—covariates entering the model without careful consideration—and warn that this will result in more model misspecification than no adjustment at all. Wooldridge (2005) provides a similar argument (and a short derivation) that the ignorability of the treatment (assumption of "as if random") is violated by adding too many covariates to control, which results in overcontrolling.

For the RV approach, we can condition on the covariates that are not balanced, but we are still assuming that there are no "unmeasured confounders" or that the covariate we are conditioning on is the only one causing the imbalance. For the CS approach, this latter assumption is less strict: we only assume that there are no "time-varying unmeasured confounders".

However, this doesn't mean that the CS approach is the default approach when the groups differ at pretreatment covariates. In addition to arguments I summarized in this section, in Section 1.1.7, I discuss one crucial point in choosing an appropriate approach—this criteria is related to the concept of exchangeability and to the concept of the regression to the mean**.**

1.1.6 Debates on reliability

Measurement of change became a favorite topic of psychometricians in 1970s. The use of CS approach has been criticized for years, and mainly was maligned through the 1960s and 1970s (Willett, 1997). Lord himself argued about unreliability of the change score (Lord, 1956). Cronbach & Furby (1970) end their paper with the following suggestion: "It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways." (p. 80)
On a similar cautionary note, Linn & Slinde (1977) noted that: "Problems in measuring change abound and the virtues in doing so are hard to find. Major disadvantages in the use of change scores are that they tend to conceal conceptual difficulties and they can give misleading results." (p. 147)

---

[23] … and different types of exchangeability (e.g., full exchangeability, partial exchangeability)—see Greenland & Robins (2009) and Weisberg (2010) for more on this.

Some argue that change scores represent an accumulated error: a combination of two "not-perfectly-reliable" terms, as they argue, becomes "even-less-reliable".This argument is based on Gulliksen (1950, p. 353) who provides a formula for the combination of two scores with errors. Gulliksen (1950) showed that errors accumulate when we combine two "not-perfectly-reliable" scores (e.g., pretest and posttest administered to students). In particular, reliability of the change score (from pretest, y_1, to posttest, y_2), is:

$$r_{y_2-y_1} = \frac{r_y - r_{y_1y_2}}{1 - r_{y_1y_2}}, \quad \textbf{(16)}$$

in which $r_{y_1y_2}$ is the correlation between the pretest and posttest and r_y is the average of two reliability coefficients.

Cattell (1982) and Gollwitzer et al. (2014) note that Gulliksen's formula assumes that reliability is stable and that two tests (pre and post) have the same variance[24]. In an educational setting, for instance, variation in abilities might be decreasing among students as they learn and thus variance might be decreasing (i.e., the treatment might be making students less heterogeneous). It can be easily shown from the formula provided by Gulliksen (1950), that the larger the difference in variances between the pre and the post measures, the higher the reliability of the difference score (Zimmerman & Williams, 1982; 1998). This point is often overlooked in attacks on the CS approach from that particular perspective. But, is this argument relevant at all—specifically for the inference between group differences?

However, this perspective can be challenged: if the "gain" is similar among the subjects, the "gain" will seem to have low reliability (since similar gains cannot be distinguished between persons). Collins (1996) pointed that in the CS approach, the focus is on the intra-individual variability, and there is nothing in the concept of reliability that addresses that. Thus, it would be possible to have zero reliability of the measure and still have a precise measure of the change.

From a similar perspective, as noted in Allison (1990), the ideal case for the treatment to have an effect is when the control group doesn't change and all subject in the treatment group change with the same amount. But this case will result in high correlation between pre and post, and thus lower change score reliability. Thus, low reliability of the change score is irrelevant for the causal inference. If we are interested in precision, the concern should be the error variance in the CS approach (Allison, 1990).

Similar confusion (and consequently another critique of the CS approach by psychometricians) arose from the belief that pre and post measures should correlate highly for the validity (construct validity) argument. This claim was dismissed later on (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983) and it was shown that the correlation between pre and post tests can even be negative while the test is perfectly valid. This is mainly due to the potential heterogeneity in the gains. The lower the test-retest reliability, the higher the gain score reliability.

Longitudinal design provides the foundation for the causal inference because we know three things about causality: (1) covariation, (2) temporal precedence (time order of

[24] In Lord's hypothetical example, mean and standard deviation of pre and post are identical in the treatment group. The same is true for the control group.

cause-effect), and (3) elimination of other causal factors (Viswanathan, 2005). Two-wave design (pre and post) is the least informative method to measure change. The ideal design to make inferences regarding the change would involve more than two waves.

1.1.7 Regression to the mean

Gains generally have a negative correlation with the pretest (Linn & Slinde, 1977; Bereiter, 1963; Thorndike, 1966). This is due to the regression to the mean[25]—first pointed by Galton (1886) and Pearson (1930). This means, that gain scores will be higher for the person with lower pretest and this has been regarded as unfairness due to the advantage to persons with a particular pretest measure. This "regression effect" can be seen in virtually all test-retest situations (Freedman et al. 1991).

This claim has been dismissed since a change and a previous status will always be related: the current status is the product of the prior changes (Willett, 1997). The error in the measures will underestimate the true relationship (correlation between true score of pretest and true change), and Rogosa et al. (1982) proposed a correction using a method of moments. Correlation of pretest and gains is "an interesting fact of life" (Rogosa et al, 1982), but does this have any implication on the choice among the two approaches?

Regression to the mean, in my opinion, is one of the key "hints" in choosing which approach to use deciding how to interpret it and requires a very clear line of thought. Most importantly, RV and CS approaches assume "different regressions" to the mean, or to put it more accurately, regressions to "different means". This aspect in particular is, perhaps, the most crucial (and difficult) consideration in preferring one approach to another and requires a careful consideration of exchangeability between the groups.

The RV approach assumes that two groups will regress toward the grand mean. If two groups are indeed regressing to the grand mean, then the group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality, and the difference in gains between groups will simply be the result of the "regression artifact".

However, if the group with the lower mean is in fact lower due to the social demarcation of some sort (e.g., gender, race, income), is the assumption of "regression toward the grand mean"—the assumption of the RV approach—plausible at all? If groups differ at the pretest, then the exchangeability assumption (the assumption that two groups are coming from the same population) needs to be questioned. The only case in which two exchangeable groups may differ at pretest is when the random assignment is the unlucky one[26]. If this is the case, the RV approach must be the approach taken by the researcher.

The CS approach, however, does not require the exchangeability assumption the same way as the RV approach, and assumes that the posttest scores will regress to their group-specific means. Similar points were argued in Allison (1990), Kenny (1975), and Kenny & Cohen (1980). The CS approach becomes a very robust choice in such case. However, this approach can give misleading results when used to compare exchangeable

[25] Regression to the mean is sometimes referred as "regression artifact", "regression fallacy", "regression effect", or, as Galton (1886) originally put it, "regression to mediocrity".

[26] The so–called "unhappy" randomization (Kenny, 1975).

groups (e.g., groups resulting from random assignment, groups for which the treatment indicator is strongly ignorable).

It is safe to state, then, that when two groups differ at the pretest, the CS approach relies on a "nonexchangeability" assumption (e.g., that group differences are not due to an "unhappy randomization"). Regression toward the group-specific means is not an issue if the goal is to compare the nonexchangeable groups. But if the groups are exchangeable, the findings from the CS approach might be just the result of the "regression artifact"—regression to the mean. In other words, due to the regression to the mean, the group with the lower mean will gain more than the group with the higher mean even if the treatment has no effect at all.

The specification of exchangeability and "type" of the regression to the mean, then, are the two most difficult and important decisions one needs to make, and unless there is a random assignment to groups, the decision is going to require a coherent argument and evidence supporting the choice of the approach. Allison (1990) suggested that in ambiguous cases, the best strategy is to use both RV and CS approaches and trust only to conclusions that are consistent across methods.

There are, however, other issues that researcher needs to consider before interpreting the results, and the most important one—measurement error—will be discussed next.

1.1.8 Measurement Error

So far we have assumed that the pretest and the posttest measures do not contain any measurement error. However, the outcome variable is measured with error in many situations. Often in social sciences, when scores from the pre or post measures (e.g., EAP[27] scores) are used in the secondary analyses, the measurement error[28] in the scores is ignored. This section discusses the implications of ignoring the measurement error for the two approaches (CS and RV). When either approach does not account for the measurement error, between-person differences will be decreased due to the unaccounted noise. Thus, both approaches are inefficient when they ignore measurement error, and as will be elaborated below, the RV approach is biased unless the "treatment" is assigned at random.

There is an extensive literature discussing approaches to deal with measurement error (see Carroll et al., 2006; Buonaccorsi, 2010). Accounting for measurement error

[27] EAP (expected-a-posteriori) estimates incorporate distributional information of subjects. MLE estimation of subject-specific latent variables uses the responses for the subject as the only information about the subject (by maximizing the likelihood of obtaining these values). Compared to EAP estimates, MLE estimates have greater prediction-error variance. For more on comparisons between MLEs and EAPs see Rabe-Hesketh & Skrondal (2013, pg. 111). In this chapter, EAP estimates were used for all models that ignore the measurement error.

[28] The error in these variables can either be non-differential or differential. Non-differential error is when the fallible variable (e.g., EAP estimates from the test) contains no information regarding the dependent variable beyond the true score. In other words, Y|(True score, fallible score) = Y|(True score). Differential error (a comparatively rare case) is when the fallible measure has information that is not contained in the true score. This can also occur when, for instance, one uses EAP estimates from the unidimensional model (e.g., a math test which contains geometry and algebra items) as a proxy for the true score on one of the dimensions only (e.g. geometry). In that case, the unidimensional observed score (EAP math score) contains information beyond the true score on geometry.

when comparing groups is simpler in situations when the outcome variable (and one of the independent variables) is measured using surveys or achievement tests. If we want the standard errors of the treatment coefficient to reflect the measurement error (in addition to the errors due to the variance) we traditionally use a so-called *latent regression*, or item response model with manifest predictors (Mislevy, 1987; Verhelst & Eggen, 1989; Zwinderman,1991; Zwinderman, 1997). What are the consequences of not using the latent regression approach? This is discussed next.

1.1.8.1 Ignoring measurement error in the dependent variable – the naïve CS approach

In CS approach (that ignores the measurement error), the only variable on the right side of the equation is the treatment indicator, which is free of error. The dependent variable, however, contains measurement error. If measurement error in the dependent variable is independent of the treatment indicator, then the OLS estimation is consistent and coefficients are unbiased (Wooldridge, 2002), as shown below.

Assume that y^* is the true gain from pretest to posttest that we don't observe and W is the treatment indicator. The CS approach is simply:

$$y^* = \beta_0 + \beta_1 W + \epsilon. \quad \textbf{(17)}$$

Further assume that y is the observed gain score, a manifestation of the true gain, which contains the measurement error such that

$$y = y^* + e, \quad \textbf{(18)}$$

and can be re-expressed as:

$$e = y - y^*. \quad \textbf{(19)}$$

Then, the naïve CS approach can be expressed as:

$$y = \beta_0 + \beta_1 W + \epsilon + e. \quad \textbf{(20)}$$

If the residual term (ϵ) and measurement error (e) are uncorrelated, as we usually assume, then the error variance in Equation 21 is larger than the error variance in the Equation 18, since var$(\epsilon + e)$ = var(ϵ) + var(e) > var(ϵ). This only reduces the power, but does not violate any assumptions of OLS.

1.1.8.2 Ignoring measurement error in the dependent and explanatory variables – the naïve RV approach

In the RV approach, fallible measures appear on both sides of the equation**.** Assume that X^* is the true pretest score and X is the observed pretest score and

$$e_x = X - X^*, \quad \textbf{(21)}$$

is the measurement error at the pretest. Further assume that Z^* is the true posttest score and Z is the observed posttest score and

$$e_z = Z - Z^*, \quad \textbf{(22)}$$

is the measurement error at posttest. The RV approach is then:

$$Z^* = \beta_0 + \beta_1 W + \beta_2 X^* + \epsilon, \quad \textbf{(23)}$$

which can be rewritten as:

$$Z = \beta_0 + \beta_1 W + \beta_2 X + \epsilon - e_x + e_z. \quad \textbf{(24)}$$

If we assume that measurement errors (e_x *and* e_z) are uncorrelated with the residual (ϵ) and other terms in the model, then OLS gives consistent estimates even though the variance of the error increases. As long as we assume that the measurement error for the pretest (e_x) is correlated with the true pretest score only, there is nothing in the Equation 23 that violates the assumptions of the OLS.

However, if we assume that measurement error (e_x) is correlated with the observed version of itself (X), then the covariance of X and e_x is the variance of the measurement error. This case is known as the **classical errors-in-variables assumption** and violates the OLS assumptions and results in inconsistent estimates of all coefficients. The resulting outcome is attenuation bias—coefficients will be attenuated—positive coefficients will tend to be underestimated and negative coefficients will tend to be overestimated. The higher the relationship between the variable measured with error and other covariates, the worse the attenuation bias (Wooldrige, 2001).

However, in the naïve RV approach (i.e., ignoring measurement error), if the treatment was allocated to groups randomly, measurement error in the covariate (i.e., fallible pretest score) does not bias the estimates of group mean differences (coefficient of the treatment indicator) and thus inferences for the group mean differences are correct (Carroll et al., 2006, p. 52; Buonaccorsi, 2010, p.114; Carroll et al., 1985; Carroll, 1989). If we assume that the treatment indicator is independent of the pretest and the expected value of the pretest is the same for each group—guaranteed only with proper random assignment—one can still use the fallible covariate (e.g., pretest) to gain efficiency and obtain an unbiased group mean comparison. The coefficient of the pretest, however, will still be biased, and needs to be interpreted with caution.

1.1.8.3 Accounting for measurement error when comparing groups: latent regression

Structural Equation Modeling (SEM) is a general framework designed to deal with regression of (on) latent variables. By using a model from the SEM framework (e.g., latent regression), we attempt to account for measurement error in the variables when investigating the relationship to the manifest variables. SEM models aim to account and correct for the measurement error in such comparisons, but the correction is only as good as the information provided (DeShon, 1998).

The group indicator (e.g., male vs. female, treatment vs. control) is a manifest variable and the latent variable itself (or multiple latent variables) is measured using a set of items[29]—this is the measurement part of the SEM model[30]. Instead of regressing

[29] Responses to each item are modeled by having the measurement model in the equation. Items can be either categorical (binary, ordinal, nominal) or continuous, or a mix of both. When we assume that the latent variable is continuous, a set of SEM models that deal with the categorical items is known as item response theory (IRT) models, and family of models that deal with continuous items are known as factor analysis (FA) models.

predictions (e.g., EAP scores) on the group indicator, we want to regress the actual latent variable itself within the model.

The analog of the naïve RV approach in the SEM framework, the RV approach that accounts for the measurement error in items (latent RV approach), is shown in Figure 1.5. The two time-points are modeled as two different random-effects. One can then estimate the regression coefficient of the pretest "dimension" (arrow from θ_1 to θ_2 in Figure 1.5 below) in addition to the occasion-specific variances and item parameters.

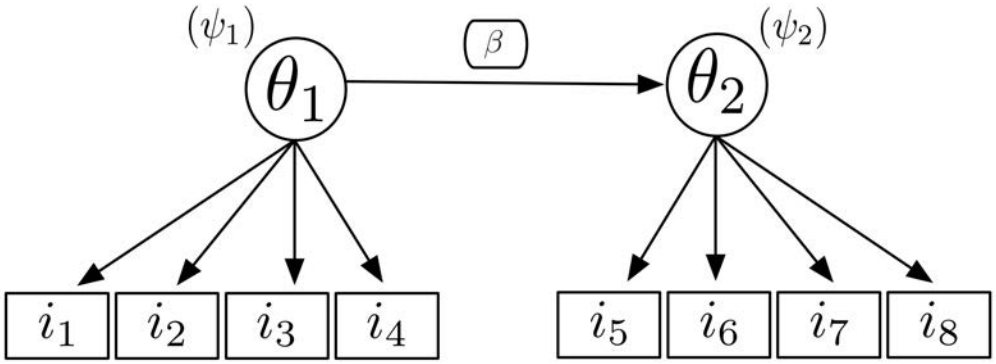


Figure 1.5. The RV approach in the SEM framework with four items administered repeatedly on two occasions.

Alternatively, instead of estimating the regression coefficient of the pretest, we can estimate the correlation between the two random-effects (correlation between dimensions, double-sided curved arrow in Figure 1.6). This model is shown in Figure 1.6 below and is known as Andersen's model (Andersen, 1985) in the Rasch literature.

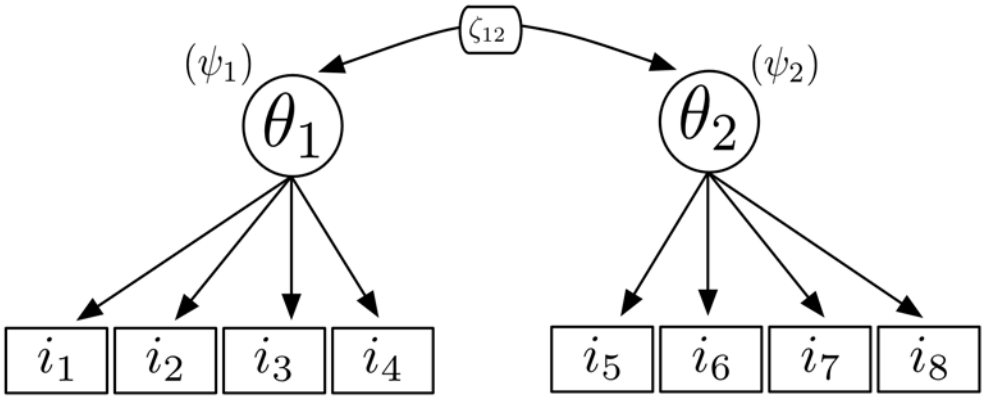


Figure 1.6. Andersen's model (Andersen, 1985) with four items administered repeatedly on two occasions.

[30] Measurement model can also be thought as a multilevel model. Responses to items are nested within person, and variation in these responses describes the level-1 variation resulting in a particular type of IRT model (e.g., Rasch model[30] if the item-specific regression coefficients are fixed to unity). For ordinal items, either adjacent-category or cumulative logit link functions are traditionally used.

The CS approach in the SEM framework is shown in Figure 1.7. For this approach, responses at pretest were loaded on pretest only while responses at posttest are loaded on both dimension resulting in the model in which estimated abilities in the posttest dimension indicate the latent change. This model is also known as Embretson's model for change (Embretson, 1991) in the Rasch literature. Item difficulty estimates are usually anchored to establish a common metric (see Section 1.2.3)

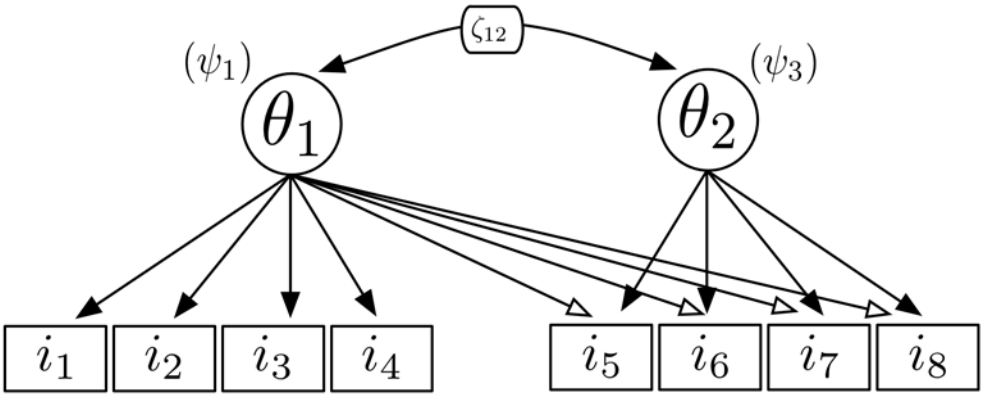


Figure 1.7. Embretson's model for change (Embretson, 1991) with four items administered repeatedly on two occasions.

The analysis of the data in this study present comparison of the RV and CS approaches for both naïve and latent versions. Note that, the paradox itself does not vanish in the latent versions of the model.

Most of the responses at the pretest and posttest were scored polytomously and were analyzed using the partial credit model (PCM; Masters, 1982) and cumulative Rasch model (CRM; Agresti & Lang, 1993)[31]. Below I briefly introduce the PCM and CRM models, and then, for the simplicity of the presentation, continue with the simple Rasch model: the special case of the PCM and CRM models[32] when items are binary. I then present the multilevel Rasch model—the main approach used in the analysis of the response data in the ADM study, discussed in Section 1.2.1.

In the PCM, the probability of person j scoring k on item i, P_{jik}, can be expressed as

$$P_{jik} = \frac{exp \sum_{l=0}^{k}(\theta_j - \delta_{ik})}{\sum_{h=0}^{M_i} exp \sum_{l=0}^{h}(\theta_j - \delta_{ik})}, k = 0,1, \dots, M_i, \quad \textbf{(25)}$$

where θ_j, and δ_{ik} are the ability of person j and the difficulty of step k of item i, respectively; $M_i + 1$ is the number of (ordered) categories for the item, and we use the following notational conventions for identification:

$$\sum_{k=0}^{0}(\theta_j - \delta_{ik}) \equiv 0, \quad \textbf{(26)}$$

and

[31] Most of the software programs designed for IRT modeling allow only either PCM or CRM modeling, thus I use both to make this study easier to replicate.

[32] PCM and CRM models are two different ways of formulating ordinal logits and have different interpretation of parameters. See Chapter 2 of this dissertation for the detailed discussion of the models.

$$\sum_{k=0}^{h}(\theta_j - \delta_{ik}) \equiv \sum_{k=1}^{h}(\theta_j - \delta_{ik}). \quad \textbf{(27)}$$

In the CRM model, the probability of person j scoring k on item i, P_{jik}, can be expressed as:

$$P_{jik} = \frac{exp(\theta_j - \delta_{ik})}{1+exp(\theta_j - \delta_{ik})} - \frac{exp(\theta_j - \delta_{ik+1})}{1+exp(\theta_j - \delta_{ik+1})}, k = 0,1, \dots, M_i. \quad \textbf{(28)}$$

For binary items, the PCM and the CRM simplify to the Rasch model:

$$P_{ji1} = \frac{\exp(\theta_j - \delta_i)}{1+\exp(\theta_j - \delta_i)}, \quad \textbf{(29)}$$

or

$$logit(P_{ji1}) = \tau_{ji1} = \theta_j - \delta_i, \quad \textbf{(30)}$$

in which $\theta_j$and δ_i are the ability of person j and the difficulty of item i, respectively.

Following Raudenbush et al. (2003) and Kamata & Cheong (2007), the Rasch model can also be expressed using a two-stage formulation, as follows:
The item level (level 1) of the model can be expressed as:

$$\tau_{ij} = \pi_j + \sum_{q=1}^{I-1}\pi_{qj}X_{qj}, \quad \textbf{(31)}$$

where X_{qj} is the q^{th} item dummy variable for the subject j with value equal to unity when $i = q$ and 0 otherwise; π_j is the intercept term, and π_{qj} is the coefficient associated with X_{qj}. The formulation in Equation 30 allows elements of **X** and **τ** to vary across subjects (by incorporating the subject subscript). By restricting these to be invariant at level-2, we implement the Rasch assumption regarding the item slopes (i.e., equal discriminations). Thus, the subject-level (level-2) is expressed as

$$\pi_j = \beta_0 + u_{0j}, \quad \textbf{(32)}$$

$$\pi_{qj} = \beta_q, q = 1, \dots, I-1, \quad \textbf{(33)}$$

in which u_{0j} is the latent variable at the respondent level ("ability"). We assume u_{0j} to be distributed as N(0, ξ). This is in distinction to the original Rasch approach, which does not require the assumption of normality as it uses the conditional maximum likelihood (CML) estimation[33] (by conditioning on the sufficient statistics). However, the CML approach does not readily handle covariates, which is an important part of the current formulation.

Note that

$$\tau_{ij} = \beta_0 + \beta_q + u_{0j}, \quad \textbf{(34)}$$

which is similar to the expression in Equation 29 if we express:

$$u_{0j} = \theta_j, \quad \textbf{(35)}$$

and

$$\delta_i = -(\beta_0 + \beta_q). \quad \textbf{(36)}$$

[33] The Rasch model that uses the CML estimation is closely related to the CS approach, as shown in Appendix A.1.

In addition, by adding a third level subscript (e.g., for school or classroom) to the expression in Equation 30, we can express the three-level Rasch model using the three-stage formulation, with level-1 expressed as:

$$\tau_{ijs} = \pi_{js} + \sum_{q=1}^{I-1} \pi_{qjs} X_{qjs}, \quad \textbf{(37)}$$

and level-2 expressed as:

$$\pi_{js} = \beta_{0s} + u_{0js}, \quad \textbf{(38)}$$

$$\pi_{qjs} = \beta_{qs}, \quad \textbf{(39)}$$

and level-3 can be expressed as:

$$\beta_{0s} = \gamma_{00} + \nu_{0s}, \quad \textbf{(40)}$$

$$\beta_{qs} = \gamma_{q0}, \quad \textbf{(41)}$$

with ν_{0s} as a level-3 (e.g., school or classroom) latent variable, assuming $\nu_{0s} \sim N(0, \psi)$.

Thus, the response of person j who is in school s to item i:

$$\tau_{ijs} = \gamma_{00} + \gamma_{q0} + u_{0js} + v_{0s}, \quad \textbf{(42)}$$

with

$$u_{0js} + v_{0s} = \theta_j, \quad \textbf{(43)}$$

and

$$\delta_i = -(\gamma_{00} + \gamma_{q0}), \quad \textbf{(44)}$$

in which q is the item indicator ($q = 1, 2, \ldots I - 1$; there is no dummy variable for the I^{th} item to achieve full rank of the design matrix).

The level-2 and level-3 equations can be extended by including person-level (level-2) covariates such as pre-test score, and cluster-level (level-3) covariates such as treatment indicator in a cluster-randomized study or other cluster level indicators.

Pretest score, essentially, is also obtained using the Rasch model within the same model, thus accounting for measurement error in pretest scores, resulting in multilevel versions of either Andersen's (Andersen, 1985) or Embretson's (Embretson, 1991) models shown in Figures 1.6 and 1.7 respectively. The resulting model can be considered as two-dimensional multilevel Rasch model—a special case of a multilevel SEM approach.

By incorporating the item response model into the multilevel design, one can directly link items with respondents and account for the variation in the responses to items within respondents. In such a model, level-1 represents the variation in responses within subjects (within-subject between-item); level-2 represents the latent variables varying among subjects within clusters; and level-3 accounts for the variation between clusters (e.g., schools or classrooms).

1.1.9 Derivation of bias of RV model if CS model is the correct model

In Section 1.1.4, I showed that the CS approach is not a special case of the RV approach and that these two approaches are different models. Next, similar to Allison

(1990), I derive the bias that follows from incorrectly using the RV model when the model implied by the CS approach is the true model.

To demonstrate that the paradox is inevitable when the RV approach is used incorrectly, I assume that the CS model is the correct model and that the treatment does not have an effect (i.e., $\beta_2 = 0$). As shown in Equation 8, the incorrectly used RV approach[34], in such case is:

$$Y_{2j} = \alpha + \beta_2 W_j + \beta_3 Y_{1j} + \epsilon^{\Delta}. \quad \textbf{(45)}$$

Let ρ_{1W} be the correlation between the pretest (Y_{1j}) and the treatment indicator (W_j) and let ρ_{2W} be the correlation between the posttest (Y_{2j}) and the treatment indicator and let ρ_{21} be the correlation between the pretest and the posttest. For the clarity of the presentation I will drop the j subscript. The partial regression coefficient of Y_1 controlling for W in the RV approach is (see for instance Sen & Srivastava, 1990; Jobson, 1991)

$$\beta_{21\cdot W} = \frac{\sigma_2\sqrt{1-\rho_{2W}^2}}{\sigma_1\sqrt{1-\rho_{1W}^2}} \frac{\rho_{21}-\rho_{2W}\rho_{1W}}{\sqrt{1-\rho_{21}^2}\sqrt{1-\rho_{1W}^2}} = \frac{\beta_{21}-\rho_{2W}\rho_{1W}}{1-\rho_{1W}^2}, \quad \textbf{(46)}$$

in which the β_{21} is the regression coefficient of the pretest when the posttest is regressed on the pretest only. If the treatment does not have an effect, as we assumed above, the above expression can be re-expressed as:

$$\beta_{21\cdot W} = \frac{\beta_{21}-\rho_{1W}^2}{1-\rho_{1W}^2}. \quad \textbf{(47)}$$

We see from the Equation 46 above that $\beta_{21\cdot W}$ will be less than unity—this is consistent with the findings in practically all pre-post studies. The partial regression coefficient for W, controlling for Y_1, can be expressed as:

$$\beta_{2W\cdot 1} = \frac{\beta_{2W}-\rho_{21}\rho_{W1}}{1-\rho_{1W}^2}. \quad \textbf{(48)}$$

Since we assume that there is no treatment effect, we can rewrite the above Equation 47 as

$$\beta_{2W\cdot 1} = \frac{\beta_{1W}-\rho_{21}\rho_{W1}}{1-\rho_{1W}^2}, \quad \textbf{(49)}$$

which, in turn, we can expand as:

$$\beta_{2W\cdot 1} = \frac{\sigma_2\sqrt{1-\rho_{21}^2}}{\sigma_W\sqrt{1-\rho_{1W}^2}} \frac{\rho_{1W}-\rho_{21}\rho_{1W}}{\sqrt{1-\rho_{21}^2}\sqrt{1-\rho_{1W}^2}} = \frac{(1-\rho_{21})\rho_{1W}}{1-\rho_{1W}^2}. \quad \textbf{(50)}$$

Let's decompose the residual at pretest, ϵ_1 in Equation 3, into stable and dynamic components, such that

$$\epsilon_{1j} = U_j + V_{1j}. \quad \textbf{(51)}$$

In Equation 50, U_j is the stable component of the construct and V_{1j} is the dynamic component (e.g., state at pretest, which can change at posttest). Recall that δ in Equations

[34] Recall that estimates are biased since Y_{1j} is correlated with ϵ^{Δ} by construction.

3 and 4 is the preexisting difference between the two groups and $cov(W,U)/var(W)$ is the correlation between the stable component (U) and the treatment indicator (W). Then, the correlation between the treatment indicator and the pretest, ρ_{1W}, can be expressed as:

$$\rho_{W1} = \delta + \frac{cov(W,U)}{var(W)}. \quad \textbf{(52)}$$

If we substitute ρ_{1W} in the above equation for ρ_{W1} in the Equation 49, we obtain:

$$\beta_{2W\cdot 1} = \frac{(1-\rho_{21})[\delta+\frac{cov(W,U)}{var(W)}]}{1-\rho_{1W}^2}. \quad \textbf{(53)}$$

The expression above (Equation 52) will be non-zero even when there is no treatment effect—it will be zero only when: (1) there is no preexisting difference between the two groups ($\delta = 0$) AND (2) the treatment assignment is independent of the stable component of the construct measured repeatedly (when $cov(W,U) = 0$).

1.1.10 Implications of RV

As was noted previously, in almost all applications, the coefficient of the pretest in the RV approach will be between 0 and 1. Let's assume that the RV approach is the correct model and that the treatment does not have an effect. Then,

$$Y_{2j} = \beta_1 + \beta_2 W_j + \beta_3 Y_{1j} + \epsilon_j, \quad \textbf{(54)}$$

can be rewritten as:

$$Y_{2j} - \beta_3 Y_{1j} = \beta_1 + \beta_2 W_j + \epsilon_j. \quad \textbf{(55)}$$

If the β_2=0 (coefficient of the treatment indicator is zero), as we assumed, then

$$Y_{2j}^{W_j=0} - \beta_3 Y_{1j}^{W_j=0} = Y_{2j}^{W_j=1} - \beta_3 Y_{1j}^{W_j=1}, \quad \textbf{(56)}$$

which can be re-expressed as

$$\beta_3 Y_{1j}^{W_j=1} - \beta_3 Y_{1j}^{W_j=0} = Y_{2j}^{W_j=1} - Y_{2j}^{W_j=0}, \quad \textbf{(57)}$$

and consequently,

$$\beta_3 (Y_{1j}^{W_j=1} - Y_{1j}^{W_j=0}) = Y_{2j}^{W_j=1} - Y_{2j}^{W_j=0}. \quad \textbf{(58)}$$

This implies that the mean difference on the posttest will be less than the mean difference in the pretest. In other words, two group means will come closer to the grand mean. This implication is not plausible when the exchangeability of the two groups is not a reasonable assumption, as I will discuss below.

Allison (1990) provides a hypothetical example: the two groups are males and females, and all males are assigned to the treatment condition and all females are assigned to the control condition. Suppose that the outcome variable is the productivity at work, which is measured at pre and post. Further suppose that the correlation between productivity at the pretest and the posttest is 0.5 and variances are same at both occasions. If the treatment has no effect, the RV approach implies that the gender gap in the productivity on the posttest should be only half of the gender gap at the pretest—an unintuitive and unjustifiable implication. The CS approach, in turn, would show that the

gender gap is same at both time-points, as there is no treatment effect.

1.1.11 "It depends"

Lord's Paradox is considered "by far, the most difficult paradox to disentangle and requires clear thinking" (Wainer & Brown, 2007, p.25)[35]. To the question of "which approach to use" the safest answer is usually "it depends."

The first question that needs to be answered is whether groups a researcher wants to compare are exchangeable with respect to the outcome of interest. For instance, it might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym, and not exchangeable if the outcome of interest is the mean number of calories burned. Subject matter expertise is a must to answer this question.

As I elaborated above, the exchangeability assumption is related to the assumption of the regression to the grand mean. If a researcher suspects that, instead, the regression to the group-specific means is more plausible assumption, then s/he should prefer the CS approach. If, however, the measure at pretest is the counfounder, then a researcher should use the RV approach.

If the groups are not a result of random assignment, the RV approach has an additional drawback: measurement error. In that case, the regression model that does not account for the measurement error will produce biased estimates.

"Making sure that findings are consistent when both approaches are used" is perhaps the safest suggestion when a researcher is confused. But this shouldn't result in discarding an important insight into the intervention. By formulating the research question clearly and by considering the assumptions of each of the two approaches, a researcher can find answers that may be useful to stakeholders. This is what I attempt to demonstrate in the next section using some exemplary real data.

1.2.1 ADM study

This section examines the effects of a Data Modeling curriculum designed to improve statistical reasoning skills, as well as, general math achievement when compared to the existing curriculum. To answer that question, we conducted an ADM Efficacy Study, in which schools were randomly assigned to either the treatment or the control condition and pre- and post-tests were administered to students at these schools before and after the treatment[36].

The treatment indicator is manipulated at the macro-level (i.e., the school level), in what is known as a cluster-randomized trial (CRT). The cluster-randomized trial design helps in avoiding possible contamination between the treatment and control conditions in contexts where the within-macro randomized assignment would likely lead to this type of problem, and where there are important macro-level aspects of the treatment.

One alternative to cluster-randomized trials is a multi-site trial design (MST),

[35] In comparison to Simpson's Paradox and Kelley's Paradox (see Wainer & Brown, 2007 for details).

[36] Clearly, the allocation to either the treatment or the control conditions was done before the pretest was administered. Therefore,,the treatment assignment does not depend on the pretest.

where the treatment assignment is randomized to individual units (e.g., students) within each cluster (e.g., school). Multisite trials are more efficient (Moerbeek, van Breukelen, & Berger, 2000), but at the price of the risk of interference between units in different treatment conditions.[37]

There are also multi-site cluster-randomized trials (MSCRT), in which treatments are randomized for subgrous within each cluster such as treatments randomly assigned to teachers within each school (see Wijekumar, Hitchcock, Turner, Lei, & Peck, 2009).

It is instructive to briefly discuss the advantages and disadvantages of the cluster-randomized trials and assumptions of the design used in the study. Before that, I will briefly discuss the psychometric properties of the ADM pre- and post assessments and the procedure linking (e.g., putting on the same scale) these two tests.

1.2.2 ADM assessment

The measure that was administered as both a pretest and posttest is the ADM Statistical Reasoning Measure developed by Rich Lehrer at Vanderbilt in conjunction with the Berkeley Evaluation and Assessment Research (BEAR) Center. The measure has five sub-dimensions (domains): Data Display (DAD), Models of Variability (MOV), Chance (CHA), Concepts of Statistics (COS), and Informal Inference (INI) (see Appendix A.2 for the description of these constructs). The overall composite and domain-specific scores for each student were produced based on IRT analyses—Rasch model in particular. For a detailed description of each of the levels of these constructs, along with learning progressions, see Schwartz (2012).

Table 1.2 summarizes the analysis and psychometric properties of the pretest instrument consisting of 23 items, 18 of which are common with the "Post 2011" test (described below). These 18 items were anchored to the difficulty estimates obtained from the calibration of "Post 2011" to establish the common scale (discussed in the next section). The analysis shows a reasonable (between $3/4$ and $4/3$) item fit[38] for the all of the non-anchored items. The reliability statistics were good: the EAP/PV person separation reliability estimate was estimated at .89 and Cronbach's alpha was estimated at .84.

The analysis also showed that the range of easy, moderate, and difficult items provided good coverage of the student proficiency distribution.

Table 1.3 summarizes the analysis and psychometric properties of the postest instrument consisting of 25 items, 13 of which are common with the "Post 2011" instrument and thus were anchored to obtain the common scale. The analysis shows reasonable item fit for the all of the estimated items. None of the step or item parameters are showing problems with the fit. The reliability statistics were again good: EAP/PV person separation reliability estimate was estimated at .87 and Cronbach's alpha was estimated at .87.

[37] For more on power calculations and optimal design for cluster-randomized and multisite trials and comparison of the two, see Moerbeek, van Breukelen, & Berger (2000; 2001a; 2001b; 2008), Raudenbush & Liu (2000), Snijders & Bosker, (2012), and Ryan, (2013).

[38] A common convention of 3/4 (0.75) and 4/3 (1.33) is used as an acceptable lower and upper bounds (Adams & Khoo, 1996).

Table 1.2. Analysis results for Pre 2013 test

Sample Size	894
Number of items in calibration	23
Number of common (anchored) items	18
Number of polytomous items	21
Number of dichotomous items	2
Missing data [39]	none
Model	PCM
Weighted Fit MNSQ >1.33, T sig. (Item Parms)	none
Weighted Fit MNSQ >1.33, T sig. (Step Parms)	none
Reliability estimates:	
Estimated *aprior*/person variance reliability (EAP/PV)	.89
Cronbach's Alpha	.84

Table 1.3. Analysis results for Post 2013 test

Sample Size	789
Number of items in calibration	25 total
Number of common (anchored) items	13
Number of polytomous items	25
Number of dichotomous items	none
Missing data	none
Model	PCM
Weighted Fit MNSQ >1.33, T sig. (Item Parms)	none
Weighted Fit MNSQ >1.33, T sig. (Step Parms)	none
Reliability estimates:	
Estimated *aprior*/person variance reliability (EAP/PV)	.87
Cronbach's Alpha	.87

The analysis of the postest also showed that the range of easy, moderate, and difficult items provided good coverage of the student proficiency distribution. The population standard deviation was .65.

Once pretest and posttest were linked through the "Post 2011" test (see next section), the gains in EAP estimates ($EAP_{Post} - EAP_{Pre}$) were computed for students who were administered both pretest and posttest. These gains were also computed for

[39] Missing responses that are not missing systematically were coded as incorrect (zero).

each of the relevant domain scores, namely for DAD, COS, CHA, MOV, and INI domains.

Correlations between the dimensions obtained from the multidimensional analysis in which the posttest and "Post 2011" response data sets were calibrated together (for larger sample size and more number of items) are shown in the Table 1.4 below.

Table 1.4. Correlations between domains and variance for each domain

	DAD	MRC	COS	CHA	MOV
DAD					
MRC	0.84				
COS	0.77	0.81			
CHA	0.79	0.78	0.78		
MOV	0.69	0.78	0.81	0.87	
INI	0.91	0.87	0.93	0.88	0.83
variance	0.26 (0.01)	0.73 (0.02)	0.90 (0.03)	0.61 (0.02)	1.25 (0.04)

As we see from the Table 1.4, the DAD dimension has the highest correlation with INI dimension and lowest with the MOV dimension—this finding will be useful in the coming sections.

Equating the tests is necessary to achieve a common scale and comparability among the tests. Two tests need to have sufficient number of common items to allow common-item equating. Another option for achieving the common scale between tests without any common items (as is the case here) is to have a third test of reference with which two tests of interest share common items. This approach was used in linking the pre- and post- tests. The "Post 2011" test administered to students of similar grade level as part of a larger study was chosen as the reference scale. The pre- and post-tests were calibrated by anchoring the item parameters to values obtained from the analysis of the "Post 2011" reference test, as discussed in the next section.

1.2.3 Test design and linking of the tests

In order to avoid possibilities of cross-contamination of the pre- and post- tests results, these two tests do not have any common items. In order to obtain change scores, however, we need these two measures to be on the common scale. The common scale for the two tests are linked using a pre-calibrated data set, a large set of ADM items administered in 2011 ("Post 2011"). The pretest has 23 items and posttest has 25 items: of these 48 items, 32 are common with the test administered to the larger sample in 2011.

Specifically, for the pretest, the item parameters were calibrated by (1) anchoring the 18 items that were administered in 2011 to the estimates obtained at that time, and (2) estimating the item parameters for the remaining five (new) items using the current data set. Similarly, for the post-test, the items were calibrated by (1) anchoring the 14 items that are common with the 2011 test, and (2) estimating the parameters of the remaining 11 (new) items using the current data set.

As a result, parameters for all of the 48 items used in the analyses were obtained and anchored for all of the subsequent analyses. Linking was done using both PCM and CRM models. This approach to analyze the pre- and post- test provides a strong factorial

measurement invariance (Millsap, 2011).

1.2.4 Cluster-randomized trials

CRT is less efficient compared to the classical randomization (Cornfield, 1978) and multisite trials (Moerbeek, van Breukelen, & Berger, 2000). However, in the CRT, the interference between units within the cluster does not bias the inference and internal validity of the study. Nevertheless, in CRT, we still assume no interference between clusters—that is, schools in different treatment conditions do not interact with each other. In addition to the robustness against within-cluster contamination, cluster-randomized trials are usually financially and logistically more convenient to implement. In settings where clusters are schools or classrooms, sometimes the only feasible option is to administer the treatment or control conditions to the whole cluster, mainly due to fairness and ethical considerations.

"Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception, however, and should be discouraged" (Cornfield, 1978, p.101-102). Therefore, in such designs, one needs to analytically account for clusters in the estimation, generally, by using hierarchical linear modeling (HLM) techniques. Including dummy variables for clusters in the regression model (the so-called fixed-effects or dummy-variable approach) is not possible since the treatment indicator does not vary within clusters, as opposed to multisite trials, and thus perfectly collinear with the combination of cluster dummies. However, in comparison to multisite trials, statistical power will be lower even after we account for the clustering when analyzing the data.

In cluster-randomized trials, risk of imbalance between treatment and control groups on important covariates is higher compared to multisite trials—the number of randomization units (number of clusters) in CRT will always be smaller than the number of randomization units in multisite trials (number of individuals within clusters).

Assume a CRT design with schools as clusters. The treatment is at the school level while the outcome is at the student level. For students $i = 1,2,3,..N$ nested within clusters $j = 1,2,3,\dots,M$, the cluster-level treatment is denoted by $T_j \in \{0,1\}$, and the outcome is expressed as $Y_{ij} = Y_{ij}(T_j)$. Proper random assignment guarantees independence, expressed as:

$$Y_{ij}(1), Y_{ij}(0) \perp T_j. \quad \textbf{(59)}$$

Then, the population average treatment effect (PATE) is simply $E[Y_{ij}(1) - Y_{ij}(0)]$. Assuming $N_j = N$ for all j (i.e., equal cluster sizes), the difference-in-means estimator, then, is:

$$\hat{\tau} \equiv \frac{1}{M_1}\sum_{j=1}^{m} T_j \bar{Y}_j - \frac{1}{M_0}\sum_{j=1}^{m}(1 - T_j)\bar{Y}_j, \quad \textbf{(60)}$$

in which $\bar{Y}_j$ is:

$$\bar{Y}_j \equiv \sum_{i=1}^{N_j} Y_{ij}/_{N_j}. \quad \textbf{(61)}$$

1.2.5 Cluster-randomized assignment of treatment in the ADM study

Schools in four districts from cities in southwestern U.S. agreed to participate in the study. In this design, schools within a particular district might be similar with respect to the outcome of interest or other variables due to the common district they belong to. Therefore, random assignment to the treatment or control groups was done within districts. For instance, schools in district A were randomized separately from schools in district B. This type of assignment is known as blocking or stratified randomization[40].

Blocking ensures that district membership of schools do not affect our inferences. As a result, treatment and control groups are similar with respect to district composition and the imbalance between treatment and control groups arising from the district membership is greatly reduced. In other words, district membership is no longer a confounding variable after we randomly allocate to treatment or control groups within each district. This reduces variability between treatment conditions associated with particular district-specific polices, practices, and student body composition that may influence these learning-related outcomes.

When comparing randomization with and without blocking, randomization with blocking reduces estimation variance of the causal effect (Imai, King, & Stuart, 2008). However, blocked randomization has fewer degrees of freedom and thus lower power in small samples, but this is only an issue when you have a very small number of units within blocks, and is still preferable to classical randomization (Imai, King, & Stuart, 2008). Sample size within blocks should be sufficiently large to fully utilize the advantages of the stratified randomization.

The blocked randomization design, as described above, is an *unmatched* design within each stratum, with the strata being the district. Schools within districts can vary greatly as well, and if they do, random assignment of the treatment within districts will not be very efficient. Recall that cluster-randomization reduces efficiency compared to multisite trials (Moerbeek, van Breukelen, & Berger, 2000).

However, this loss in efficiency can be limited by pairing similar schools (clusters) within each stratum (i.e., district) and assigning the treatment randomly within each matched pair. Matching before assigning the treatment ("pre-randomization matching")[41], is a preferable approach from the perspective of efficiency, power, bias, and robustness (Imai, King, Nall, 2009). This is also known as matched pair design with non-exchangeable pair members (Raudenbush, 2008). This can be seen as an additional stratification within each strata with only two units within strata[42].

With the small number of clusters, the risk of imbalance between treatment and control groups is greater since clusters (and not students) are assigned at random. Therefore, a matched pair design is vital in such studies. Such design also improves precision and power of the trial (Hayes & Moulton, 2009), and results in reduced power only with very small samples (e.g., ten pairs) assuming equal cluster sizes (Martin, Diehr,

---

[40] "Block what you can and randomize what you cannot" wrote Box, Hunter, & Hunter, 2005 (p. 93).

[41] Another criteria to consider when matching can be cluster sizes.

[42] The idea of matching is similar to stratification, but in paired matching, only one school is allocated to the treatment group and one to the control group. In stratification, (e.g., when district is the strata), in each district, more than one school is allocated to treatment.

Perrin & Koepsell, 1993). Our study used the matched pair design, and we discuss advantages and disadvantages of this design next.

1.2.6 Matched pairs

Within each district, schools were matched to form pairs. Matching[43] of schools was performed based on previous year's aggregate test scores. Within each matched pair (e.g., for two matched schools within a particular district), one treatment school was randomly chosen. This approach minimizes the chances of a poor split especially if the matching criteria—previous year aggregate test score—is correlated highly with the outcome variable.

If the pairing was appropriate, the variance between similar schools in previous year aggregate test scores within a matched pair will be less than the variance between all schools. Matching itself (in the design stage, before the randomization) did not result in any change in the participation of any of the schools, but merely served to reduce the covariate imbalance (previous year's aggregated test score of the school) between the groups as much as possible.

The process used to assign schools within each district, with some additional considerations of the schools size, was as follows (DMS report, 2014): (1) random numbers were assigned to each school; (2) the group of schools were then sorted by their 6th grade mean NCE Math score and divided into blocks; (3) within each block, schools were sorted by the total number of students who had completed the state's standardized Math test (AIMS) in the sixth grade (this sorting variable was viewed as an indicator of school size); (4) each block was then split evenly into two groups: one group included the schools with the lowest numbers of students who had completed the AIMS Math test, and the other included those schools with the highest numbers; (5) groups with two schools in each were then created by pairing a school from the group with the highest number of AIMS test-takers with one from the group with the lowest number; (6) for each pair, the school with the lowest randomly assigned number was designated as a control school. The school with the highest random number was assigned to the Data Modeling treatment group.

1.2.7 Dropouts

There were a number of complications that compromised the randomization design of the study. A number of schools opted out after they heard the randomization results. In particular, four schools assigned to the treatment group opted out from the study immediately after the randomization results were known, as well as three more schools a bit later. Five schools from the control group opted out right after the randomization results were announced.

Reasons for opting out are unknown. It's possible schools that opted out after randomization decided that the new curriculum is disruptive. The usual practice in CRT randomized trials is to obtain a signed consent of the administrators to stay in the study regardless of the randomization result in order not to bias the study, but it is difficult to

[43] "Matching" discussed here is different from the matching generally done after the treatment has been assigned (e.g., propensity score matching).

implement and enforce such procedures in practice.

These drop-outs of clusters, particularly after learning their intervention group, in addition to loss of power and precision, raise the concern of selection bias. This is particularly problematic if schools that drop out differ from schools in the study on important relevant (and possibly unknown) covariates. Most importantly, however, missing data due to some clusters dropping out raises additional complication due to the specific design of the randomization that was based on pairing. Specifically, when a cluster (school) is lost from the trial, the entire matched pair is lost, and as a result the remaining cluster does not have its matched cluster for the comparison. This is particularly damaging for inference when there is a small number of clusters. The additional concern due to the loss of participants (due to the complications related to the timing of the pre-test) is discussed below in detail.

Analyzing the data matched by design as if not matched, or the so-called *breaking the matches* approach (Hayes and Moulton, 2009), might not be appropriate since variance estimators tend to be biased (Donner, Taljaard, & Klar, 2007; Imai, 2008). However, in the case of dropouts and loss of clusters due to further complications, this is the only feasible approach and, to some degree, reduces the implications of the dropouts in the matched pair design. In addition, we avoid a reduction in power (less loss of degrees of freedom) that usually comes with matched analysis. Matched design followed by unmatched analysis was suggested by Diehr, Koepsell, & Cheadle (1995a) for the cases with fewer clusters. We therefore employ matched design with unmatched analysis, and this was specified before the data analysis to avoid data snooping. In addition, the team analyzing the data (BEAR Center) was not provided the pair indicators that would allow matched analysis. A more serious complication is discussed in Section 1.2.9 below.

1.2.8 Students, teachers, and schools in the data

Some students in the sample either changed schools during the course of the study or were absent during the pre- or post- administrations. We assumed that students' switching schools or being absent from the test administration is unrelated to the test outcome (missing at random [MAR]). There were a total of 893 students who took the pretest and 798 students who took the posttest out of total of 914 students. These 914 students are nested in 40 teachers from 21 schools that agreed to participate in the study.

After students with either missing pretest or posttest were eliminated, there remained 768 students who were included in the analysis. Of these students, 406 are in the control group and 362 are in the treatment group. In the treatment group, the sample size decreased from 456 to 362. The loss in the sample was smaller in the control group—from 458 to 406. A summary of the sample size for each group is shown in the Table 1.5 below.

Table 1.5. Students, teachers, and schools in the treatment and control groups.

	treatment group			control group		
	students	teachers	schools	students	teachers	schools
initial sample	456	19	10	458	21	11
nonmissing pre and post	362	19	10	406	21	11

There were a total of 40 teachers from 21 schools within four districts: six schools with 11 teachers (five teachers from three schools assigned to the treatment) in District A; four schools with five teachers (three teachers from two schools assigned to the treatment) in District B; five schools with 11 teachers (four teachers from two schools assigned to the treatment) in District C; and six schools with 13 teachers (seven teachers from three schools assigned to the treatment) in District D.

1.2.9 Late pretest

While the initial plan was to administer the pretest to all of the students before the treatment initiates, a subset of students from the treatment group were pre-tested after the treatment has initiated. In particular, nine of the 19 teachers in the treatment group were administered the pretest *after* they had taught Units 1 and 2 of the Data Modeling curriculum due to the delayed IRB approval. Units 1 and 2 of the ADM curriculum correspond to the DAD module of the curriculum[44].

Due to the late pretest, the pretest score for a subset of the treated students is not a "pre-treatment variable" anymore and this complication needs to be addressed properly. We did not completely eliminate students who took the pretest late from the analysis sample; however, we separated them from students who were pretested before the treatment started. We will still use this subset of the treatment group data to support our findings on the treatment effect. In particular, three groups[45] were formed:

Group A (treatment_A): schools assigned to the treatment condition and pretest administered before the treatment,

Group B (treatment_B): schools assigned to the treatment condition and pretest administered shortly after the treatment was initiated (potentially biasing the pretest), and

Group C (control): schools assigned to the control condition.

Summary of sample size for each the three groups is shown in the Table 1.6 below

Table 1.6. Students, teachers, and schools in three groups

	students	teachers	schools
Treatment Group A	189	10	6
Treatment Group B	173	9	6
Control Group C	406	21	11

The main approach in estimating the treatment effect will consist of comparing the treatment Group A (consisting of 189 students from 6 schools) with the control Group C (consisting of 406 students from 11 schools). However, the comparison of (i) Group B with Group A, and (ii) Group B with Group C may provide additional insights to support findings.

[44] The new curriculum (treatment) consists of the modules that instruct students on topics related to the five related dimensions (domains): DAD, COS, CHA, MOV, and INI, discussed in Section 1.2.2 and Appendix A.2.

[45] The difference between groups A and B in the pretest is that group B received the DAD module and group A did not. That "pseudo-treatment" (being in Group B vs Group A) can be assumed to be assigned randomly since we have no reason to assume that schools that were delayed due to IRB approval differ systematically on important background characteristics.

RV and CS approaches used in the study are illustrated in Figures 1.8 and 1.9 respectively, where three lines are used to represent the three groups. Distances between the lines indicate the magnitude of the treatment effect for the three groups. Notice that the distance between the lines in CS approach (Figure 1.8) differs from the distances in the RV approach (Figure 1.9). One can visually observe from Figure 1.8 that group A (treatment_A, blue line) has higher gains than group C (control group, red line). In the RV approach, in contrast, this difference disappears. This is one form of Lord's paradox.

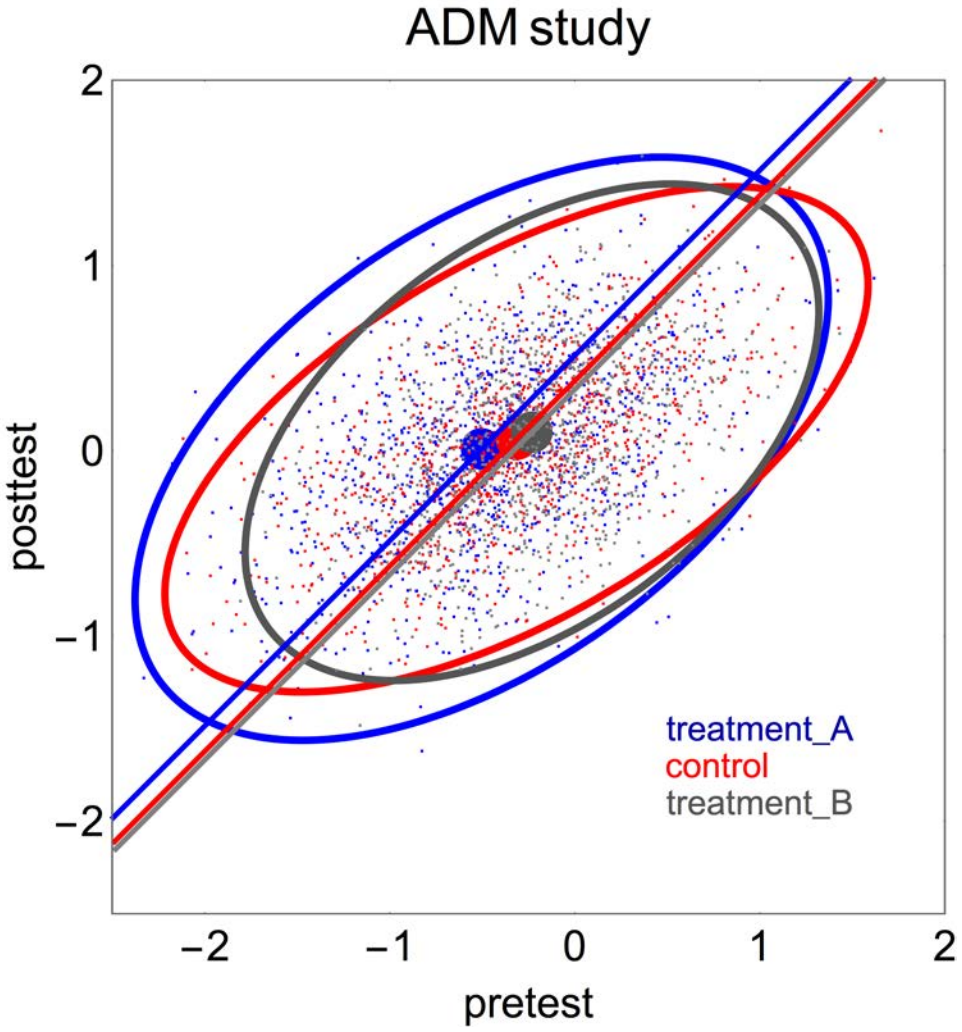


Figure 1.8. CS approach to the ADM study: three lines represent 45° line for three groups.

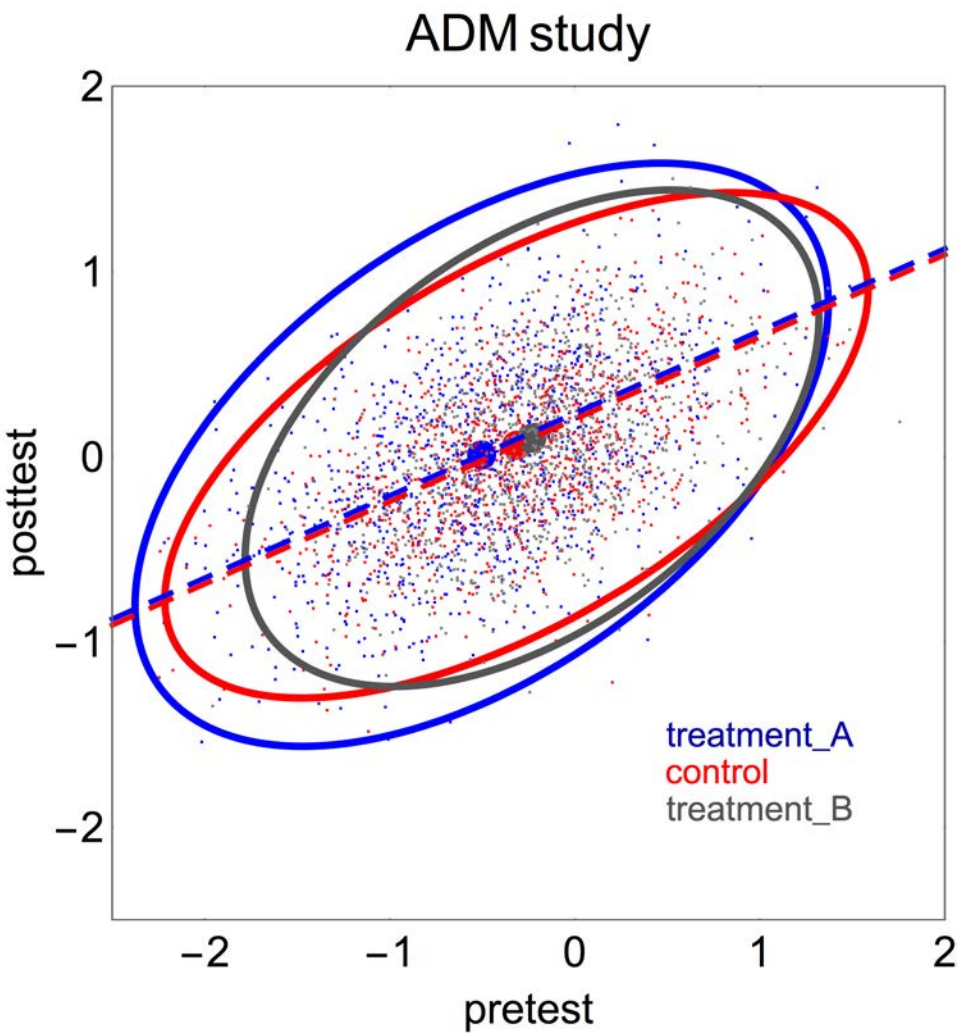


Figure 1.9. RV approach to the ADM study: three lines represent the coefficient of pretest (slope of line = 0.44). The red and gray lines are on top of each other.

It is clear what we would expect to find from the comparison of groups A and the control group: if the treatment has an effect, gains in group A will be higher than gains in group C. This statement can be formulated with further specificity: if the randomization was proper and if the treatment has an effect, groups A and C should not differ on the pretest but group A should be higher than group C on the posttest.

How, then, can we formulate our hypotheses when the group B is involved in pair-wise comparisons? If the treatment has an effect, we would expect differences in both pretest and posttest when comparing groups B and C. This is because both pretest and posttest of group B is a post-treatment measure, specifically for the module that was taught before the pretest in group B (DAD domain).

When comparing group B with group A, however, we would expect no difference in means at posttest but difference in means at pretest. This is because at posttest, groups A and B both have completed the same treatment and we wouldn't expect to see any differences. At the pretest, however, group B was already exposed to the treatment and we would expect to see differences in means at pretest, if the treatment has an effect and if "natural" assignment to groups A and B was not biased.

If the treatment has an effect, the difference in change scores between groups B and C is expected to be less than the difference in change scores between group A and C. This is because, if the treatment has an effect, group B should already have gained scores between their actual start of the treatment (2-3 weeks before the pretest) and the actual pretest date.[46] A different formulation of this hypothesis is: if the treatment has an effect, gains for group the A should be higher than gains for the group B.

To address the issue of the late pretest, our collaborators on the project suggested eliminating items measuring the DAD dimension from the estimation of the composite score in the pretest and posttest instead of eliminating students who were pretested late, However, as shown in Section 1.2.2, the DAD dimension is highly correlated with the other dimensions and there is a risk that Units 1 and 2 of the curriculum had impact not only on the performance of students at pretest on DAD related items but also on performance on items measuring other dimensions. In addition, the EAP estimates of the composite score with and without DAD items are highly correlated (more than .92).

Also notice that the elimination of the group B from the pure treatment group (group A) is not due to the noncompliance or any other factor related to the outcome. One might argue that we should still keep group B together with Group A and obtain the intention-to-treat (ITT) estimator (Fisher et al., 1990), which gives an unbiased estimate in the case of noncompliance or missing outcomes. Teachers in group B haven't been noncompliant (i.e., didn't violate any intervention protocols). The main motivation to separate group B from the pure treatment group (group A) is that the pretest scores of students in group B are not valid for the estimation of treatment effect utilizing the pre-post design and gains of students in group B will bias the estimate of the treatment effect by indicating smaller gains from the treatment.

The multidimensional nature of the construct will be utilized in the next section to provide evidence for the treatment effect. In particular, given that Units 1 and 2 are aimed at increasing the performance in the DAD dimension, the implication of the comparisons of the groups A and C with the group B will be different whether the outcome being compared is the DAD score or a non-DAD score.

1.2.10 Pre-treatment balance

Considering the design of the ADM study described above, if the treatment and control groups differ significantly on important covariates at the pretest, we would conclude that either the randomization wasn't successful, or the dropouts after learning their condition resulted in the imbalance. The latter case implies incomparability (nonexchangeability) of the two groups. There is a chance for unhappy randomization in this study—randomization relies on chance and number of units (schools) is small.

Balance in covariates, however, is not necessary for the CS approach as long as we assume that the covariates that are not balanced do not affect the dynamic component of the main outcome. Table 1.7 below shows percentages of males, students with disability, English language learners, and proportions of Hispanic and White students by each group (treatment groups A and B, and the control group).

[46] It is not straightforward to formulate analogous hypothesis using the RV approach.

Table 1.7. Covariates by each of the three groups.

	group A	group B	control group	total
male	46%	51%	49%	48%
disability	7%	7%	5%	6%
ELL	8%	5%	5%	6%
hispanic	53%	64%	34%	45%
white	31%	22%	50%	39%

As can be seen in Table 1.7, there are differences in proportions of Hispanic and white students between the groups. Considering all the factors that can potentially undermine the virtues of randomization (e.g., dropouts of clusters after learning the randomization results), the design we ended up with may well be considered the so-called nonequivalent control group design.

Tables 1.8 and 1.9 below show balance on composite and component EAP scores at pretest between the three groups. Scores are obtained from the partial credit model (PCM) and the cumulative Rasch (CRM) models respectively.

Table 1.8. Composite and component scores at the pretest using the partial credit model.

PCM	group A		group B		control group	
	mean	var	mean	var	mean	var
composite	-0.53 (0.05)	0.44 (0.05)	-0.21 (0.04)	0.25 (0.03)	-0.31 (0.04)	0.45 (0.04)
DAD	-0.60 (0.07)	0.53 (0.10)	-0.03 (0.05)	0.24 (0.06)	-0.63 (0.05)	0.59 (0.08)
CHA	-0.64 (0.09)	1.38 (0.21)	-0.41 (0.07)	0.71 (0.12)	-0.16 (0.06)	1.02 (0.11)
COS	-1.18 (0.14)	3.13 (0.50)	-0.74 (0.13)	2.55 (0.41)	-0.81 (0.10)	3.08 (0.33)
MOV	-0.09 (0.10)	1.29 (0.23)	0.02 (0.10)	1.23 (0.23)	0.09 (0.07)	1.42 (0.17)
INI	-0.31 (0.07)	0.31 (0.11)	0.00 (0.05)	0.11 (0.06)	-0.03 (0.03)	0.14 (0.04)

Table 1.9. Composite and component scores at the pretest using the cumulative Rasch model.

CRM	group A		group B		control group	
	mean	var	mean	var	mean	var
composite	-1.07 (0.09)	1.54 (0.18)	-0.58 (0.09)	1.12 (0.14)	-0.63 (0.07)	1.60 (0.13)
DAD	-1.14 (0.11)	1.23 (0.27)	-0.17 (0.11)	0.88 (0.24)	-1.18 (0.08)	1.50 (0.21)
CHA	-1.22 (0.16)	4.07 (0.57)	-0.81 (0.13)	2.29 (0.36)	-0.41 (0.10)	2.92 (0.29)
COS	-2.00 (0.21)	6.69 (1.00)	-1.33 (0.20)	5.50 (0.83)	-1.39 (0.14)	6.54 (0.66)
MOV	-0.39 (0.14)	2.76 (0.44)	-0.21 (0.14)	2.73 (0.46)	-0.10 (0.10)	2.94 (0.32)
INI	-0.76 (0.13)	1.42 (0.37)	-0.05 (0.12)	0.68 (0.28)	-0.16 (0.08)	0.95 (0.21)

As we see in Tables 1.8 and 1.9, the mean of the treatment B group is higher than the mean of the treatment A group at pretest in all of the scores. The mean of the treatment B group is higher than the mean of the control group at pretest in composite, DAD, and INI domains. Also, variances of all scores in the group B are lower than in the group A and control groups. We will return to the comparison of scores between the three groups at pretest to gather evidence for the treatment effect.

1.2.11 Exploiting the multidimensional nature of the test to investigate the treatment effect

William Cochran had two major pieces of advice when drawing causal conclusions from observational studies (cited in Rubin, 2006, p. 11). First, he advised researchers to speculate about sources and directions of residual bias. One should ask what unmeasured disturbing variable might be affecting the conclusions of the study and how might these conclusions change if the variables were controlled.

And second, he advised researchers to formulate complex and interrelated causal hypotheses. This implies that one, for instance, should ask: if there is a causal effect of the treatment, which outcome variables should be most affected and which least, and for which groups should the effect be largest and smallest? This is what I attempt to do next.

In this section, investigate and gather evidence by posing elaborate questions by exploiting the three of the following: (1) the multidimensionality of the pretest and posttest, (2) the multi-module[47] nature of the treatment curriculum, and (3) the late pretest for the group B. All comparisons of groups in this section are based on the so-called latent regression model (see Section 1.1.8).

We have seen arguments that make the reliance on the random assignment of the treatment questionable (see Section 1.2.10). However, since groups A and B both ended up in the treatment group, and the formation of the groups A and B can be assumed to be at random (i.e., we assume that the IRB approval being late is not related to the outcome or any other relevant covariate). Then we can assume that the comparison of the groups A and B are free of confounders, barring covariate imbalance purely due to the number of clusters (since clusters, not students, are assigned to treatment or control, and therefore chances for an unlucky split are higher)[48].

For the ease of presentation and interpretation, let $L_j = 1$ if the student j received the pretest after the DAD module (domain) of the treatment (Units 1 and 2) was introduced and $L_j = 0$ otherwise. As we pointed out earlier, the subgroup that started the treatment curriculum before the pretest cannot be ignored since this may confound results in two ways: (1) pretest score on the DAD dimension will be reflecting not only the pre-existing differences but also the "low dose" post-treatment differences; (2) since performances in domains are correlated, skills taught in DAD module of the curriculum might help students in their pretest subscores in the composite score and other domains of the construct.

Let $W_j = 1$ if the student is in the treatment group and $W_j = 0$ if the student is in the control group. Three distinct groups can be expressed as:

group A: $W_j = 1|L_j = 0$;

group B: $W_j = 1|L_j = 1$;

group C: $W_j = 0|L_j = 0$ (control group).

Using the above arguments, I investigate differences between groups A and B in the DAD domain in the next section.

---

[47] The treatment curriculum consists of the modules that instruct students on topics related to five dimensions (CHA, COS, DAD, INI, MOV) described in Section […].

[48] CRT reduces the chance of nonequivalent groups but does not eliminate it.

1.2.11.1 Effect of the "DAD instruction" vs. "no instruction"

If the DAD module of the treatment curriculum has an effect on the DAD scores, as is the aim of the curriculum, we would expect to see differences between groups A and B on the pretest most pronounced in the DAD dimension and less pronounced in the remaining (CHA, COS, INI, and MOV) dimensions. At posttest, however, we would expect to see no significant differences between groups A and B (assuming these groups are balanced in all the ways that matter) in all dimensions since both groups A and B receive the same treatment by the time they take the posttest. Comparison of groups A and B at pretest (using both PCM and CRM models) and accounting for clustering at the teacher level is shown in Table 1.10 below.

Table 1.10. Difference at pretest between groups B and A (reference group).

domain	model	treatment effect	z-value	p-value
DAD	PCM	0.50 (0.12)	4.12	0.00
	CRM	0.90 (0.23)	3.93	0.00
CHA	PCM	0.09 (0.23)	0.38	0.70
	CRM	0.15 (0.39)	0.38	0.70
COS	PCM	0.26 (0.37)	0.70	0.48
	CRM	0.41 (0.54)	0.76	0.45
MOV	PCM	0.01 (0.23)	0.05	0.96
	CRM	0.05 (0.33)	0.15	0.88
INI	PCM	0.24 (0.11)	2.14	0.03
	CRM	0.59 (0.26)	2.26	0.02

Note: Z-values are shown to demonstrate the magnitude of the difference between the groups. Comparison of groups is based on latent regression and includes a random effect for the teacher level (to account for clustering).

As we see from the Table 1.10, differences between groups A and B at pretest are most pronounced in the DAD dimension (group B is significantly higher in DAD than group A at 0.01 level). Differences in INI dimension are also significant at 0.05 level. Note that the INI dimension has the highest correlation with the DAD dimension among all other dimensions—correlation is 0.91 as reported in Table 1.4 of the Section 1.2.2 (and 0.90 as reported in Schwartz, 2012). Assuming that groups A and B are comparable (i.e., formed at random), this finding serves as an evidence of the effect of the DAD module of the treatment compared to no instruction at all.

Comparison of groups A and B at posttest, accounting for clustering at the classroom level, are shown in Table 1.11 below (using both PCM and CRM models).

At posttest, differences between groups A and B are not significant for any of the dimensions. Notice that the difference in DAD and INI dimensions found at pretest does not persist at posttest.

Table 1.11. Difference at posttest between groups B and A (reference group).

module	model	treatment effect	z-value	p-value
DAD	PCM	0.00 (0.10)	0.04	0.97
	CRM	0.02 (0.26)	0.09	0.93
CHA	PCM	-0.06 (0.09)	-0.66	0.51
	CRM	-0.02 (0.18)	-0.10	0.92
COS	PCM	0.10 (0.15)	0.66	0.51
	CRM	0.19 (0.29)	0.65	0.51
MOV	PCM	0.39 (0.26)	1.54	0.12
	CRM	0.66 (0.42)	1.59	0.11
INI	PCM	0.02 (0.14)	0.16	0.87
	CRM	0.08 (0.26)	0.32	0.75

Note: Z-values are shown to demonstrate the magnitude of the difference between the groups. Comparison of groups is based on latent regression and includes a random effect for the teacher level (to account for clustering).

The research question investigated above can also be approached from the following perspective: if the treatment has an effect, we would expect to see higher gains (from pretest to posttest) in group A's DAD scores compared to group B's DAD scores (since group B's DAD scores are higher at the pretest and thus should not increase substantively from the pretest to posttest. If there are any gains in the DAD dimension for the group B, this may be attributed to the effect of other modules on the DAD score due to positive correlation. Results of this comparison are shown in the Table 1.12 below.

Table 1.12. Difference in gains between groups B and A (reference group) in the DAD domain using PCM and CRM models. Comparison of groups is based on latent regression.

	treatment effect	z-value	p-value
PCM	-0.49 (0.09)	-5.60	0.00
CRM	-0.83(0.18)	-4.65	0.00

Note: Comparison of groups is based on latent regression.

As we see in the Table 1.12, group A is significantly higher in gains from pretest to posttest in DAD domain than group B.

Similarly, due to the positive correlation between dimensions, we would expect to see slightly higher gains in CHA, COS, INI and MOV in group A compared to gains in group B. Additional evidence would be if gains for the group B on dimensions that correlate with the DAD domain the highest were found to be smaller than gains on dimensions that correlate with the DAD domain the lowest. This hypothesis is supported by summaries in Table 1.13 below.

Table 1.13. The difference in gains between groups B and A (reference group) in CHA, COS, MOV and INI domains using PCM and CRM models. Comparison of groups is based on latent regression.

		treatment effect	z-value	p-value
CHA	PCM	-0.27 (0.12)	-2.22	0.03
	CRM	-0.40 (0.22)	-1.83	0.07
COS	PCM	-0.27 (0.19)	-1.43	0.15
	CRM	-0.37 (0.28)	-1.31	0.19
MOV	PCM	0.32 (0.16)	1.95	0.05
	CRM	0.57 (0.25)	2.32	0.02
INI	PCM	-0.22 (0.09)	-2.32	0.02
	CRM	-0.51 (0.19)	-2.65	0.01

As we see in the Table 1.13, the gains for dimensions CHA, COS, and INI are higher for the group A (reference group) and gains in INI are significantly higher in group A at 0.05. Surprisingly, gains in MOV dimension are higher in the group B (compared to group A), significant at 0.05 level. Note that MOV has the lowest correlation with DAD (see Table 1.4 in Section 1.2.2). Also recall that the INI dimension has the highest correlation with the DAD dimension.

1.2.11.2 Full-dose vs. partial-dose

We can compare gains of the group B with gains of group A to obtain the effect of "non-DAD modules" (partial dose of the treatment) vs. entire curriculum (full dose of the treatment) on the composite score. The composite score is obtained by loading all items on a single dimension. Alternatively, the composite score can be obtained by excluding the DAD items. This can be helpful for the following comparison: we would expect the difference in gains in the composite score to be more apparent when the DAD items are included, compared to the comparison of gains in the composite score when the DAD items are excluded. This is because the effect of the partial dose (all modules except DAD) shouldn't be substantively different from the full dose when the outcome variable is the composite score that does not include the DAD items. However, when the DAD items are included in the estimation of the composite score, the "full dose" group (group A) should have a more apparent advantage in gains compared to the "partial dose" (group B). This hypothesis is confirmed in the Table 1.14 below.

Table 1.14. Difference in gains between group B and group A (reference group) in composite domains using PCM and CRM models. Comparison of groups is based on latent regression.

		treatment effect	z-value	p-value
composite (all items)	PCM	-0.21 (0.06)	-3.24	0.00
	CRM	-0.24 (0.13)	-1.81	0.07
composite (DAD items excluded)	PCM	-0.11 (0.08)	-1.32	0.19
	CRM	-0.13 (0.15)	-0.83	0.40

As we see from the table above, the group A (full-dose) has higher gains than the group B (partial-dose) when the composite score includes the DAD items, and this difference is significant at the 0.01 level for the PCM and at 0.1 level for the CRM models. Also note that the difference between groups A and B is smaller (and not significant) when the comparison is made on the composite score that does not include the DAD items.

1.2.12 Comparison of the treatment groups with the control group: does the treatment have an effect?

In this section, I present results from the comparison of groups A, B and C for both the change score (CS) and regressor variable (RV) approaches. For all of the comparisons, I present estimated treatment effects when we: (1) ignore the measurement error in the scores (regular regression of EAP scores on a treatment dummy), and (2) attempt to account for the measurement error in these scores—by using latent regression.

I include the results that ignore the measurement error to demonstrate the differences or similarities of CS and RV approaches across types of regression methods (regular vs. latent regression). Interpretation of the group difference, however, is based on findings from the models that account for the measurement error (latent regression). Comparisons of these two approaches are accounting for the clustering at the teacher level. Since the ways that CS and RV approaches account for clustering are different, comparisons that ignore clustering at the teacher level are presented in Appendix A.3.

We have seen some results of comparisons between groups A and B. The main interest of the study, however, is in the comparison of the groups A (treatment group) and C (control group). However, the difference between groups B and C in gains from the pretest to posttest would give the effect of the partial treatment (effect of CHA, COS, INI, and MOV modules) of the new ADM curriculum vs. regular curriculum. If the treatment has an effect (i.e., difference in gains between groups A and C is positive and significant), then the effect of the partial treatment (difference in gains between groups B and C) would be expected to be smaller in magnitude.

We find a significant treatment effect when we compare composite scores of the group A with the control group using the CS approach. As shown in the Table 1.15 below, the treatment group is significantly higher (at 0.05) than the control group in the composite score using the CS approach for both PCM and CRM models. As we would expect, the difference between groups B and C is smaller than the difference between groups A and C (i.e., partial dose vs. full dose).

The RV approach, however, does not indicate any significant differences between the groups. Note that findings from the RV approach will be closer to the findings from the CS approach only for the outcome variables for which there is no difference at the pretest.

Table 1.15. Comparison of the treatment groups with the control group (reference group) in the composite construct using CS and RV approaches and accounting for the teacher level random effect[49].

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	-0.04 (0.08)	0.09 (0.17)	0.00 (0.07)	0.08 (0.16)
latent regression	-0.05 (0.07)	0.06 (0.22)	0.01 (0.10)	0.08 (0.20)
group A vs control				
regular regression	0.11 (0.08)	0.25 (0.16)	0.03 (0.07)	0.05 (0.15)
latent regression	0.17 (0.07)**	0.35 (0.15)**	-0.03 (0.12)	-0.04 (0.24)

*** <0.01, ** <0.05, * <0.1

Table 1.16. Comparison of groups B and A (reference group) on composite construct using the CS and the RV approaches and accounting for the teacher level random effect.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.15 (0.10)	-0.16 (0.23)	-0.03 (0.09)	0.03 (0.20)
latent regression	-0.24 (0.09)***	-0.31 (0.17)*	0.02 (0.11)	0.10 (0.22)

*** <0.01, ** <0.05, * <0.1

Table 1.16 above shows results from the comparison of groups B and A using both the CS and the RV approaches. As we see from the table, these two groups differ on gains significantly at the 0.01 level when the PCM model is used.

Tables 1.17 and 1.18 below show findings when we compare these groups on the DAD domain only. For the DAD domain, both the CS and the RV approaches indicate that the treatment does have a negative effect for the group B when we compare it to both control group (Table 1.17) and the group A (Table 1.18), which is what we would expect since the pretest of the group B may already be reflecting the effect of the training in the DAD module. We do not find a significant treatment effect on the DAD domain when we compare the group A (new curriculum) with the control group (existing curriculum).

Notice that the CS and the RV approaches give similar results when we compare groups A and C (Table 1.17) in the DAD domain. This is because groups A and C do not differ at pretest on the DAD domain.

[49] Here we account only for the teacher cluster: variance of the school level random-effect was either not significant or likelihood-ratio test preferred the simpler model (the one that accounts only for the teacher-level clustering).

Table 1.17. Comparison of the treatment groups with the control group (reference group) in the DAD domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	-0.43 (0.09)***	-0.67 (0.14)***	-0.06 (0.06)	-0.13 (0.14)
latent regression	-0.56 (0.09)***	-0.98 (0.20)***	-0.01 (0.11)	-0.05 (0.24)
group A vs control				
regular regression	-0.04 (0.09)	-0.07 (0.14)	-0.01 (0.05)	-0.02 (0.13)
latent regression	-0.04 (0.09)	-0.07 (0.19)	0.00 (0.10)	-0.01 (0.24)

*** <0.01, ** <0.05, * <0.1

Table 1.18. Comparison of groups B and A (reference group) in the DAD domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.39 (0.09)***	-0.61 (0.16)***	-0.03 (0.06)	-0.08 (0.16)
latent regression	-0.52 (0.10)***	-0.90 (0.21)***	0.00 (0.09)	0.03 (0.23)

*** <0.01, ** <0.05, * <0.1

Table 1.19. Comparison of the treatment groups with the control group (reference group) in the CHA domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	0.20 (0.15)	0.33 (0.25)	-0.04 (0.02)	-0.05 (0.04)
latent regression	0.07 (0.11)	0.12 (0.20)	-0.18 (0.11)	-0.29 (0.20)
group A vs control				
regular regression	0.29 (0.16)*	0.46 (0.27)*	-0.01 (0.02)	-0.03 (0.04)
latent regression	0.36 (0.12)***	0.54 (0.20)***	-0.09 (0.12)	-0.20 (0.22)

*** <0.01, ** <0.05, * <0.1

Table 1.20. Comparison of groups B and A (reference group) in the CHA domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.08 (0.19)	-0.12 (0.33)	-0.02 (0.02)	-0.02 (0.04)
latent regression	-0.28 (0.13)**	-0.41 (0.23)*	-0.08 (0.10)	-0.07 (0.21)

*** <0.01, ** <0.05, * <0.1

We also found a significant (at 0.01 level) effect of the treatment on CHA domain when we compare group A with the control group (Table 1.19). In addition, we found

that group A is significantly higher than group B in changes in the CHA domain, shown in Table 1.20 above.

Tables 1.21 and 1.22 below shows finding when we compare three groups on the COS domain. We found that the treatment has a positive effect when compared to the control group, significant at 0.1 level when using the CS approach. As we see in the Table 1.21, there are no significant differences between groups A and B on the COS domain.

Table 1.21. Comparison of the treatment groups with the control group (reference group) in the COS domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	0.02 (0.23)	0.08 (0.33)	0.04 (0.07)	0.07 (0.15)
latent regression	0.01 (0.18)	0.09 (0.27)	0.06 (0.13)	0.09 (0.23)
group A vs control				
regular regression	0.20 (0.22)	0.35 (0.31)	-0.01 (0.07)	-0.02 (0.15)
latent regression	0.31 (0.17)*	0.51 (0.27)*	-0.04 (0.13)	-0.07 (0.24)

*** <0.01, ** <0.05, * <0.1

Table 1.22. Comparison of groups B and A (reference group) in the COS domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.18 (0.31)	-0.25 (0.45)	0.05 (0.09)	0.09 (0.18)
latent regression	-0.30 (0.20)	-0.41 (0.31)	0.09 (0.15)	0.16 (0.28)

*** <0.01, ** <0.05, * <0.1

Tables 1.23 and 1.24 show results from the comparison of the groups on MOV domain. For the MOV domain at pretest, the mean of the group B is closer to the mean of the control group (as shown in the Table 1.8). Therefore, CS and RV approaches produce similar results when we compare the group B with the control group.

We found that the group B is significantly higher than the control group on the MOV domain, as shown in Table 1.23 below. However, we didn't find a significant difference between group A and the control group.

Table 1.23. Comparison of the treatment groups with the control group (reference group) in the MOV domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	0.28 (0.15)*	0.51 (0.24)**	0.19 (0.07)***	0.38 (0.14)***
latent regression	0.46 (0.15)***	0.78 (0.23) ***	0.37 (0.14)***	0.63 (0.23)***
group A vs control				
regular regression	0.12 (0.14)	0.18 (0.21)	0.02 (0.07)	0.01 (0.16)
latent regression	0.18 (0.16)	0.26 (0.25)	0.01 (0.16)	-0.01 (0.26)

*** <0.01, ** <0.05, * <0.1

Table 1.24. Comparison of groups B and A (reference group) in the MOV domain using the CS and the RV approaches and accounting for the teacher level random effect

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	0.15 (0.18)	0.33 (0.29)	0.17 (0.11)	0.37 (0.24)
latent regression	0.29 (0.18)	0.53 (0.30)	0.37 (0.21)*	0.65 (0.35)*

*** <0.01, ** <0.05, * <0.1

Tables 1.25 and 1.26 show comparisons of groups on the INI domain. We found that the group A is significantly higher (at 0.05 level) in gain then the control group. We also found that group A is significantly higher (at 0.05 level) in gain than group B.

Table 1.25. Comparison of the treatment groups with the control group (reference group) in the INI domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	-0.01 (0.07)	0.02 (0.12)	0.00 (0.08)	0.04 (0.14)
latent regression	-0.01 (0.09)	0.01 (0.19)	0.01 (0.11)	0.09 (0.22)
group A vs control				
regular regression	0.05 (0.06)	0.17 (0.12)**	0.02 (0.07)	0.06 (0.14)
latent regression	0.23 (0.09)**	0.57 (0.20)***	0.00 (0.11)	0.03 (0.24)

*** <0.01, ** <0.05, * <0.1

Table 1.26. Comparison of groups B and A (reference group) in the INI domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.06 (0.08)	-0.15 (0.16)	-0.02 (0.09)	-0.01 (0.17)
latent regression	-0.24 (0.11)**	-0.57 (0.22)**	0.02 (0.12)	0.06 (0.24)

*** <0.01, ** <0.05, * <0.1

Using the CS approach and accounting for the clustering at the teacher level, I found that the treatment has a significant effect on the composite score (significant at $\alpha = 0.05$ level). Mean change in the composite score of students in the treatment group was 0.17 logits (0.35 for the CRM model) higher than the mean change in the composite score of students in the control condition. Similarly, mean change for the treatment group was estimated higher in the CHA domain (significant at $\alpha = 0.01$ level); COS domain (significant at $\alpha = 0.1$ level); and INI domain (significant at $\alpha = 0.05$ level). These findings were consistent for both PCM and CRM models.

I provided arguments on why the RV approach is not reliable. In particular, I argued that the exchangeability assumption is questionnable due to: (1) imbalance between treatment and control groups on important covariate (race); and (2) dropout of clusters after learning the results of randomization.

I also speculated on differences between the three groups with respect to various outcomes and found evidences that support those conjectures.

1.3 Discussion

In the ADM study, groups A, B, and C differ on pretest and other pretreatment covariates. If we think that the lack of balance is the result of the unhappy randomization, then the RV approach is the right approach since groups are still exchangeable and the assumption of the regression to the grand mean is plausible. If the lack of the balance between groups at pretest might be a result of any other factor such as nonignorable dropouts and any form of selection to/out of the treatment, then the RV approach might be misleading.

The purpose of almost every evaluation study, including this two-wave design discussed in this chapter, is to make a causal statement. For the causal statement, the dilemma of comparing oranges with apples is the most important one to consider. The only method that guarantees comparability of groups is randomization—the greatest contribution of Ronald Fisher. Sir Ronald Fisher, the man who single-handedly established the foundations of modern statistics, consistently claimed that "*you cannot prove anything without randomized experimental design*" (Salsburg, 2002).

There is a price that needs to be paid when there is no randomization, but that price needs to be made as small as possible. That price can be expressed in terms of the assumptions: we want a minimum number of assumptions and we need to avoid "heroic" assumptions. The justification for the assumption should be as complete as possible (and assumptions need to be laid out clearly in the first place). The decision on exchangeability should be justified (and the research community should find it acceptable).

By assessing balance on available pre-treatment covariates in the ADM study, I showed that the assumption of exchangeability is questionable, and therefore the RV approach is not reliable. While it is difficult to provide evidences for the "nonexchangeability" (so that we can discard the possibility of the regression to the mean), the CS approach is still the preferred method among the two when exchangeability is questionnable. Conditioning on covariates on which the groups are not balanced can be another alternative and I didn't discuss implication of additional

covariates in RV and CS approaches in order not to dilute the attention from the central argument in the paradox.

I attempted to explain the Lord's paradox and provide guidelines on choosing between the two approaches. One important message of the chapter is that measurement error is not the source of the paradox. Second finding/message is that we have "measurement" analogues of the CS and RV approaches in the IRT literature, and same guidelines apply in the IRT framework. The third finding is that, when the groups being investigated are not results of randomization, the pre-treatment balance on important covariates require careful investigation and judgment on the plausibility of the exchangeability assumption. "Regression to the grand mean" or "regression to the group-specific means" are two important considerations in selecting among the approaches. I presented arguments why the CS approach relies on "nonexchangeability" with respect to response patterns when groups are not balanced at pre-treatment.

In summary, when two groups are not formed by the means of randomization, researcher needs to justify and provide evidences for: (1) the assumption of exchangeability if the RV approach is to be used; (2) the assumption of nonexchangeability when groups are not balanced on pretreatment covariates (to make sure that lack of balance is not due to the "unhappy randomization") if the CS approach is to be used, especially when the number treatment-level units (e.g., number of schools in cluster-randomized trials) is small.

Limitations of the study. This chapter did not provide a survey of which types of evidences need to be provided for exchangeability assumption (e.g., tests of means, medians, or distributions of relevant pretreatment covariates between the two groups). This study also fell short of providing alternative approaches to RV and CS—these two are the two simplest ones (it might be the case that more elaborate methods are necessary). The comparison of apples with oranges is definitely not a good idea, but if one still needs to compare, the RV approach is the tricky one. In such cases, the CS approach, which assumes that the change would have been the same in both groups had both received the treatment, is preferred. For the case with clustered data, this chapter didn't focus on implications of exchangeability when randomization is at the person level vs. cluster level.

Appendix A.1: First-differencing (CS approach) and Rasch model (with CML)

Rasch model with CML estimation procedure is an analog of the differencing estimator (see for instance Skrondal and Rabe-Hesketh, 2007). In particular:

$$P\left(y_{ij}=1\middle|x_{1j},\dots x_{Ij},c_j\right)=\frac{\exp(\beta x_{ij}+c_j)}{1+\exp(\beta x_{ij}+c_j)}$$

where y_{ij} are independent conditional on c_j and x_{ij} are item dummy indicators. Consider two items (y_{1j} and y_{2j}):

$$P\left(y_{1j}=0,y_{2j}=1\middle|x_1,x_2,c_j,y_{1j}+y_{2j}=1\right)^{50}$$

$$=\frac{\Pr\left(y_{1j}=0,y_{2j}=1\right)}{\Pr\left(y_{1j}=0,y_{2j}=1\right)+\Pr\left(y_{1j}=1,y_{2j}=0\right)}$$

$$=\frac{\frac{1}{1+\exp\left(\beta x_{1j}+c_j\right)}*\frac{\exp\left(\beta x_{2j}+c_j\right)}{1+\exp\left(\beta x_{2j}+c_j\right)}}{\frac{1}{1+\exp\left(\beta x_{1j}+c_j\right)}*\frac{\exp\left(\beta x_{2j}+c_j\right)}{1+\exp\left(\beta x_{2j}+c_j\right)}+\frac{\exp\left(\beta x_{1j}+c_j\right)}{1+\exp\left(\beta x_{1j}+c_j\right)}*\frac{1}{1+\exp\left(\beta x_{2j}+c_j\right)}}$$

$$=\frac{\frac{\exp\left(\beta x_{2j}+c_j\right)}{\left[1+\exp\left(\beta x_{1j}+c_j\right)\right]*\left[1+\exp\left(\beta x_{2j}+c_j\right)\right]}}{\frac{\exp\left(\beta x_{2j}+c_j\right)+\exp\left(\beta x_{1j}+c_j\right)}{\left[1+\exp\left(\beta x_{2j}+c_j\right)\right]*\left[1+\exp\left(\beta x_{1j}+c_j\right)\right]}}$$

$$=\frac{\exp\left(\beta x_{2j}+c_j\right)}{\left[\exp\left(\beta x_{2j}+c_j\right)+\exp\left(\beta x_{1j}+c_j\right)\right.}=\frac{\exp\left(\beta x_{2j}\right)}{\left[\exp\left(\beta x_{2j}\right)+\exp\left(\beta x_{1j}\right)\right]}$$

$$=\frac{\frac{\exp\left(\beta x_{2j}\right)}{\exp\left(\beta x_{1j}\right)}}{\left[\exp\left(\beta x_{2j}\right)+\exp\left(\beta x_{1j}\right)\right]/\exp\left(\beta x_{1j}\right)}=\frac{\exp\left[\beta\left(x_{2j}-x_{1j}\right)\right]}{1+\exp\left[\beta\left(x_{2j}-x_{1j}\right)\right]}$$

$$=G[\beta(x_{2j}-x_{1j})]$$

in which

$$G[w]=\exp(w)/(1+\exp(w))$$

[50] With the only two items, condition $y_{1j}+y_{2j}=1$ is necessary for identification.

Appendix A.2: ADM Constructs

Data Display (DAD): DAD domain represents students' ability to read and interpret graphical representation of the data with focus on reasoning related to the properties of aggregate.

Conceptions of Statistics (CoS): COS domain represents students' ability to recognize that statistics are summary measures of data that are developed to answer research questions about distributions' measures of central tendency and dispersion that reflects the sample-to-sample variation.

Chance (Cha): CHA domain represents students' ability to understand that concepts such as chance, probability, and uncertainty are related to produce distributions of outcomes.

The Models of Variability (MOV): MOV domain represents students' ability to reason about the role and importance of chance to model a distribution of measurements and observations.

Informal Inference (InI): INI domain represents students ability to make an inference based on outcomes obtained from single or multiple samples.

Appendix A.3: Comparison of three groups using CS and RV approaches and ignoring the clustering of students.

Table 1.13b. Comparison of the treatment groups with the control group (reference group) in the composite construct using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	-0.05 (0.05)	0.07 (0.09)	0.00 (0.03)	0.10 (0.07)
latent regression	-0.04 (0.04)	0.08 (0.10)	0.00 (0.04)	0.11 (0.08)
group A vs control				
regular regression	0.14 (0.05) ***	0.28 (0.09) ***	0.03 (0.03)	0.06 (0.07)
latent regression	0.15 (0.05) ***	0.32 (0.10) ***	0.04 (0.04)	0.07 (0.08)

*** <0.01, ** <0.05, * <0.1

Table 1.14b. Comparison of groups B and A (reference group) in the composite construct using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.18 (0.06)***	-0.21 (0.12)*	-0.03 (0.05)	0.05 (0.09)
latent regression	-0.19 (0.06)***	-0.23 (0.12)**	0.00 (0.04)	0.05 (0.11)

*** <0.01, ** <0.05, * <0.1

Table 1.15b. Comparison of the treatment groups with the control group (reference group) in the DAD domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	-0.43 (0.05)***	-0.66 (0.10)***	-0.07 (0.03)***	-0.15 (0.07)**
latent regression	-0.56 (0.07)***	-0.95 (0.14)***	0.02 (0.04)***	0.07 (0.12)**
group A vs control				
regular regression	-0.03 (0.05)	-0.06 (0.10)	-0.02 (0.03)	-0.05 (0.06)
latent regression	-0.05 (0.08)	-0.10 (0.14)	-0.02 (0.04)	-0.08 (0.11)

*** <0.01, ** <0.05, * <0.1

Table 1.16b. Comparison of groups B and A (reference group) in the DAD domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.40 (0.07)***	-0.60 (0.12)***	-0.02 (0.03)	-0.04 (0.08)
latent regression	-0.49 (0.09)***	-0.83 (0.18)***	0.05 (0.05)	0.14 (0.14)

*** <0.01, ** <0.05, * <0.1

Table 1.17b. Comparison of the treatment groupswith the control group (reference group) in the CHA domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	0.16 (0.08)**	0.25 (0.13)*	-0.04 (0.01)**	-0.05 (0.03)**
latent regression	0.08 (0.10)	0.12 (0.17)	-0.17 (0.06)***	-0.28 (0.10)***
group A vs control				
regular regression	0.34 (0.08)***	0.56 (0.14)***	-0.01 (0.02)	-0.03 (0.03)
latent regression	0.35 (0.11)***	0.53 (0.18)***	-0.13 (0.06)**	-0.27 (0.11)**

*** <0.01, ** <0.05, * <0.1

Table 1.18b. Comparison of groups B and A (reference group) in the CHA domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.18 (0.09)**	-0.30 (0.16)*	-0.02 (0.02)	-0.02 (0.03)
latent regression	-0.27 (0.12)**	-0.40 (0.22)*	-0.05 (0.07)	-0.00 (0.14)

*** <0.01, ** <0.05, * <0.1

Table 1.19b. Comparison of the treatment groups with the control group (reference group) in the COS domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
groupB vs control				
regular regression	0.00 (0.14)	0.06 (0.21)	0.04 (0.03)	0.08 (0.07)
latent regression	0.02 (0.17)	0.10 (0.25)	0.08 (0.06)	0.13 (0.11)
group A vs control				
regular regression	0.25 (0.13)*	0.41 (0.20)**	-0.02 (0.03)	-0.04 (0.07)
latent regression	0.29 (0.16)*	0.47 (0.24)**	-0.07 (0.06)	-0.15 (0.12)

*** <0.01, ** <0.05, * <0.1

Table 1.20b. Comparison of groups B and A (reference group) in the COS domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs groupA				
regular regression	-0.26 (0.16)	-0.35 (0.23)	0.06 (0.04)	0.12 (0.08)
latent regression	-0.27 (0.20)	-0.37 (0.28)	0.15 (0.07)**	0.28 (0.14)**

*** <0.01, ** <0.05, * <0.1

Table 1.21b. Comparison of the treatment groups with the control group (reference group) in the MOV domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	0.25 (0.09)***	0.48 (0.14)***	0.20 (0.04)***	0.40 (0.08)***
latent regression	0.49 (0.14)***	0.80 (0.21)***	0.41 (0.08)***	0.69 (0.14)***
group A vs control				
regular regression	0.14 (0.09)	0.20 (0.14)	0.01 (0.04)	0.00 (0.08)
latent regression	0.20 (0.14)	0.25 (0.21)	0.01 (0.09)	-0.05 (0.15)

*** <0.01, ** <0.05, * <0.1

Table 1.22b. Comparison of groups B and A (reference group) in the MOV domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	0.11 (0.11)	0.28 (0.17)*	0.18 (0.05)***	0.39 (0.10)***
latent regression	0.32 (0.16)**	0.57 (0.25)**	0.43 (0.12)***	0.76 (0.20)***

*** <0.01, ** <0.05, * <0.1

Table 1.23b. Comparison of the treatment groups with the control group (reference group) in the INI domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs control				
regular regression	0.01 (0.04)	0.05 (0.07)	0.01 (0.04)	0.07 (0.07)
latent regression	0.00 (0.07)	0.04 (0.15)	0.03 (0.06)	0.13 (0.12)
group A vs control				
regular regression	0.05 (0.04)	0.18 (0.08)**	0.03 (0.04)	0.06 (0.07)
latent regression	0.21 (0.07)***	0.54 (0.16)***	-0.04 (0.06)	-0.05 (0.12)

*** <0.01, ** <0.05, * <0.1

Table 1.24b. Comparison of groups B and A (reference group) in the INI domain using the CS and the RV approaches.

	CS approach		RV approach	
	PCM	CRM	PCM	CRM
group B vs group A				
regular regression	-0.04 (0.05)	-0.13 (0.09)	0.00 (0.04)	0.03 (0.09)
latent regression	-0.22 (0.09)**	-0.51 (0.19)***	0.08 (0.07)	0.19 (0.14)

*** <0.01, ** <0.05, * <0.1

Appendix A.4: A note on clustering

Measures taken from students at two time-points are nested within students, while students in the study are nested within classrooms, which in turn are nested within schools and schools are nested within districts. There can be variability between and within clusters (students, classrooms, schools, district) at each level, due to (1) random variation among units nested within different clusters at each level, (2) non-random variation among units nested within different clusters, including factors that are not correlated with the treatment, and factors that are associated with the treatment.

Dependence induced by clustering can be accounted for by using random-effects or fixed-effects approaches. However, the only cluster that we can account for using the fixed-effects approach is a clustering at the district level. This is due to the fact that the treatment indicator is at the school-level, and thus does not vary within schools or classrooms: this makes the treatment indicator perfectly collinear with classroom or school dummies. Therefore, for classrooms and schools, random effects need to be included if we want to account for clustering. Note that this necessity to account for clustering does not apply to the CS approach described earlier (see Section 1.1.4).

The choice of random effect vs. fixed effect also relates to the inferences we are attempting to make. In particular, if the interest lies in the inference for all schools from the relevant population of schools, then a random-effects approach is necessary (which requires assumptions regarding the distribution of these random-effects).

Using the CS approach, however, we eliminate random-effects and avoid additional assumptions. The only assumption we make when using the CS approach is, however, that the random-effects are independent of the dynamic factors between two time-points. In other words, we assume that the coefficient of the time dummy variable in Equation 9 does not vary randomly over clusters.

Before applying the first-differencing operator (CS approach), it is worth specifying the full model with all the clustering detailed, and discussing how dependence induced by clustering is handled for each level separately. The fully random-effects approach to this nested structure results in a five-level model. These five levels are: two occasions (level-1) nested in students (level-2), which in turn are nested in classrooms (level-3), which are nested in schools (level-4), which are finally nested in districts (level-5). The equation for this model specification is shown below in Equation 61:

$$y_{ijklm} = \beta_1 + \beta_2 W_{1ilm} + \beta_3 a_{1ijklm} + \zeta^{(2)}_{1jklm} + \zeta^{(3)}_{1klm} + \zeta^{(4)}_{1lm} + \zeta^{(5)}_{1m} + \epsilon_{ijklm}. \qquad \textbf{(61)}$$

In the Equation 61, the elements are:

Level 1: (occasion), i
y_{ijklm}: ADM test,
a_{1ijklm}: occasion (1, 2)
ϵ_{ijklm}: occasion specific random error

Level 2: (student), j
$\zeta^{(2)}_{1jklm}$: person-specific random effect

Level 3: (classroom), k

$\zeta^{(3)}_{1klm}$: teacher (classroom) specific random effect

Level 4: (school), l
$\zeta^{(4)}_{1lm}$: school-specific random effect
W_{1ilm}: treatment dummy

Level 5: (district), m
$\zeta^{(5)}_{1m}$: district-specific random effect

The model above assumes that all random effects are normally distributed with zero mean, and are independent of each other and of the remaining terms in the model. In Section 1.1.4, I demonstrated that the CS approach is equivalent to first-differencing the above model and eliminates terms that are not varying across occasions (all terms without the i subscript).

Prior to employing the multilevel modeling, however, one should consider the number of clusters and the number of units within clusters to evaluate the appropriateness and feasibility of specifying random vs. fixed effects. With only four districts, for instance, it is not feasible to specify a random effect for the district and thus it is imperative to include district dummies if one wishes to account for the district membership. Note, however, if the random assignment of the treatment to schools was blocked at the district level, such inclusion of the district dummies may not be necessary after all.

For the analysis that includes groups A and C, the number of students is 595, and the number of classrooms is 31, and the number of schools is 17. Schools cannot be included as fixed effects (as dummies) since the treatment (main variable of interest) is at the school level. However, even if the treatment were assigned within schools, an option to include schools as fixed-effects would not be feasible option due to small number of classrooms within each school.[51]The model shown in Equation 61 is not estimated due to the restrictions discussed above and since it is not the main focus of the chapter.

Since the treatment effect is at the school-level, standard error for the treatment dummy is important for the inference. Raudenbush (1997) discusses the degree to which the standard errors of the experimental condition coefficient can be decreased (and thus power increased) when an added covariate is strongly related to the outcome variable. Note that in a multilevel approach, the standard error of the treatment coefficient depends more strongly on the number of clusters, rather than the cluster sizes (Raudenbush, 1997).

[51] See Rabe-Hesketh & Skrondal, 2012, p. 159, for suggested cluster sizes and number of clusters for each random-effects and fixed-effects approaches.

Chapter 2

Data expansion for ordinal modeling

Ordinal response data obtained from surveys and tests are often modeled using cumulative, adjacent-category, or continuation-ratio logit link functions. Instead of using one of these specifically designed procedures for each of these formulations of logits, we can modify the structure of the data in such a way that methods designed for dichotomous outcomes (i.e., binary logistic regression) allow us to achieve the targeted polytomous contrasting (cumulative, adjacent-category, or continuation-ratio). Thus, one can implement procedures designed for dichotomous outcomes on appropriately expanded data. The techniques presented in this chapter, which we refer to as *data expansion techniques*, represent this approach.

In the psychometrics literature, little is known about using data expansion for fixed and random effects estimators when applied to polytomous items. Data expansion provides a practical solution for modeling ordinal data without using specialized statistical packages. This is particularly important for using complex latent variable models for which software for the ordinal responses may not even exist. In addition, data expansion techniques have not been investigated for the adjacent-category logit model—a model that is widely used in psychometrics.

In the case of multiple binary responses from each subject, consistent estimates can be obtained by conditioning on the sufficient statistics without making any distributional assumptions about subject-specific effects. In most polytomous IRT models for ordinal responses, sufficient statistics do not exist. Therefore, another advantage of data expansion techniques is that a version of conditional maximum likelihood estimation can be used by applying conditional logistic regression on the subsets of the data.

We briefly review the biostatistics and econometrics literature on data expansion methods for the continuation-ratio and cumulative logit models, along with their analogous counterparts in psychometrics, for both fixed and random effects approaches. We then investigate the potential of data expansion for the adjacent-category logit model and demonstrate that this works with a fixed-effects approach only.

We also provide an explanation as to why data expansion for the adjacent-category logits does not work for the random-effects approach. We demonstrate consistency of the methods that should work theoretically by applying the estimator to "population data" that can be thought of as datasets with infinite sample size. Lastly, using the example dataset, we present a comparison of methods based on data expansion techniques to methods specifically designed to ordinal data.

2.1 Introduction

Data collected from psychological, sociological, and marketing surveys and educational tests are often categorical in nature and are therefore coded either dichotomously or polytomously. Often, the primary purpose of analyzing such data is to estimate the relationship between response probabilities and observed or latent variables.

Variables can be either nominal (unordered) or ordinal. Nominal variables do not have any inherent ordering in the levels of the variable, such as political party affiliations

or blood type. Ordinal response variables—the focus of this chapter—as the name implies, arise from natural ordinal ranking or have explicit ordering in its levels (from smallest to largest). Nominal variables are sometimes referred as qualitative variables, and from that perspective, ordinal models are considered quantitative, although they are somewhere in between qualitative and metric variables.

Dichotomous variables have exactly two levels: for instance, *success/failure*, or *pass/fail*. Polytomous variables have three or more levels: for instance, performance on a task or on an essay in an educational test might be scored as 0, 1 and 2. Such scoring of constructed-response items is common in educational assessments as well as other fields using survey-based data collection.

Rating scales are another example of how ordinal variables arise. For instance, a respondent in a survey might be asked to choose from options such as s*trongly disagree/disagree/agree/strongly agree* to express (self-reported) agreement with a particular statement. Response types of this sort are known as Likert scores (Likert, 1932). They are mostly used to measure attitudes, preferences, or opinions in scales that contain multiple statements and have no "correct" response/answer.

Developments in the modeling of ordinal variables, particularly thanks to the contributions of Goodman (1979), Nerlove & Press (1973), Bishop, Feinberg & Holland (1975), McCullagh (1980), Agresti (1984), made these models accessible. On a parallel front, polytomous item-response models gained popularity in psychometrics thanks to contributions by Samejima (1969), Bock (1972), Andrich (1978), Masters (1982), Tutz (1990), Wilson (1992), and Muraki (1992), to name a few.

In the behavioral and social sciences, researchers often collapse polytomous outcomes into two categories to avoid additional burdens in modeling and interpretation. Indeed, in some cases a lack of software that handle ordinal response variables makes this necessary. However, such practices result in a loss of information. Armstrong & Sloan (1989) report a ~25% loss in efficiency when trichotomous variables are dichotomized, and 25–50% loss in efficiency when a five-category ordinal variable is dichotomized. In addition, collapsing multiple categories into two categories adds unnecessary subjectivity in choosing the cut-point for dichotomization.

At the other extreme, ordinal variables are sometimes treated as unordered. However, there is also a loss in interpretation when methods designed for nominal variables are applied to ordinal variables (without any constraints). It is important to preserve the ordinal nature of the variables when modeling to avoid: (1) loss of power to detect significant relationship between variables; (2) maintain parsimony in interpreting results; and (3) avoid many parameters (as with nominal variables; Agresti, 2015, p.209).

Perhaps the most common shortcut is to treat ordinal variables as continuous variables. Such approaches will give misleading answers (Long, 1997). Indeed, a linear modeling approach to categorical dependent variable: (1) is not efficient in the statistical sense and (2) does not justify distributional assumptions of the linear model. However, most importantly, moment structure hypotheses are violated resulting in inconsistent estimates (Bollen & Curran, 2006, p. 231).

In a multivariate setting (when multiple ordinal variables are collected from a single subject), a common practice is to assume multivariate normality for responses and proceed with standard factor analysis techniques. However, such an approach may lead to biased estimates, incorrect standard errors, and incorrect goodness of fit tests (Moustaki,

2007).

In spite of all of these arguments against analyzing ordinal variables as if they were interval-scale variables, this is still the most commonly used approach in the social sciences. This paper proposes methods that are easy to implement and preferable to the shortcuts described above.

2.1.1 Methods for ordinal variables

Methods applied to dichotomous variables are always based on contrasting one level with another, such as *dead vs. alive*, or *correct vs. incorrect*. Ordinal variables, however, are focused on mainly three different formulations of contrasts (or coding schemes) that utilize the ordering of the variables (Agresti, 1984).

One method is based on contrasting higher level(s) with the remaining lower level(s). For instance, for the trichotomous variable with levels of poor/average/good, two possible contrasts (preserving the strict ordering of categories) are *poor vs. average/good,* and *poor/average vs. good*. Parameters obtained from these contrasts are known as *thresholds*. This type of contrast is used in cumulative logit[52] models (McCullagh, 1980). The structure of this contrast is shown in the second panel of Figure 2.1 where bold lines between categories represent the cut-points for the contrasts. The cumulative logit model has also some unique properties, which is discussed later in the chapter and articulated in Appendix B.4. Thresholds for this model must be strictly ordered.

A second alternative is to contrast each individual level of the variable with the next higher category. For instance, for categories poor/average/good, two contrasts are *poor vs. average*, and *average vs. good*. Models using this type of contrasting are mainly known as an adjacent-category logit models (Goodman, 1983) and are shown in the second panel of Figure 2.1. Adjacent-category logit models can be expressed as log-linear models and can also be estimated by fitting constrained nominal (or baseline category) models (Adams, Wilson, & Wang, 1997).

A third approach, mostly used when it is believed that the occurrence of the levels is the result of a sequential (or underlying stage-like) process, is to contrast each individual level with the level(s) above. For example, with three ordered categories poor/average/good, two possible contrasts are, *poor vs. average/good* and *average vs. good*. This type of contrasting is known as continuation-ratio model (Feinberg, 1980; McCullagh & Nelder, 1983) and is shown in the third window of Figure 2.1. This model is identical to Cox's proportional logit hazard's model (Cox, 1972) for survival data in discrete time[53].

Notice that for the particular ordinal variable, numbers of comparisons (number of threshold parameters—bold vertical lines between two ordered categories in Figure 2.1) are the same for all three types of contrasts; the interpretation, however, differs.[54]

---

[52] There is also cumulative probit (Aitchison & Silvey, 1957; McKelvey & Zavoina, 1975) model, however we limit our discussion only to models using the logit link.

[53] When the complementary log-log link is used, this is proportional hazards model. Use of clog-log link does not require data expansion. One then obtains hazard ratios instead of odds ratios.

[54] For models that deal with multinomial (a.k.a. baseline, nominal) variables see Bradley & Terry (1952), Luce (1959), McFadden (1974), a.k.a. conditional logit or discrete-choice models.

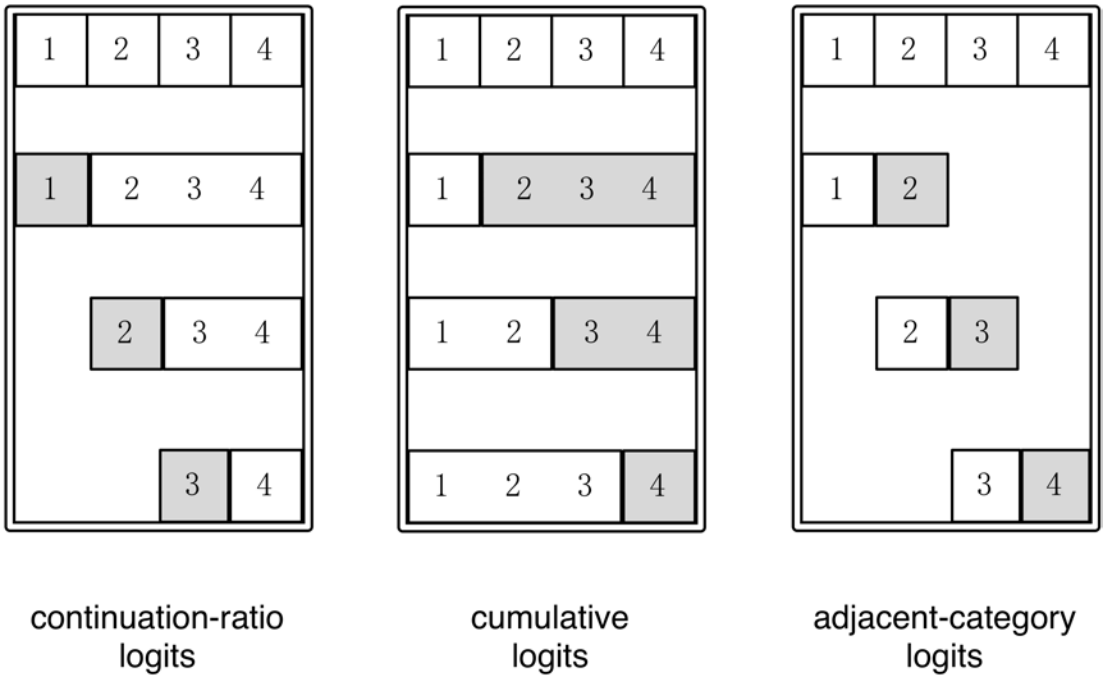


Figure 2.1. Ordinal contrasts for the four-category ordinal variable. Numerators of the expressions for each of the model are highlighted.

2.1.2 Multiple ordinal responses

When multiple ordinal responses are obtained from a particular individual, within-person dependence needs to be accounted for. In item response theory, independence of responses is assumed to be conditional on the continuous person-specific effect, usually denoted as θ and interpreted as "ability".[55]

For logit models for multiple binary responses from each subject, consistent estimates can be obtained without making any distributional assumptions about θ by conditioning on the total score. This is achieved by conditional maximum likelihood (CML), which conditions on the total score (sufficient statistic) to avoid the incidental parameters problem[56](Rasch, 1960, Breslow & Day, 1980).

Unlike in marginal maximum likelihood (MML), in CML, subject-specific effects (θ) are considered parameters (and not random variables), but are conditioned out of the likelihood function instead of using joint estimation. In contrast, marginal ML method, the most commonly used method in IRT, assumes that person-specific parameters are random variables with a particular distribution. The marginal likelihood has no closed form solution and involves computing intractable integrals. See Skrondal and Rabe-Hesketh (2004, chapter 6) for a detailed explanation.

There are also nonparametric approaches for the analysis of ordinal variables. For

[55] When categories are ordered, higher score reflects higher ability and lower score reflects lower ability. This is the requirement of the monotonicity assumption in IRT. For the results to be meaningful, we need to make sure that ordered categories are in the same direction regardless of the mixture of positively and negatively worded items.

[56] See Andersen (1970) and Pfanzagl (1993) on consistency of the CML and Neyman and Scott (1948) for incidental parameter problem.

nonparametric approaches when the dependent variable is at the ordinal level, see Weinberg & Goldberg, (1990, chapter 19) for univariate responses; Cliff (1996), Cliff & Keats (2013) for multivariate responses; Sijtsma & Molenaar (2002, chapters 7 and 8) for nonparametric polytomous IRT methods. For log-linear approaches to categorical data—which are closely related to adjacent category logit, as we will elaborate on later—see Haberman (1974), Bishop, Feinberg, & Holland (1975), Goodman (1985), Clogg & Shihadeh (1994), Bergsma & Croon (2005). Lastly, see Anderson & Yu (2007) for connections of log-multiplicative models with IRT models.

Thissen & Steinberg (1986) provide a taxonomy of polytomous IRT models that might help one to understand the focus of ordinal IRT models and the underlying response process they attempt to characterize. See also Nering & Ostini (2010), van der Linden & Hambleton (1997) and Ostini & Nering (2006) for more detailed treatment of polytomous IRT models.

2.2 Models for ordinal responses

Let the response variable y_i have *S* ordered categories, $s = 1 \ldots S$. Category-specific odds (defined differently for the different models described in Section 2.1.2) can be modeled as $H_s = \exp(\alpha_s + \boldsymbol{x}_i'\boldsymbol{\beta})$ where α_s is category-specific intercept, $\boldsymbol{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a set of regression coefficients assumed to be constant across the categories for each covariate i and $\boldsymbol{x}_i'\boldsymbol{\beta}$ implies that covariate effects are additive.[57] Define I_s as the inverse logit transformation of the linear predictor $\alpha_s + \boldsymbol{x}_i'\boldsymbol{\beta}$, so that

$$H_s = \frac{I_s}{1-I_s}. \quad \textbf{(1)}$$

To keep the notation simple, we drop the i subscript from y_i and do not show the conditioning on $\boldsymbol{x_i}$.

2.2.1 Continuation-ratio logit model

The continuation-ratio odds are modeled as

$$\frac{P(y=s)}{P(y>s)} = H_s, \;\; s < S. \quad \textbf{(2)}$$

It follows that the conditional response probability that y equals s, given that y is at least s is:

$$P(y = s|y \geq s) = \frac{H_s}{1+H_s} \equiv I_s, \;\; s < S, \quad \textbf{(3)}$$

where I_s can also be interpreted as the conditional probability of not progressing beyond level s. When categories represent discrete survival times, I_s is called a discrete hazard. The unconditional response probabilities are given by

[57] The assumption that these coefficients are constant across categories for a particular covariate can be relaxed with $\exp(\alpha_s + \boldsymbol{\beta_s}\boldsymbol{x_i'})$. However this saturated model will result in more parameters. See Fullerton (2009) for a general framework and Fullerton & Xu (2012) for compromise between fully constrained and unconstrained formulations.

$$P(y=s) = P(y=s|y \geq s)\underbrace{P(y \neq 1)P(y \neq 2|y \geq 2) \ldots P(y \neq s-1|y \geq s-1)}_{P(y \geq s)}, \textbf{(4)}$$

and can be expressed as [58]

$$P(y=s) = I_s \prod_r^{s-1}(1-I_r), \;\; D_S = 1. \textbf{(5)}$$

To fit the continuation ratio logit model with the correct likelihood function, we can fit a binary logistic regressions after expanding the data based on the form of the probabilities in Equation 5. In particular, the product of I_s and $1 - I_r$ represented in (6) can be seen as probabilities of positive and negative response in binary logistic regression when we define a new binary response variable d for each of $s - 1$ contrasts. The expanded data has a row of data for each term in the product shown in Equation 5: when $y = 2$, expand to 2 rows of data $r = 1, \ldots, s$, making the linear predictor $\alpha_s + \boldsymbol{x_i'\beta}$ and the response variable $d_1 = 0$ and $d_2 = 1$. For $y = S$, delete the last row of data since I_s can be obtained from $I_1, \ldots, I_{s-1}$. Logistic regression then gives probabilities $P(d_r = 0) = 1 - I_r$ for $r = 1, \ldots, s - 1$ and $P(d_r = 1)$ for $r = s$. This expansion methods is shown in the first panel of Figure 2.2 for $S = 4$ ordered categories.

In Figure 2.2, C_{ry} is the likelihood contribution from the binary logistic regression model for row r and response category y. The data can be fully represented by a table such as the one in Figure 2.2 for each distinct set of covariate values, with frequency weights N_s denoting the number of subjects having $y = s$ for that set of covariate values. Concentrating on one set of covariate values, the required likelihood for that subset of the data is

$$\prod_s P(y=s)^{N_s}. \textbf{(6)}$$

Using logistic regression on the expanded data, the total likelihood is

$$\prod_s \left(\prod_{r=1}^{R_s} C_{rs}\right)^{N_s}. \textbf{(7)}$$

This likelihood is correct since $P(y = s) = \prod_{r=1}^{R_s} C_{rs}$ where the product is over the R_s values r that exist for category s in the data expansion as seen from Equation 5. In other words, the likelihood of the expanded data from binary logistic regression is identical to likelihood from the continuation-ratio logit model[59].

Allison (1982) suggested using this data expansion technique in the context of discrete-time survival analysis. Related ideas were proposed in Wu & Ware (1979) in which they followed a similar approach to model data consecutively as it became available. D'agostino, Lee, Belanger, Cupples, Anderson, & Kannel, (1990) pooled the observations collected at different time-points and applied a similar technique. See also Armstrong & Sloan (1989), Ananth & Kleinbaum (1997), Greenland (1994), and Cole & Ananth (2001) for more on this topic.

Recall that when two or more ordinal responses are obtained from each subject, the within-subject dependence needs to be accounted for in the model. To model such

[58] If we were to include covariates $\boldsymbol{x}$, this expression is:
$P(y = s|\boldsymbol{x}) = I_{s|x} \prod_r^{s-1}(1 - I_{r|x})$

[59] Likelihoods has to be identical since strata implied by dichotomizations are conditionally independent in the continuation-ratio logit model.

responses, Tutz (1990) proposed using a sequential one-parameter logistic model—to represent the hypothesized underlying stage-like response process and showed how conditional and unconditional (joint)[60] ML estimates can be obtained via the dichotomization of steps. Tutz (1990) and Mellenberg (1995) discuss the rating scale version of this sequential model. These models can be estimated using the data expansion technique to obtain consistent estimates.

Note that for the continuation-ratio logits model, there are two ways of expressing ordered contrasts: categories could be grouped in an increasing order or in a decreasing order. These two produce different results, thus do not have a palindromic invariance (McCullagh, 1978) property. See Sijtsma & Hemker (2000, p. 284) for further discussion on reversibility. See Appendix B.1 for the alternative formulation of the continuation-ratio logits model.

2.2.2 Cumulative logit model

The cumulative logit model is the most widely used model in ordinal regression modeling due to its ease of interpretation. In addition, in this model, regression parameters are invariant to grouping of categories: meaning, when adjacent categories are merged, this does not affect the remaining parameters. This model is also called proportional odds model due to the assumption of proportionality of cumulative odds across all cut-points. In other words, odds ratios are constrained to be the same for all partitionings of response variables.

The cumulative odds are modeled as

$$\frac{P(y>s)}{P(y\leq s)} = H_s,\ s < S \qquad \textbf{(8)}$$

in which we are comparing the odds of being above versus below any point on the response scale. The corresponding cumulative probabilities are

$$P(y > s) = I_s,\ \ s < S. \qquad \textbf{(9)}$$

It follows that the individual response probabilities are given by

$$P(y = s) = P(y > s - 1) - P(y > s) = I_s - I_{s-1},\ \text{where}\ \ I_0 = 1, I_S = 0. \qquad \textbf{(10)}$$

Cole et al. (2004), Choi & Cole (2004), and Wellman (2006) use the "person-threshold" expansion shown in the second table of Figure 2.2. This expansion, however, produces an incorrect likelihood since

$$P(y = 1) \neq (1 - I_1)(1 - I_2)(1 - I_3). \qquad \textbf{(11)}$$

In other words, the probabilities in Equation 10 do not factorize into products of terms of the form I_r and $1 - I_r$. This expansion, however, produces correct odds,

$$\frac{(N_2+N_3+N_4)}{N_1} = \frac{I_1}{1-I_1}, \frac{(N_3+N_4)}{N_1+N_2} = \frac{I_2}{1-I_2}, \frac{N_4}{N_1+N_2+N_3} = \frac{I_3}{1-I_3}. \qquad \textbf{(12)}$$

[60] Note that unconditional ML method—method still used by some IRT software packages—has the incidental parameter problem (Neyman & Scott, 1948) and is not consistent, resulting in bias in parameter estimates that does not disappear as sample size increases

Thus, using this expansion, we can obtain correct point estimates when the model for these odds is saturated for each set of covariate values.[61] Note that each of the subtables in the second panel of the Figure 2.2 represent each of the odds expressed in Equation 12. Since these odds are assumed to be equal, each of the subtables could also be analyzed independently, but this approach is less efficient.

Notice that in the second panel of Figure 2.2 (cumulative logit model), the number of rows for the expanded data for each category is the same since all S logits are involved in $s-1$ contrasts—this is why these are sometimes called global logits[62] (see for instance Bartolucci et al., 2007). In other words, unlike the other two models, data expansion for the cumulative model uses the full data in obtaining the parameters for each cut-point.

A correct likelihood can be obtained by combining data expansion with a composite link function, as shown in Rabe-Hesketh & Skrondal (2007).

Cole, Allison, & Ananth (2004) presented the data expansion technique for this model with a single outcome. Their main goal was to relax the proportional odds assumptions in order to obtain threshold-specific log-odds ratios. When this assumption is relaxed only partially, or not relaxed at all, the dependence induced by rows of the same subject in the expanded data needs to be accounted for. Cole et al. (2004) handled this dependence using generalized estimating equations with an autoregressive working covariance structure. Note that within-person variability does not cause any issues when the proportional odds assumption is relaxed completely (Cole et al, 2004).

Sufficient statistics do not exist for cumulative logit model to obtain conditional ML estimates. A solution is, then, to translate the ordinal data into binary dichotomizations ("pseudo-responses")—for which sufficient statistics do exist—and simultaneously fit conditional ML on binary data. Agresti & Lang (1993) proposed this transformation into binary dichotomizations to obtain estimates similar to conditional ML estimates. In particular, they used maximum likelihood with additional constraints to fit a quasi-symmetry model[63] on the binary collapsings of the ordinal variables. Binary collapsings in their model were based on a cumulative logit formulation. Their approach to binary data gives identical estimates as the conditional ML estimates obtained from the one-parameter logistic (1-PL) model, first pointed out by Tjur (1982).

Mukherjee, Ahn, Liu, Rathouz, & Sanchez (2008), using similar expansion techniques to Agresti & Lang (1993), demonstrated the use of data expansion to obtain conditional maximum likelihood estimates, which they called amalgamated conditional logistic regression. In Mukherjee et al. (2008), the dependence induced by multiple dichotomizations from each person's response in the expanded data was accounted for using generalized estimating equations.

More recently, in the economics literature, Baetschman (2012)—likely unaware of developments in disjoint literature—proposed similar ideas to Mukherjee et al. (2008) and Agresti & Lang (1993). Baetschman (2012) used cluster standard errors to account

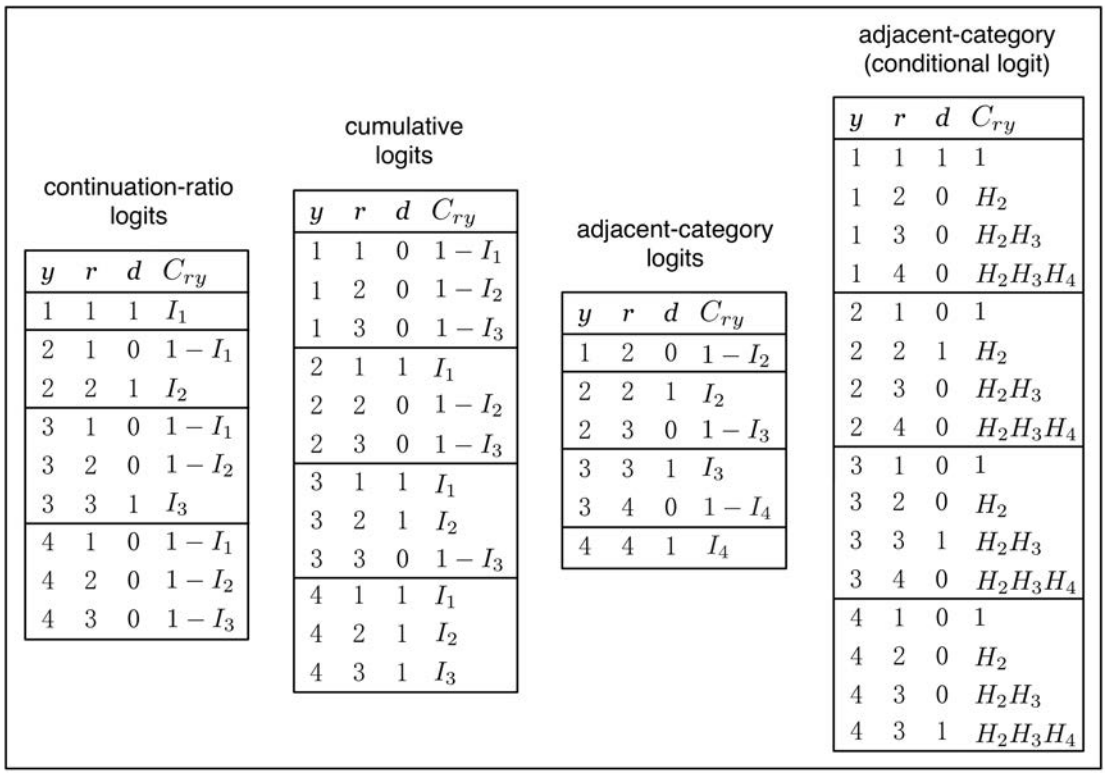
---

[61] IRT models generally do not include covariates, except item dummies

[62] In contrast to global logits, adjacent-category logits are sometimes referred as "local" logits.

[63] For a two-way table, the quasi-symmetry model holds if odds ratios contrasting rows i and i' and columns j and j' $\theta_{ij(i'j')} = \frac{F_{ij}F_{i'j'}}{F_{i'j}F_{ij'}} = \theta_{i'j'(ij)}$ for $i \neq i', j \neq j'$. In particular, a symmetric interaction is imposed by $log(F_{ij}) = \lambda + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{AB(ij)}$ where $\lambda_{AB(ij)} = \lambda_{AB(ji)}$.

for the dependence induced by the same individual in the expanded data. This model is identical to the 1-PL model with amalgamated conditional ML estimates. Muggeo & Aiello (2011) proposed a similar data expansion technique to obtain random-effects (i.e., marginal ML) estimates of the cumulative logit model with alternative link functions.



continuation-ratio logits

y	r	d	C_{ry}
1	1	1	I_1
2	1	0	$1-I_1$
2	2	1	I_2
3	1	0	$1-I_1$
3	2	0	$1-I_2$
3	3	1	I_3
4	1	0	$1-I_1$
4	2	0	$1-I_2$
4	3	0	$1-I_3$

cumulative logits

y	r	d	C_{ry}
1	1	0	$1-I_1$
1	2	0	$1-I_2$
1	3	0	$1-I_3$
2	1	1	I_1
2	2	0	$1-I_2$
2	3	0	$1-I_3$
3	1	1	I_1
3	2	1	I_2
3	3	0	$1-I_3$
4	1	1	I_1
4	2	1	I_2
4	3	1	I_3

adjacent-category logits

y	r	d	C_{ry}
1	2	0	$1-I_2$
2	2	1	I_2
2	3	0	$1-I_3$
3	3	1	I_3
3	4	0	$1-I_4$
4	4	1	I_4

adjacent-category (conditional logit)

y	r	d	C_{ry}
1	1	1	1
1	2	0	H_2
1	3	0	H_2H_3
1	4	0	$H_2H_3H_4$
2	1	0	1
2	2	1	H_2
2	3	0	H_2H_3
2	4	0	$H_2H_3H_4$
3	1	0	1
3	2	0	H_2
3	3	1	H_2H_3
3	4	0	$H_2H_3H_4$
4	1	0	1
4	2	0	H_2
4	3	0	H_2H_3
4	3	1	$H_2H_3H_4$

Figure 2.2. Data expansion rules for the variable with four ordered categories.
Note: y indicates the original response category and d is the constructed binary response variable for the expanded data, and r indicates the subtable in the expanded data.

In Figure 2.2, r indicates the sub-data (sub-table)—subset of the data on which binary logistic regression is performed. Note that for the continuation-ratio and adjacent category logit models, d in the table represents the numerator in each of the models presented Equations 2, 8, and 13 for continuation-ratio, cumulative, and adjacent category logit models respectively. In other words, it indicates whether $r = y$ for continuation-ratio and adjacent-category, while in cumulative logits model, d indicates whether $r > y$. Also note that C_{ry} indicates the likelihood contribution.

The cumulative logit model for polytomous item is unique among the three models in that it can expressed using the latent response formulation, as shown in Appendix B.4.

2.2.3 Adjacent-category logit model

In adjacent-category logit models, logits are formed locally. In particular, adjacent-category odds are modeled as

$$\frac{P(y=s)}{P(y=s-1)} = H_s, \;\; s > 1 \tag{13}$$

and the corresponding conditional probabilities are

$$P(y = s|y \in \{s-1, s\}) = I_s, \;\; s > 1. \tag{14}$$

The unconditional response probabilities can be derived by

$$\frac{P(y=s)}{P(y=1)} = \frac{P(y=s)}{P(y=s-1)}\frac{P(y=s-1)}{P(y=s-2)} \ldots = \prod_{r=1}^{s} H_r \,, \;\; s > 1, E_1 = 1. \tag{15}$$

It follows that

$$P(y = s) = P(y = 1)\prod_{r=1}^{s} H_r \,, \;\; s > 1. \tag{16}$$

Since the probabilities must sum to 1, we obtain

$$P(y = s) = \frac{\prod_{r-1}^{s} H_r}{\sum_{r=1}^{S}\prod_{r-1}^{s} H_t}. \tag{17}$$

Choi & Cole (2004) use the data expansion shown in the third table in Figure 2.2. However, this produces an incorrect likelihood since $P(y = 1) \neq 1 - I_2, P(y = 2) \neq I_2(1 - I_3)$. The probabilities in (17) do not factorize into products of terms of the form I_r and $(1 - I_r)$ because the denominator, $1 + H_2 + H_2H_3 + H_2H_3H_4 + \cdots$ does not factorize into products of terms having the form $1 + H_r$. For example, for $S = 3$, the product $(1 + H_2)(1 + H_3) = 1 + H_2 + H_3 + H_2H_3$ gives an extra term H_3.

However, this data expansion produces correct adjacent-category odds asymptotically. In particular, since

$$\frac{N_2}{N_1} = \frac{I_2}{1-I_2}, \frac{N_3}{N_2} = \frac{I_3}{1-I_3}, \frac{N_4}{N_3} = \frac{I_4}{1-I_4}, \tag{18}$$

the estimates are correct when the model for these odds is saturated for each set of covariate values (i.e., with all possible response patterns). Equation 18 essentially implies that when the response is in the, say, second category, only those who responded in the second or first category are involved in estimating the step parameter.

The data expansion for conditional logistic regression to obtain conditional ML estimates for adjacent categories is given in the last table in Figure 2.2 where "term" is the numerator of the unconditional probabilities, each value of y is a different group, and we condition on the sum of responses being 1 for each group.

Figure 2.3 shows the first three tables shown in Figure 2.2 with reordered rows so that contrasts that are shown reflect contrasts presented in Figure 2.1 for the respective ordinal model.

continuation-ratio logits

y	r	d	C_{ry}
1	1	1	I_1
2	1	0	$1-I_1$
3	1	0	$1-I_1$
4	1	0	$1-I_1$
2	2	1	I_2
3	2	0	$1-I_2$
4	2	0	$1-I_2$
3	3	1	I_3
4	3	0	$1-I_3$

cumulative logits

y	r	d	C_{ry}
1	1	0	$1-I_1$
2	1	1	I_1
3	1	1	I_1
4	1	1	I_1
1	2	0	$1-I_2$
2	2	0	$1-I_2$
3	2	1	I_2
4	2	1	I_2
1	3	0	$1-I_3$
2	3	0	$1-I_3$
3	3	0	$1-I_3$
4	3	1	I_3

adjacent-category logits

y	r	d	C_{ry}
1	2	0	$1-I_2$
2	2	1	I_2
2	3	0	$1-I_3$
3	3	1	I_3
3	4	0	$1-I_4$
4	4	1	I_4

Figure 2.3. Data expansion rules for the variable with four ordered categories reflecting contrasts presented in Figure 2.1 sorted by r (i.e., subtable).

The adjacent-category logit model can be viewed as a constrained multinomial logit model which in turn can be estimated by expanding the data using conditional logistic regression (Agresti, 1993); Poisson regression (Chen and Kuo, 2001); or stratified Cox regression (Allison & Christakis, 1994; Chen & Kuo, 2001).

In the psychometrics literature, Masters (1982) proposed an IRT analogue of this model—a partial credit model (PCM)—with joint and conditional ML estimation; Wilson & Adams (1993) proposed marginal ML estimation. The PCM is unique among the polytomous IRT models due to its sufficiency property. In particular, the counts of subjects completing each step of an item are jointly sufficient for the item steps (Wright & Masters, 1982). The rating scale model (RSM; Andrich, 1978) is identical to PCM in the underlying response structure: it assumes however, that the residual thresholds across items are constrained to be same. This model is useful when items have the same ordered categories as Likert scales. The difference between the PCM and RSM models is illustrated in Appendix B.2.

Agresti (1993) used the rating scale model and obtained conditional maximum likelihood estimates by fitting a log-linear model. Note that PCM can be considered a special case of the nominal response model (Mellenbergh, 1995). Another related model is Wilson's (1992) ordered partition model (OPM). The OPM allows nominal levels within any given ordered level.[64]

[64] Wilson (1992; Wilson & Adams, 1993) developed the ordered partition model (OPM) as an extension of the PCM. The OPM is designed to model data that is neither entirely nominal nor completely ordered. In particular, it is viewed as appropriate model for items that have ordered levels, but where there may be a number of nominal response categories within any given level. Wilson suggests that the OPM may be particularly relevant for data resulting from the performance assessments.

2.2.4 Data expansion for the PCM model when marginal maximum likelihood method is used

Recall that each subtable provides information about the random intercept and different subsets of subjects are represented in different subtables. If the latent variable of the subject is high, subject's response will be in the higher category. In other words, for the lowest two categories in the adjacent-category model, the mean of the latent variable will be lower than in the highest two categories. Therefore, when there is a random-effect (latent variable), each subset of the data will have a different latent variable mean and variance that is different from the mean and variance of the entire population—the variance of the latent variable, will be smaller in each subset of the data than the variance in the population (entire data). Data expansion will not produce consistent estimates for the PCM model when marginal maximum likelihood method is used (see Table 2 below for the direction of the bias).

Also note that this issue does not apply to the cumulative logit model since the data for each contrast is not a subset of the data (as can be seen from Figure 2.1) –it is just the data itself but with additional variables indicating the contrasts.

2.3 Population study

To evaluate the consistency of the estimators and to demonstrate the performance of the data expansion technique asymptotically, we generated a sample with all possible response patterns (i.e., saturated data) using probabilities as frequency weights and used probabilities of these response patterns as weights for the log-likelihood contribution[65]. Similar approaches were taken in Breinegaard, Rabe-Hesketh, & Skrondal (2015) and Jeon & Rabe-Hesketh (2015)[66].

Before presenting the conditions of the population parameters, I briefly present models from which the true parameters were generated (i.e., "gold standard"): cumulative 1-PL and partial credit models.

In the partial credit model, the probability of person j scoring k on item i, P_{jik}, can be expressed as

$$P_{jik} = \frac{exp\sum_{l=0}^{k}(\theta_j - \beta_{ik})}{\sum_{h=0}^{M_i} exp\sum_{l=0}^{h}(\theta_j - \beta_{ik})}, \quad k = 0,1, \dots, M_i, \tag{19}$$

where θ_j, and β_{ik} are the ability of person j and the difficulty of step k of item i respectively; $M_i + 1$ is the number of (ordered) categories for the item, and we use the following notational conventions for identification:

$$\sum_{k=0}^{0}(\theta_j - \beta_{ik}) \equiv 0, \tag{20}$$

and

$$\sum_{k=0}^{h}(\theta_j - \beta_{ik}) \equiv \sum_{k=1}^{h}(\theta_j - \beta_{ik}). \tag{21}$$

[65] Analyses of the population data was carried out using gllamm (Rabe-Hesketh, Skrondal, & Pickles, 2005) package in Stata (see Appendices B.3a and B.3b).

[66] For earlier applications of this approach see Rotnitzky & Wypij (1994) and Heagerty & Kurland (2001).

In the cumulative 1-PL model, which is the special case of the graded response model with all discrimination parameters fixed at one, the probability of person j scoring k on item i, P_{jik}, can be expressed as[67]:

$$P_{jik} = \frac{\exp(\theta_j - \gamma_{ik})}{1+\exp(\theta_j - \gamma_{ik})} - \frac{\exp(\theta_j - \gamma_{ik+1})}{1+\exp(\theta_j - \gamma_{ik+1})}, \quad k = 0,1,\dots,M_i. \tag{22}$$

where θ_j, and γ_{ik} are the ability of person j and the difficulty of step k of item i, respectively.

For binary items, both of the models above simplify to the Rasch model:

$$P_{ji1} = \frac{\exp(\theta_j - \delta_i)}{1+\exp(\theta_j - \delta_i)}, \tag{23}$$

or

$$\text{logit}\left(P_{ji1}\right) = \tau_{ji1} = \theta_j - \delta_i, \tag{24}$$

in which θ_j and δ_i are the ability of person j and the difficulty of item i, respectively.

To investigate the data expansion for the cumulative 1-PL model, we generated data from this model for three hypothetical items with three categories in each, with step parameters $\gamma_{11} = -2.0$, $\gamma_{12} = -1.0$; $\gamma_{21} = 0.0$, $\gamma_{22} = 0.5$; $\gamma_{31} = -1.0$, $\gamma_{32} = 1.0$, and a variance for the latent variable, $\psi = 1.0$. Note that in this model, step parameters within a given item should be ordered with $\gamma_{i1} < \gamma_{i2} < \cdots < \gamma_{iM_i}$.[68]

Using the generated "population" data (with the cumulative 1-PL model as the true model), we applied data expansion technique by using probabilities of each response pattern as weights.

Results from the population study with the cumulative 1-PL model are shown in Table 2.1 below. As can be seen from the table, by using the data expansion technique on the population data, we were able to recover parameter estimates of the cumulative 1-PL model using both marginal and conditional ML methods. The third column provides "population" standard errors—these are standard errors for the sample size of 1. To obtain the standard error had the sample size been, say, 1000, we would have to divide the respective standard error with $\sqrt{1000}$. These standard errors for N=1000 are provided with the obtained point estimates for the marginal ML method (for both Tables 2.1 and 2.2).

Note that data expansion with conditional ML method fixes the step parameter for the first item at zero. Thus, parameters from the "True" model were adjusted to have first two parameters constrained at zero.

[67] See for instance Samejima (1997) for the 2-PL version of the model. Although I call the cumulative 1-PL model as the cumulative Rasch model in the first and third chapters of this dissertation, some might disagree that such a model "qualifies" as the Rasch model.

[68] Ordering of the threshold parameters in cumulative 1-PL and partial credit models has been the topic of recent discussions (Adams, Wu, & Wilson, 2012; Andrich, 2013).

Table 2.1. Population parameter estimates for the cumulative 1-PL model.

Parameters	True	marginal ML expanded	population SE	True (adj.)	conditional ML expanded
γ_{11}	-2.0	-2.00 (0.11)	3.31	0.0	0.00
γ_{12}	-1.0	-1.00 (0.08)	2.65	0.0	0.00
γ_{21}	0.0	0.00 (0.08)	2.42	-2.0	-2.00 (0.78)
γ_{22}	0.5	0.50 (0.08)	2.48	-1.5	-1.50 (0.70)
γ_{31}	-1.0	-1.00 (0.08)	2.65	-1.0	-1.00 (0.77)
γ_{32}	1.0	1.00 (0.08)	2.65	-2.0	-2.00 (0.77)
ψ	1.0	1.00 (0.13)	4.12	1.0	

Note: Robust standard errors for the marginal ML were obtained by converting probability weights from the population data into frequency weights by multiplying each probability (i.e., "frequency" of each response pattern) by 1000 (i.e., N=1000). Conditional ML estimates using data expansion with cluster robust standard errors were obtained using clogit command in Stata.

To investigate the data expansion for the partial credit model, we generated data from the PCM model for three hypothetical items with three categories in each, with step parameters: $\beta_{11} = 0.5$, $\beta_{12} = 0.2$; $\beta_{21} = 0.0$, $\beta_{22} = 1.0$; $\beta_{31} = 0.6$, $\beta_{32} = 0.2$, and a variance for the latent variable, $\psi = 1.0$.

Results from the population study with the partial credit model are shown in Table 2.2 below. Note that estimates obtained from the data expansion using marginal ML method are inconsistent for the reasons detailed in Section 2.2.4. We can observe from Table 2.2 that parameters for the first step within each of the three items ($\beta_{11}, \beta_{21}, \beta_{31}$) are underestimated while parameters for the second step within each item ($\beta_{12}, \beta_{22}, \beta_{32}$) are overestimated. This is because the mean of the latent variable of those who "pass" the second threshold in each item is higher than the mean of those who do not pass the second threshold. As a result, first step parameters appear "more difficult" than they actually are (i.e., estimated "easiness" is lower than the true "easiness"). Similarly, second step parameters appear "less difficult" than the true estimates (estimated "easiness" is higher than the true "easiness").

Table 2.2. Population parameter estimates for the partial credit model.

Parameters	True	Marginal ML expanded	population SE	True (adj.)	conditional ML expanded
β_{11}	0.5	0.09 (0.10)	3.09	0.0	0.00
β_{12}	0.2	0.46 (0.09)	2.76	0.0	0.00
β_{21}	0.0	-0.52 (0.11)	3.49	-0.5	-0.50 (0.80)
β_{22}	1.0	1.28 (0.10)	3.19	0.8	0.80 (0.86)
β_{31}	0.6	0.19 (0.10)	3.12	0.1	0.10 (0.82)
β_{32}	0.2	0.44 (0.09)	2.73	0.0	0.00 (0.95)
ψ	1.0	0.73 (0.15)	4.82	1.0	

Note: Robust standard errors for the marginal ML were obtained by converting probability weights from the population data into frequency weights by multiplying each probability (i.e., "frequency" of each response pattern) by 1000. CML estimates using data expansion with cluster robust standard errors were obtained using clogit command in Stata.

For both cumulative 1-PL and partial credit models, data expansion technique provides consistent estimates. Stata code for the simulation of the population data is provided in Appendices B.3a and B.3b for both cumulative and adjacent-category logit models.

We also confirmed that the data expansion for both cumulative and adjacent-category logit models perfectly recovers the parameters when there is no latent variable or random effect (See Appendices B.3c and B.3d for the population study code).

2.4 Example: HADS depression dataset

The demo dataset contains responses of 201 oncological patients to 14 ordinal polytomous items that measure anxiety (7 items) and depression (7 items), according to the Hospital Anxiety and Depression Scale questionnaire developed by Zigmond & Snaith (1983). The data is publicly available (see Appendix B.3e for details) and was collected by Newell, Sanson-Fisher, Girgis, & Ackland, 1999, within a larger cross-sectional study, which in addition to assessing anxiety and depression aimed to assess the prevalence and predictors of physical symptoms and perceived needs among oncological patients of the academic outpatient medical oncology department (see Newell et al, 1999). The research was conducted in general medical outpatient clinics on adults of both genders between the ages of 16 and 65 who suffered from a wide variety of complaints and illnesses.

For this demo example, only items measuring depression from the HADS instrument were selected (see Appendix B.5). All items have 4 response categories: the minimum value 0 corresponds to a low level of depression, whereas the maximum value 3 corresponds to a high level of depression.

Estimates of item step parameters for the cumulative1-PL model (i.e., 1-PL version of the graded response model) are shown in the Table 2.3 below.

The first column in Table 2.3 represents estimates obtained using data expansion method with cluster-robust standard errors. The second column in Table 2.3 represents estimates for the cumulative Rasch model using cumulative logit link (i.e., exact method, benchmark).

The third column in Table 2.3 represents estimates from the exact method with rescaled step parameters. In particular, to be able to compare to the conditional ML method, we adjusted estimates obtained from the cumulative 1-PL by setting three step parameters of the first item to zero (by subtracting γ_{11}, γ_{12}, and γ_{13} from γ_{i1}, γ_{i2}, and γ_{i3} respectively).

The fourth column in Table 2.3 represents estimates obtained using data expansion method with amalgamated conditional ML estimates and cluster-robust standard errors.

As we see from the Table 2.3 above, data expansion technique worked well to obtain estimates for the cumulative 1-PL model.

Table 2.3. Estimates for the cumulative 1-PL (1-PL graded response) model using depression items in the HADS instrument

	Marginal MLE			Conditional MLE
parameters	Data expansion	exact (1-PL GRM)	adjusted exact	Data expansion
γ_{11}	-0.88 (0.29)	-0.89 (0.21)	0.00	0.00
γ_{12}	2.83 (0.33)	2.82 (0.28)	0.00	0.00
γ_{13}	4.32 (0.48)	4.31 (0.42)	0.00	0.00
γ_{21}	-0.23 (0.28)	-0.27 (0.20)	0.62	0.62 (0.24)
γ_{22}	1.15 (0.29)	1.04 (0.21)	-1.79	-1.72 (0.31)
γ_{23}	3.19 (0.35)	3.10 (0.30)	-1.21	-1.18 (0.42)
γ_{31}	-2.01 (0.32)	-2.05 (0.24)	-1.16	-1.07 (0.33)
γ_{32}	1.15 (0.28)	1.09 (0.21)	-1.73	-1.72 (0.30)
γ_{33}	3.71 (0.41)	3.69 (0.35)	-0.62	-0.63 (0.45)
γ_{41}	-3.48 (0.43)	-3.54 (0.33)	-2.65	-2.51 (0.47)
γ_{42}	-0.14 (0.29)	-0.16 (0.20)	-2.98	-3.02 (0.35)
γ_{43}	3.43 (0.39)	3.40 (0.32)	-0.91	-0.92 (0.58)
γ_{51}	0.37 (0.27)	0.43 (0.21)	1.32	1.22 (0.29)
γ_{52}	2.65 (0.34)	2.61 (0.27)	-0.21	-0.19 (0.39)
γ_{53}	4.18 (0.44)	4.06 (0.39)	-0.25	-0.14 (0.60)
γ_{61}	-0.46 (0.29)	-0.48 (0.20)	0.41	0.40 (0.29)
γ_{62}	2.43 (0.31)	2.40 (0.26)	-0.43	-0.42 (0.35)
γ_{63}	4.32 (0.45)	4.29 (0.42)	-0.02	0.00 (0.61)
γ_{71}	-1.16 (0.31)	-1.19 (0.21)	-0.29	-0.27 (0.36)
γ_{72}	1.11 (0.28)	1.04 (0.21)	-1.78	-1.75 (0.33)
γ_{73}	4.67 (0.51)	4.60 (0.47)	0.29	0.34 (0.63)

Note: Estimates for the exact method was obtained using clogit command in Stata. Code the exact and data expansion methods is provided in Appendix B.3e. Standard errors are larger than expansion method. Robust standard errors are provided in column 3 and they never differ from model-based standard errors more than 0.02.

Estimates of item step parameters for the partial credit model are shown in the Table 2.4. The first column in Table 2.4 shows conditional ML estimates obtained using data expansion method for partial credit model. The second column shows conditional ML estimates for the partial credit model using specialized software for polytomous items (eRm package in R;). Third column in Table 2.4 shows adjusted values of the exact methods similar to that of Table 2.3 (adjusting for the step parameters of the first item constrained to zero).

Table 2.4. Conditonal ML estimates of the item location parameters using the partial credit model applied on depression items in the HADS instrument.

parameters	Expansion	Exact PCM	Adjusted exact PCM
β_{11}	0.00	0.00	0
β_{12}	0.00	3.18 (0.36)	0
β_{13}	0.00	2.80 (0.51)	0
β_{21}	1.20 (0.32)	1.09 (0.26)	1.09
β_{22}	-2.58 (0.42)	1.09 (0.29)	-2.09
β_{23}	-0.72 (0.56)	2.61 (0.36)	-0.19
β_{31}	-0.98 (0.38)	-0.86 (0.27)	-0.86
β_{32}	-1.92 (0.34)	1.62 (0.26)	-1.56
β_{33}	-0.22 (0.55)	3.10 (0.39)	0.30
β_{41}	-2.25 (0.51)	-2.08 (0.37)	-2.08
β_{42}	-3.09 (0.39)	0.56 (0.24)	-2.62
β_{43}	-0.28 (0.69)	3.23 (0.35)	0.43
β_{51}	1.44 (0.31)	1.24 (0.24)	1.24
β_{52}	-0.79 (0.48)	2.61 (0.34)	-0.57
β_{53}	-1.05 (0.98)	2.84 (0.49)	0.04
β_{61}	0.42 (0.31)	0.45 (0.24)	0.45
β_{62}	-0.64 (0.43)	2.59 (0.31)	-0.58
β_{63}	0.09 (0.82)	3.13 (0.49)	0.33
β_{71}	-0.04 (0.40)	0.04 (0.25)	0.04
β_{72}	-2.11 (0.38)	1.30 (0.26)	-1.88
β_{73}	1.08 (0.78)	4.12 (0.49)	1.32

Note: Estimates for the exact method was obtained using clogit command in Stata. Code for the exact and data expansion methods is provided in Appendix B.3e.

Notice that due to the small sample size (N=201), data expansion for the partial credit model didn't work as well as in the cumulative 1-PL model. As was shown in the population study, estimates from the data expansion method are consistent and will hence be closer to the estimates from the exact method as the sample size increases. Next we analyze the data using the rating scale model (RSM), which has less number of parameters.

Recall that the rating scale model is a special case of PCM, which assumes that the threshold parameters across the items are constrained to be the same (see Appendix B.2 for the graphical illustration). By slightly modifying the expression for the PCM model (Equation 19), the RSM can be expressed

$$P_{jik} = \frac{exp\sum_{l=0}^{k}(\theta_j-(\beta_i+\tau_k))}{\sum_{h=0}^{M_i} exp\sum_{l=0}^{h}(\theta_j-(\beta_i+\tau_k))}, \quad k = 0,1,\dots,M_i. \qquad \textbf{(24)}$$

Similar to the PCM, data expansion with the conditional ML method can be applied to the data to approximate the estimates from the RSM model. However, step

parameters across all items (which are equal across items in the RSM model) are not estimated when using data expansion technique.

Item location parameters from the rating scale model for the items in the demo dataset (using conditional ML method for both exact and data expansion methods) are shown in the Table 25 below. Notice that, unlike estimates in the previous table (Table 2/4, PCM), estimates from the data expansion method in Table 2.5 are closer to the estimates from the exact method. This is because the RSM has fewer parameters and thus the data expansion method is more efficient than for the unconstrained PCM.

Table 2.5. Conditonal ML estimates of the item location parameters using the rating scale model applied on depression items in the HADS instrument.

parameters	Exact method	Data expansion
β_2	-0.32 (0.15)	-0.25 (0.20)
β_3	-0.83 (0.15)	-0.96 (0.18)
β_4	-1.53 (0.16)	-1.82 (0.22)
β_5	0.50 (0.16)	0.61 (0.22)
β_6	0.08 (0.16)	0.09 (0.19)
β_7	-0.53 (0.15)	-0.57 (0.22)

Note: Location of the first item is constrained to zero for the identification of the model. Conditional ML estimates for the exact method was obtained using eRm package in R. Data expansion estimates were obtained using Stata (see Appendix B.3e for the code).

2.5 Discussion

In choosing among the three coding schemes mentioned above, researchers might prefer the one that is easiest to interpret, or more simply, the one that is most often used in their field. For instance, adjacent-category logits model is mostly used in educational measurement, while cumulative logits model is commonly used in biostatistics and economics, and continuation-ratio logits model is often used in biostatistics and sociology. However, the choice of which model to use ought be based on the specific research question, research design, and a careful consideration of the underlying response processes.

Indeed, if the focus is on investigating the odds of advancing beyond a particular level conditional on reaching that level, a continuation ratio should be preferred. If, on the other hand, the focus in on investigating odds of being at or beyond a particular level, then a cumulative logits model would be preferred. And similarly with a focus on odds of being at a higher category of two adjacent categories, then the adjacent-category logits model would be a suitable option. When the ordering of the category parameters is the matter of investigation, however, then adjacent-category logit model will likely be most appropriate since it does not have order constraints[69], as is the case with the cumulative logit model. Statistical tests can be conducted by comparing if more parsimonious model (adjacent-category) fits as good as baseline category logit model (Tuerlinckx & Wang,

[69] Though not requirement of the rating scale or partial credit model, ordering of the threshold parameters has been the topic of recent discussions (Adams, Wu, & Wilson, 2012; Andrich, 2013).

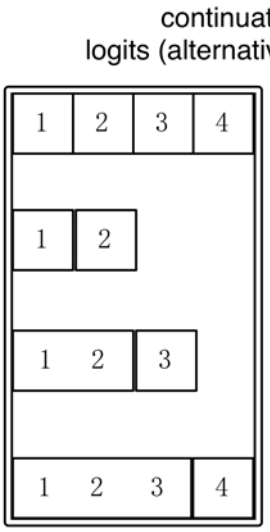
2004). Another direction of research might be in developing methods for checking assumptions and goodness of fit of the polytomous IRT models.

The data expansion techniques discussed in this chapter are useful tools, in particular when used on the datasets with larg sample size. These techniques can be easily employed when the software that allows the exact methods is not available. One can also use these techniques to obtain a version of the conditional ML estimates with the cumulative logit model. Using the population data, we showed that results obtained from the data expansion are asymptotically correct. We demonstrated how to apply data expansion for both marginal and conditional ML methods. We also explained why one should not use data expansion for the partial credit model with marginal ML method. Results obtained from the data expansion technique are more reliable when the sample size is large.

Appendix B.1: Continuation-ratio model with decreasing order.

Recall that continuation-ratio logit model can be in either increasing or decreasing order of categories. The former was presented in Figures 2.1 and 2.2. Below I present the latter, in particular:

$$\frac{P(y=s)}{P(y<s)} = H_s, s < S,$$



y	r	d	C_{ry}
2	1	0	$1 - I_1$
2	2	1	I_1
3	1	0	$1 - I_2$
3	2	0	$1 - I_2$
3	3	1	I_2
4	1	0	$1 - I_3$
4	2	0	$1 - I_3$
4	3	0	$1 - I_3$
4	4	1	I_3

Figure 2.4. Continuation-ratio logit model with decreasing order.

Appendix B.2: Hypothetical estimates from the partial credit and rating scale models.

Assume four items with four categories ("strongly disagree", "disagree", "agree", "strongly agree").

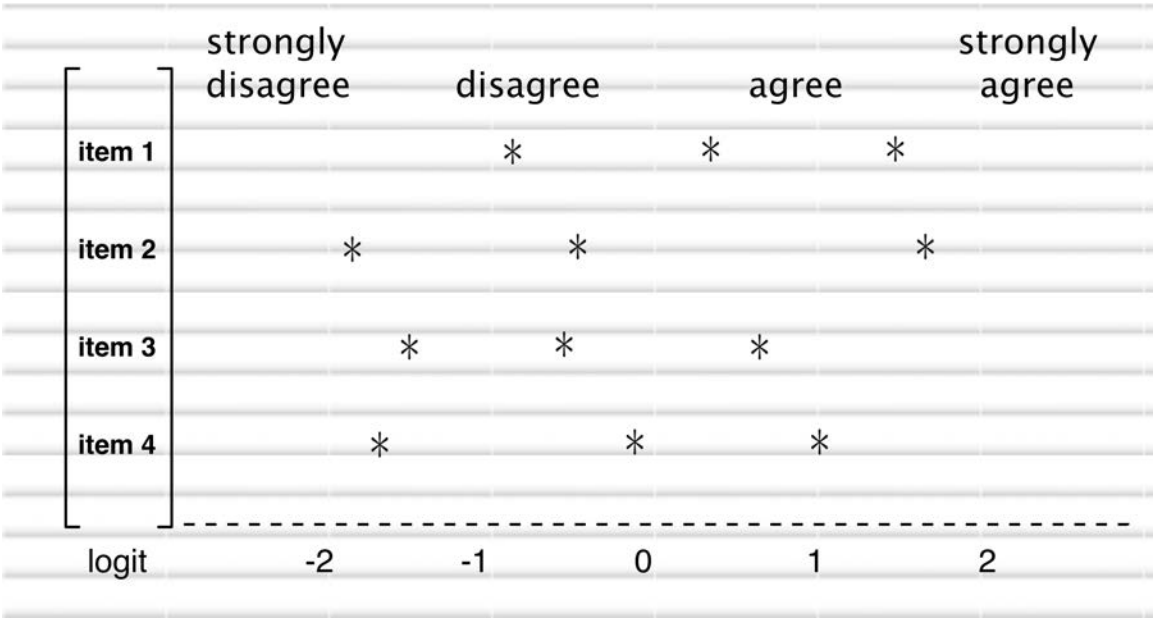


Figure 2.5. Hypothetical estimates of the three step parameters for each of the four items in the partial credit model. Note that distances between points are allowed to vary within each item.

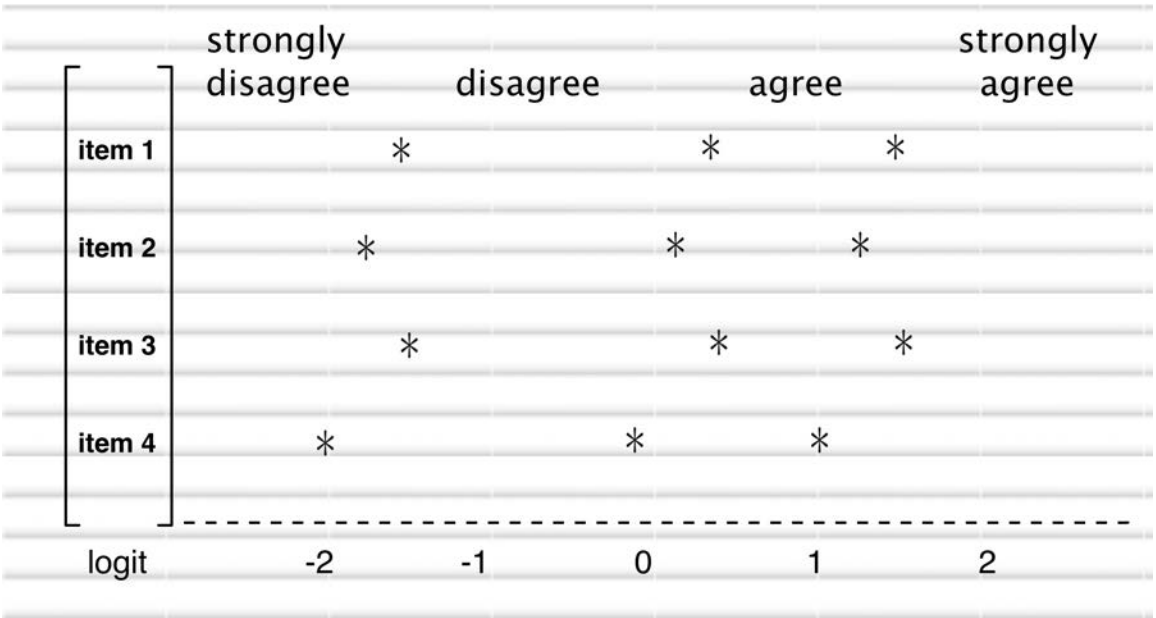


Figure 2.6. Hypothetical estimates of the three step parameters across four items. Note that distances between the points within each item is fixed across items.

Appendix B.3a: Population study: STATA code for data expansion for cumulative logit model

```
**** Graded-response model
* 3 items, 3 categories
global nitem = 3
global ncat = 3
* generate all possible response patterns
clear
set obs 1
generate pattern=1
forvalues i=1/$nitem{
      expand $ncat
      by pattern, sort: generate y`i'=_n
      replace pattern=_n
}
reshape long y, i(pattern) j(item)
* make up model parameters & calculate log-likelihood contributions for
* each pattern
tabulate item, generate(i_)
matrix a=(-2, -1, 0, .5, -1, 1, 1)
gllamm y, i(pattern) link(ologit ologit ologit) lv(item) adapt nip(30)
/// from(a) copy eval
gllapred loglik, ll
* calculate probabilities of response patterns
generate double prob = exp(loglik)
********* analyze model by MLE for population data
generate double wt2 = prob
gllamm y, i(pattern) link(ologit ologit ologit) lv(item) adapt nip(30)
/// weight(wt)
****** analyze by data expansion for population data
save gradedp, replace
* >1
use gradedp, clear
generate set=1
generate d = y>1
gllamm d i_*, nocons i(pattern) link(logit) fam(binom) adapt nip(30)
/// weight(wt)
save graded_g1, replace
* amalgamated CML
clogit d i_2 i_3 [pweight=prob], group(pattern)

* >2
use gradedp, clear
generate set=2
generate d = y>2
gllamm d i_*, nocons i(pattern) link(logit) fam(binom) adapt nip(30)
/// weight(wt)
* amalgamated CML
clogit d i_2 i_3 [pweight=prob], group(pattern)

**** pool datasets
append using graded_g1
forvalues i=1/3 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      drop i_`i'
```

```
}
egen pattset = group(pattern set)
gllamm d i_*, nocons i(pattset) link(logit) fam(binom) adapt nip(30)
/// weight(wt)
* amalgamated CML
clogit d i_2* i_3* [pweight=prob], group(pattset) vce(cluster pattern)
* to obtain robust standard errors
generate freq = round(1000*prob, 1)
drop wt2
generate double wt2 = freq
gllamm d i_*, nocons i(pattset) link(logit) fam(binom) adapt nip(30)
weight(wt) robust
```

Appendix B.3b: Population study: STATA code for data expansion for adjacent-category logit model

```
**** partial credit model
* 3 items, 3 categories
global nitem = 3
global ncat = 3
* generate all possible response patterns
clear
set obs 1
generate pattern=1
forvalues i=1/$nitem{
      expand $ncat
      by pattern, sort: generate y`i'=_n
      replace pattern=_n
}
reshape long y, i(pattern) j(item)
* expand data for estimation using gllamm
generate pers_it = _n
expand $ncat
by pers_it, sort: generate r=_n
forvalues i=1/$nitem{
      forvalues c=2/$ncat {
            generate i_`i'_`c'= (r>=`c')*(item==`i')
      }
}
generate d = (y==r)
* make up model parameters & calculate log-likelihood contributions for
* each pattern
matrix a=(0.5, 0.2, 0, 1.0, 0.6, 0.2, 1)
eq abil: r
gllamm r i_* , nocons i(pattern) link(mlogit) expanded(pers_it d o) ///
eqs(abil) adapt nip(30) from(a) copy eval trace
gllapred loglik, ll
* calculate probabilities of response patterns
generate double prob= exp(loglik)
********* analyze model by MLE for population data
generate double wt2 = prob
gllamm r i_* , nocons i(pattern) link(mlogit) expanded(pers_it d o) ///
eqs(abil) adapt nip(30) weight(wt)
* without random effects
gllamm r i_* , nocons i(pattern) link(mlogit) expanded(pers_it d o) ///
eqs(abil) init nip(30) weight(wt)
clogit d i_* [pweight=prob], group(pers_it)
* check effect of rounding
generate freq=round(10000*prob,1)
clogit d i_* [fweight=freq], group(pers_it)
drop freq
***** make unexpanded data
drop i_*
/* keep just the "data" */
keep if d==1
drop d r
save pcreditp, replace
****** analyze by data expansion for population data
* 2 versus 1
use pcreditp, clear
```

```
keep if y==1|y==2
generate set=1
generate d = y==2
tabulate item, generate(i_)
gllamm d i_*, nocons i(pattern) link(logit) fam(binom) adapt nip(30)
/// trace weight(wt)
save precredit12, replace
* without random effect
logit d i_* [pweight=prob], nocons
* fixed-effects
clogit d i_2 i_3 [pweight=prob], group(pattern)
* 3 versus 2
use pcreditp, clear
keep if y==2|y==3
generate set=2
generate d = y==3
tabulate item, generate(i_)
gllamm d i_*, nocons i(pattern) link(logit) fam(binom) adapt nip(30)
/// trace weight(wt)
save precredit23, replace
* without random effect
logit d i_* [pweight=prob], nocons
* fixed-effects
clogit d i_2 i_3 [pweight=prob], group(pattern)
***** pool datasets
use precredit12, clear
append using precredit23
forvalues i=1/3 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      drop i_`i'
}
egen pattset = group(pattern set)
gllamm d i_*, nocons i(pattset) link(logit) fam(binom) adapt nip(30)
/// trace weight(wt)
* without random effect
logit d i_* [pweight=prob], nocons
* fixed-effects
clogit d i_2* i_3* [pweight=prob], group(pattset) vce(cluster pattern)
generate freq = round(10000*prob, 1)
clogit d i_1* i_3* [pweight=prob], group(pattset) vce(cluster pattern)
disp _se[i_1_2]*sqrt(18/10000)
clogit d i_1* i_3* [fweight=freq], group(pattset) vce(cluster pattern)
* expand data by frequency weights
* 2 versus 1
use pcreditp, clear
drop pers_it
reshape wide y, i(pattern) j(item)
generate freq = round(10000*prob, 1)
expand freq
generate id=_n
reshape long y, i(id) j(item)
save pcredit_exp, replace
keep if y==1|y==2
generate set=1
generate d = y==2
tabulate item, generate(i_)
save junk, replace
```

```
* 3 versus 2
use pcredit_exp, clear
keep if y==2|y==3
generate set=2
generate d = y==3
tabulate item, generate(i_)
append using junk
forvalues i=1/3 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      drop i_`i'
}
egen idset = group(id set)
clogit d i_1* i_3* , group(idset) vce(cluster id)
drop wt2
drop freq
generate freq = round(1000*prob, 1)
generate double wt2 = freq
gllamm d i_*, nocons i(pattset) link(logit) fam(binom) adapt nip(30)
trace weight(wt) robust
```

Appendix B.3c: Population study: STATA code for data expansion for cumulative logit model without random effect

```
**** cumulative logit model
* 3 items, 3 categories
global nitem = 3
global ncat = 3
* generate all possible response patterns
clear
set obs 1
generate pattern=1
forvalues i=1/$nitem{
      expand $ncat
      by pattern, sort: generate y`i'=_n
      replace pattern=_n
}
reshape long y, i(pattern) j(item)
* make up model parameters & calculate log-likelihood contributions for
each pattern
tabulate item, generate(i_)
matrix a=(-2, -1, 0, .5, -1, 1, 0)
gllamm y, i(pattern) link(ologit ologit ologit) lv(item) adapt nip(30)
from(a) copy eval
gllapred loglik, ll
* calculate probabilities of response patterns
generate double prob = exp(loglik)
********* analyze model by MLE for population data
generate double wt2 = prob
gllamm y, i(pattern) link(ologit ologit ologit) lv(item) adapt nip(30)
weight(wt)
****** analyze by data expansion for population data
save gradedp, replace
* >1
use gradedp, clear
generate set=1
generate d = y>1
save graded_g1, replace
logit d i_2 i_3 [pweight=prob]
* >2
use gradedp, clear
generate set=2
generate d = y>2
logit d i_2 i_3 [pweight=prob]
**** pool datasets
append using graded_g1
forvalues i=1/3 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      drop i_`i'
}
egen pattset = group(pattern set)
logit d i_1* i_2* i_3* [pweight=prob],  vce(cluster pattern)
```

Appendix B.3d: Population study: STATA code for data expansion for adjacent-category logit model without random effect

```
**** adjacent-category logit
* 3 items, 3 categories
global nitem = 3
global ncat = 3
* generate all possible response patterns
clear
set obs 1
generate pattern=1
forvalues i=1/$nitem{
      expand $ncat
      by pattern, sort: generate y`i'=_n
      replace pattern=_n
}
reshape long y, i(pattern) j(item)
* expand data for estimation using gllamm
generate pers_it = _n
expand $ncat
by pers_it, sort: generate r=_n
forvalues i=1/$nitem{
      forvalues c=2/$ncat {
            generate i_`i'_`c'= (r>=`c')*(item==`i')
      }
}
generate d = (y==r)
* make up model parameters & calculate log-likelihood contributions for
each pattern
matrix a=(0.5, 0.2, 0, 1.0, 0.6, 0.2, 0)
eq abil: r
gllamm r i_* , nocons i(pattern) link(mlogit) expanded(pers_it d o)
eqs(abil) adapt nip(30) from(a) copy eval trace
gllapred loglik, ll
* calculate probabilities of response patterns
generate double prob= exp(loglik)
* without random effects (exact method)
gllamm r i_* , nocons i(pattern) link(mlogit) expanded(pers_it d o)
eqs(abil) init nip(30) weight(wt)
***** make unexpanded data
drop i_*
 /* keep just the "data" */
 keep if d==1
drop d r
save pcreditp, replace
****** analyze by data expansion for population data
* 2 versus 1
use pcreditp, clear
keep if y==1|y==2
generate set=1
generate d = y==2
tabulate item, generate(i_)
save precredit12, replace
* without random effect
logit d i_* [pweight=prob], nocons
```

```
* 3 versus 2
use pcreditp, clear
keep if y==2|y==3
generate set=2
generate d = y==3
tabulate item, generate(i_)
save precredit23, replace
* without random effect
logit d i_* [pweight=prob], nocons
***** pool datasets
use precredit12, clear
append using precredit23
forvalues i=1/3 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      drop i_`i'
}
egen pattset = group(pattern set)
* without random effect
logit d i_* [pweight=prob], nocons
```

Appendix B.3e: STATA code for HADS dataset (depression items only)

```
# To obtain HADS dataset, use R code provided below
# make sure that working directory in both R and STATA are in the same
folder
# to obtain the HADS dataset from R, use
# install.packages("MultiLCIRT")
library(MultiLCIRT)
data(hads)
hads_d <- hads[,c(1,3,4,5,9,13,14)]
write.csv(hads_d, file="hads_d.csv")
****** STATA
***** cumulative 1-PL model using HADS dataset
insheet using hads_d.csv, clear
drop v1
rename item1 ta1
rename item3 ta2
rename item4 ta3
rename item5 ta4
rename item9 ta5
rename item13 ta6
rename item14 ta7
gen one=1
collapse(sum) wt2=one, by(ta1-ta7)
gen id = _n
reshape long ta, i(id) j(item)
drop if ta==.
tab item, gen(i_)
gllamm ta, i(id) weight(wt) l(ologit ologit ologit ologit ologit ologit
ologit) lv(item) f(binom) adapt nip(30)
**** data expansion
insheet using hads_d.csv, clear
drop v1
rename item1 ta1
rename item3 ta2
rename item4 ta3
rename item5 ta4
rename item9 ta5
rename item13 ta6
rename item14 ta7
gen one=1
collapse(sum) wt2=one, by(ta1-ta7)
gen id = _n
reshape long ta, i(id) j(item)
drop if ta==.
tab item, gen(i_)
save hads_0, replace
* >0
use hads_0, clear
generate set=1
generate d = ta>0
gllamm d i_*, nocons i(id) link(logit) fam(binom) adapt nip(50)
weight(wt)
save hads_g1, replace
* >1
use hads_0, clear
generate set=2
```

```
generate d = ta>1
gllamm d i_*, nocons i(id) link(logit) fam(binom) adapt nip(50)
weight(wt)
save hads_g2, replace
* >2
use hads_0, clear
generate set=3
generate d = ta>2
gllamm d i_*, nocons i(id) link(logit) fam(binom) adapt nip(50)
weight(wt)
* pool datasets
append using hads_g1
append using hads_g2
forvalues i=1/7 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      generate i_`i'_4 = i_`i' *(set==3)
      drop i_`i'
}
egen idset = group(id set)
gllamm d i_*, nocons i(idset) link(logit) fam(binom) adapt nip(50)
weight(wt)
* to obtain cluster-robust standard errors
gllamm, cluster(id)
**** data expansion to obtain amalgamated conditional ML estimates
* >0
use hads_0, clear
generate set=1
generate d = ta>0
clogit d i_2 i_3 i_4 i_5 i_6 i_7 [fweight=wt2], group(id) vce(cluster
id)
save hads_g1_cond, replace
* >1
use hads_0, clear
generate set=2
generate d = ta>1
clogit d i_2 i_3 i_4 i_5 i_6 i_7 [fweight=wt2], group(id) vce(cluster
id)
save hads_g2_cond, replace
* >2
use hads_0, clear
generate set=3
generate d = ta>2
clogit d i_2 i_3 i_4 i_5 i_6 i_7 [fweight=wt2], group(id) vce(cluster
id)
* pool datasets
append using hads_g1_cond
append using hads_g2_cond
forvalues i=1/7 {
      generate i_`i'_2 = i_`i' *(set==1)
      generate i_`i'_3 = i_`i' *(set==2)
      generate i_`i'_4 = i_`i' *(set==3)
      drop i_`i'
}
egen idset = group(id set)
clogit d i_2* i_3* i_4* i_5* i_6* i_7* [fweight=wt2], group(idset)
vce(cluster id)
```

```
***** partial credit model using HADS dataset
insheet using hads_d.csv, clear
drop v1
rename item1 ta1
rename item3 ta2
rename item4 ta3
rename item5 ta4
rename item9 ta5
rename item13 ta6
rename item14 ta7
gen one=1
collapse(sum) wt2=one, by(ta1-ta7)
gen id = _n
reshape long ta, i(id) j(item)
*drop if ta==.
save hads0, replace
forvalues i=0/2 {
      use hads0, clear
      generate set = `i'
      local ip = `i' + 1
      keep if ta==`i'|ta==`ip'
      generate d=ta==`ip'
      clogit d i.item [fweight=wt2], group(id)
      save hads`i'_`ip', replace
}
use hads0_1, clear
append using hads1_2
append using hads2_3
sort id item ta set
tab item, gen(i)
tab set, gen(s)
forvalues k=1/7 {
      forvalues l=1/3 {
generate i`k'_s`l' = i`k'*s`l'
      }
}
egen idset = group(id set)
gllamm d i1_* i2_* i3_* i4_* i5_* i6_* i7_*, nocons i(idset) weight(wt)
link(logit) fam(binom) nip(30)
**** conditional ML
*** partial credit model
clogit d i2_* i3_* i4_* i5_* i6_* i7_* [fweight=wt2], group(idset)
vce(cluster id)
*** rating scale model
clogit d i2 i3 i4 i5 i6 i7 [fweight=wt2], group(idset) vce(cluster id)
```

Appendix B.4: Latent response formulation and relationship between three polytomous models

Rasch model, expressed as (and discussed in Chapter 1 of the dissertation):

$$P(Y=0|\theta)=\frac{1}{e^{\theta-\delta}+1},$$

$$P(Y=1|\theta)=\frac{e^{\theta-\delta}}{e^{\theta-\delta}+1},$$

can be re-expressed using the integral notation, as:

$$P(Y=0|\theta)=\int_{-\infty}^{0}\frac{e^{x-(\theta-\delta)}}{(e^{x-(\theta-\delta)}+1)^2}dx,$$

$$P(Y=1|\theta)=\int_{0}^{\infty}\frac{e^{x-(\theta-\delta)}}{(e^{x-(\theta-\delta)}+1)^2}dx,$$

in which we can see that this formulation is consistent with the so called "latent response" formulation, in which the probability of obtaining "1" represents the area above 0 and probability of obtaining "0" represents the area below 0. Below I present extension of this idea to the trichotomous item.

Cumulative logit model for ordinal items

Among the polytomous models, the cumulative 1-PL model is unique in that it can be expressed using the latent response formulation similar to the binary Rasch model. Assume a trichotomous item with two step parameters γ_1 and γ_2 with the constraint of $\gamma_1<\gamma_2$. Further assume that b_0 and b_3 are $-\infty$ and $+\infty$ respectively. Then, category-specific probabilities (see for instance Samejima, 1997, for the 2-PL version) are:

$$P(Y=0|\theta)=\frac{e^{(-(\theta-\gamma_1))}-e^{(-(\theta-b_0))}}{(e^{(-(\theta-\gamma_0))}+1)*(e^{(-(\theta-\gamma_1))}+1)}$$

,

$$P(Y=1|\theta)=\frac{e^{(-(\theta-\gamma_2))}-e^{(-(\theta-\gamma_1))}}{(e^{(-(\theta-\gamma_1))}+1)*(e^{(-(\theta-\gamma_2))}+1)}$$

,

$$P(Y=2|\theta)=\frac{e^{(-(\theta-\gamma_3))}-e^{(-(\theta-\gamma_2))}}{(e^{(-(\theta-\gamma_2))}+1)*(e^{(-(\theta-b_3))}+1)}$$

,

Expressions above can be re-expressed in the integral notation using the the latent-response formulation. We, then, obtain:

$$P(Y=0|\theta)=\int_{-\infty}^{\theta-\gamma_2}\frac{e^{x-((\theta-\gamma_1)+(\theta-\gamma_2))}}{(e^{x-((\theta-\gamma_1)+(\theta-\gamma_2))}+1)^2}dx$$

,

$$P(Y=1|\theta)=\int_{\theta-\gamma_2}^{\theta-\gamma_1}\frac{e^{x-((\theta-\gamma_1)+(\theta-\gamma_2))}}{(e^{x-((\theta-\gamma_1)+(\theta-\gamma_2))}+1)^2}dx$$

,

$$P(Y=2|\theta)=\int_{\theta-\gamma_1}^{\infty}\frac{e^{x-((\theta-\gamma_1)+(\theta-\gamma_2))}}{(e^{x-((\theta-\gamma_1)+(\theta-\gamma_2))}+1)^2}dx$$

.

Note that for all of the three expressions above, the only part of the expression that varies depending on the response are the upper and lower limits of the integral expressions.

In order to visualize how all three of the polytomous models discussed in this chapter are related, we can graph category characteristic curves for the trichotomous item for each of three polytomous models. In the Figure 2.7 below, I assume that the two step parameters are -1 and 1 for all three models.

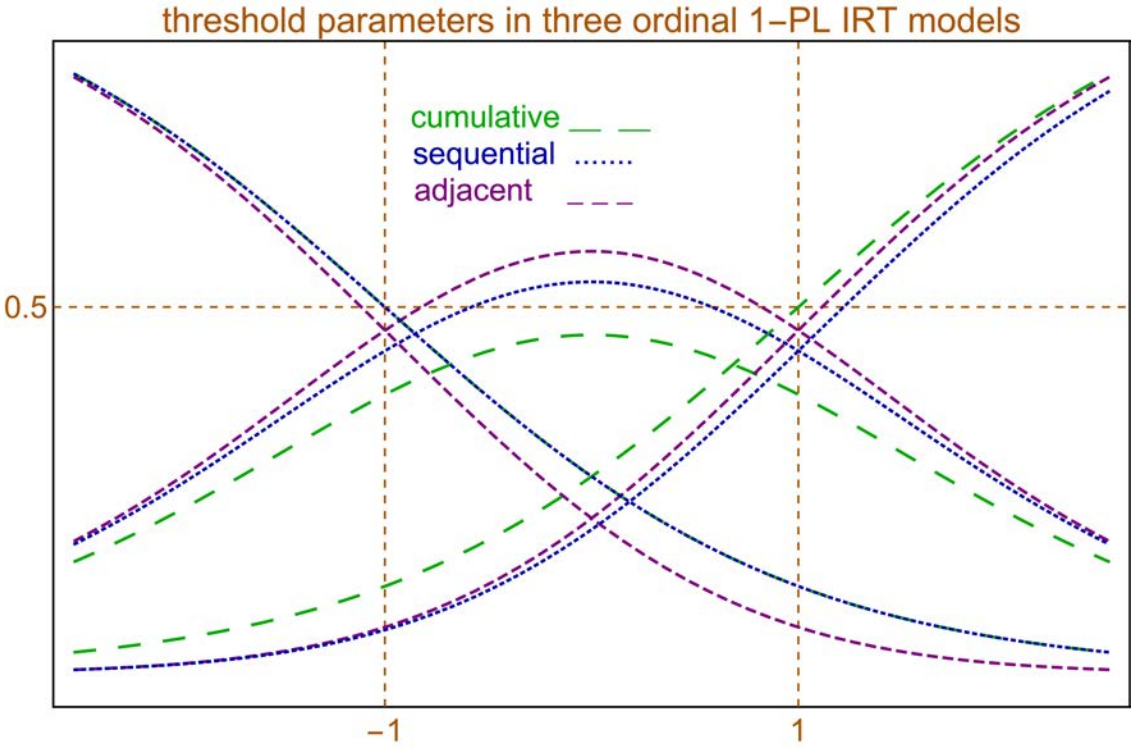


Figure 2.7. Ordinal contrasts for the four-category ordinal variable.

We can see that for the cumulative model, step parameters are locations on the scale (x-axis) on which category-specific probability functions intersect with the 0.5 probability line. For the adjacent-category logit model, these step parameters are locations where functions for the adjacent categories intersect. For the sequential (continuation-ratio) logit model, the interpretation of the first step parameter is identical to the cumulative logit model and the interpretation of the second (i.e., last) step parameter is identical to the adjacent category logit model. This can also be verified by comparing the contrasts in the

continuation-ratio logit model in Figure 2.1 to the cumulative and the adjacent-category logit model in the same figure.

Appendix B.5: Depression items on HADS questionnaire

I still enjoy the things I used to enjoy
- Definitely as much
- Not quite as much
- Only a little
- Hardly at all

I can laugh and see the funny side of things
- As much as I always could
- Not quite so much now
- Definitely not so much now
- Not at all

I feel cheerful
- Not at all
- Not often
- Sometimes
- Most of the time

I feel as if I am slowed down
- Nearly all the time
- Very often
- Sometimes
- Not at all

I have lost interest in my appearance
- Definitely
- I don't take so much care as I should
- I may not take quite as much care
- I take just as much care as ever

I look forward with enjoyment to things
- As much as ever I did
- Rather less than I used to
- Definitely less than I used to
- Hardly at all

I can enjoy a good book or radio or TV programme
- Often
- Sometimes
- Not often
- Very seldom

Chapter 3

Second-order Rasch model

This chapter aims to contribute to the estimation and interpretation of multidimensional item response theory (MIRT) models within the field of psychometrics and latent variable modeling. The main goal of the chapter is to advance the use of the second-order Rasch model.

A second-order (higher-order[70]) Rasch model assumes an overall dimension as a second order factor that explains the covariance between the first-order (component) dimensions. The model allows both dichotomous and polytomous items. Polytomous items can be estimated using the partial credit (Masters, 1982) or the cumulative Rasch (Agresti & Lang, 1993) models. The second-order solution is based on the factorization of the correlation matrix among the component (first-order) dimensions. A related model was proposed by de la Torre & Song (2009) in the 2-PL context using MCMC method for estimation. In contrast, estimates in the model discussed in this chapter are obtained using a marginal maximum likelihood method. The main purpose of the model is to provide an overall and domain scores simultaneously from the single model.

The main contribution of the chapter is to suggest ways of using the model by still preserving the advantages of the Rasch model[71]. There lacks a clear set of guidelines of the use of the second-order model for the Rasch model specifically. Historically, the main challenge in the use of such models were (1) computationally intensive estimation and (2) availability of software. In addition, it is difficult to obtain reliable and meaningful estimates in cases when a variance of one of the dimensions is low relative to other dimensions. In such cases, one first needs to re-assess if the multidimensional structure is appropriate. One, then, can use alternative parameterization of the model to avoid difficulties in the estimation, and guidelines in this chapter provide recommendations on how to achieve such parameterizations with the Rasch model.

I also present alternative parameterization of the Rasch testlet model and extension of that model and discuss how to interpret estimates obtained using second-order and bifactor models. Using a real example dataset, I demonstrate how these models are related,

I start with the motivation and literature review of the factor models with some history that led to developments of multidimensional IRT models. Then I formally present the multidimensional models within the Rasch framework.

In addition to the elaboration on the interpretational differences between the models[72], I provide comparison of these models in terms of their limitations and necessary restrictions for identification.

---

[70] I prefer to use "second-order" instead of "higher-order" although the two are used interchangeably in the literature. "Higher-order" is not specific enough since third-order factor models can also be estimated (see for instance Rijmen, Jeon, Rabe-Hesketh, & von Davier, 2014). A third-order factor is the factor that is extracted from the covariance between two or more second-order factors.

[71] See Wilson (2005) and van der Linden (1994) for the advantages of the Rasch model.

[72] Note that all of the models discussed in this chapter are appropriate when "dimensions" (i.e., constructs, latent variables) are correlated and correlations are meaningful.

3.1.1 Introduction

There is a growing interest in multidimensional IRT models (Adams, Wilson, & Wang, (1997); Brouwer, Meijer, & Zevalkink, 2013; Cai, Yang, & Hansen, 2011; de la Torre & Song, 2009; Jeon, Rijmen, & Rabe-Hesketh, 2013; Reckase, 1985, 2009; Reise, Morizot, & Hays, 2007; Rijmen 2010, 2011; Wang, Wilson, & Adams, 1997; Yao, 2010). These models are seen to be promising in modeling multicomponent data, such as data from large-scale assessments and surveys with complex structures (e.g., multiple dimensions[73], testlets). If such structures exist in the data, then the assumption of local item independence of simpler models is violated.

Often scientific questions arise about the underlying psychometric factor and dimensionality structure of the assessed constructs. Such studies are mainly concerned about the structural validity of the constructs. The main question is whether the instrument (e.g., a depression scale) should be considered as one single scale or if subscales should be distinguished.

Another question is what the hypothetical factor structure of the scales should be. Failing to account for the dimensional structure of the latent variable under consideration might have implications on inferences made from the instruments. For instance, Ackerman (1992) demonstrated that failing to account the dimensional structure of the test might result in incorrect conclusions regarding item bias.[74]

The issue of factor structure needs to be approached not only statistically, but also substantively. Given the prominent use of latent variables in social science, very little work has focused on their nature (Bollen, 2002), and the theoretical status of the latent variable has not been resolved yet (Borsboom, Mellenbergh, & Heerden, 2001). There are still no definite solutions to finding the dimensional structure of constructs for a set of response data and research in this area is greatly needed (Yao, 2010).

In addition, there is a growing interest in the models for score reporting in educational assessments. Policy makers are interested in overall scores. Teachers and parents are interested in component scores. Component scores are more interpretable and provide useful diagnostic and instructionally relevant information to guide instruction or remedial interventions.

Overall scores, in turn, serve better as summaries and allow macro level comparisons such as comparisons of schools within districts. Models that provide both types of scores simultaneously provide an elegant approach and serves all involved stakeholders. A researcher using such a model obtains scores for component and overall dimensions along with model-based standard errors.

---

[73] For instance, "achievement in math" can be seen as being comprised of achievement in, say, geometry, algebra, and arithmetic skills. All three are qualitatively distinct constructs. Math is the overall dimension. Geometry, algebra, and arithmetic skills are component dimensions. A test can be *multidimensional* when one set of items measures one factor while another distinct set of items measures another (qualitatively distinct) factor. This is known as *between-item multidimensionality*. An item can also be considered multidimensional if that item prompts more than one of these constructs. This is known as *within-item multidimensionality* (Wang et al., 1997).

[74] This is due to the fact that between-group differences in performance on the different dimensions of the multidimensional test cannot be disentangled from item bias when the factor structure of the test is not accounted for properly.

3.1.2 Factor models

Latent variables (factors, unobserved variables, constructs, dimensions) are studied in fields such as psychology (e.g. personality, self-esteem), sociology (e.g. attitude regarding abortion), political science (e.g. political efficacy), education (e.g. reading ability), marketing (e.g. brand preference). They can be continuous or categorical. Latent variables considered in this chapter will be assumed to be continuous.

Studies of factor structure have their origins in early research on intelligence and school achievement. The most prominent early theories are Spearman's Two-factor Theory (Spearman, 1904, 1927) and Thurstone's Primary Factor Theory (Thurstone, 1938). These theories represent early steps in the development of factor analysis techniques. Unlike multidimensional item response theory (MIRT) methods, which model categorical manifest variables (e.g. agree vs. disagree, treated vs. control), factor analysis methods traditionally model continuous manifest variables (e.g. age, salary).

Spearman's Two-factor theory assumes that constructs measured by tests are comprised of a general factor (g) and a specific factor. Specific factors are assumed to be unique to each test. The basic idea is to extract the general factor from the correlation matrix[75] of multiple ability tests. Spearman's Model, shown in the Figure 3.1 below, is the simplest factor structure. Arrows from g to V_1-V_6 in Figure 3.1 represent *factor loadings*. Factor loadings are the weights and correlations between each variable and the factor (the higher the load the more relevant the indicator in defining the factor).

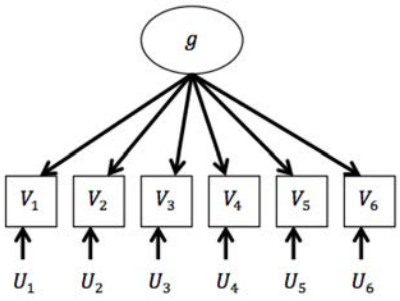


Figure 3.1. Spearman's Model. Each V_1-V_6 are tests, and U_1-U_6 are *uniquenesses* that consist of specificity and random error for each test.

Thurstone's model ("multifactor theory"), shown in Figure 3.2 below, emerged as an alternative and became popular quite rapidly in the factor analytic tradition, used mostly with Kaiser's (1960) method of rotation (varimax criterion). This model assumes multiple orthogonal group factors without a general factor. Oblique (or correlated) factors can also be modeled (Caroll, 1957; Jennrich, 1973). Note that the orthogonality is indicated in Figure 3.2 by the absence of a direct link between F_1 and F_2.

[75] The "observed data" in traditional factor analysis are correlation coefficients.

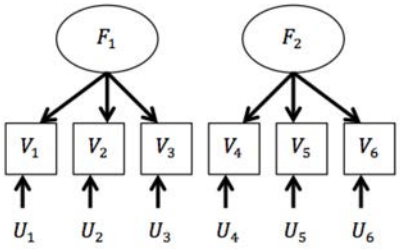


Figure 3.2. Thurstone's Model. A multiple-factor model with two independent group factors (F_1, F_2).

Thurstone was influential in the development of factor methods and formulated these methods in terms of matrix algebra (Carroll, 1993). In earlier work, Hotelling (1933) pointed to the fact that some procedures in factor analysis are related to mathematical problems of finding latent roots and vectors of a matrix and provided a new method for computations (Hotelling, 1933, 1936). These methods are related to principal components analysis, which had been developed by Pearson (1901). Multiple-factor solution, a generic term originated by Thurstone (1931), includes multiple overlapping group factors and avoidance of a general factor (Harman, 1967). In addition to Thurstone's alternative to Spearman's model, Kelley (1928) published one of the earliest works on multiple factors, which demonstrated the existence of "group factors" in addition to the general factor.

As alternatives to Spearman's and Thurstone's models, two different confirmatory factor models—the bifactor (Holzinger & Swineford, 1937) and the second-order factor models (Schmid & Leiman, 1957)—were proposed.

Bifactor models (a.k.a. *general-specific model*, *nested factor model*; Gustafsson & Balke, 1993) assume that each item is influenced by two factors—a general factor and a group factor—and that the general factor and group factors are independent first order (a.k.a., lower order) factors. The structure of a bifactor model is shown in the Figure 3.3 below. In the bifactor model, the general factor is derived directly from the observed variables.

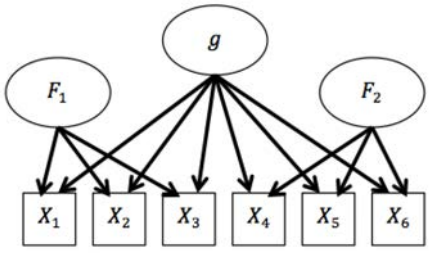


Figure 3.3. Bifactor model proposed by Holzinger & Swineford (1937) with statistically independent factors.

In contrast, a second-order model assumes a general factor as a second order factor that explains the covariance between the first order (component, group) factors. The second-order solution is the factorization of the correlation matrix among the component factors. To put it more simply, "it is from the correlations among the group

factors that the second-order *g* is derived" (Jensen, 1998). The hypothesized structure of a second-order model is shown in the Figure 3.4 below.

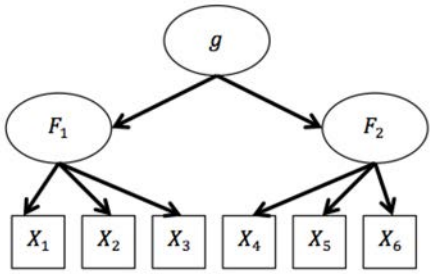


Figure 3.4. Second-order model proposed by Schmid & Leiman (1957).

Both bifactor and second-order methods can be seen as different ways of combining Spearman's and Thurstone's models. Schmid & Leiman (1957, p. 54, see also Yung, Thissen & McLeod, 1999; Rijmen, 2009) demonstrated that bifactor and second-order structures could be transformed into each other or shown to be special cases of a more general structure shown in Figure 3.5 below.

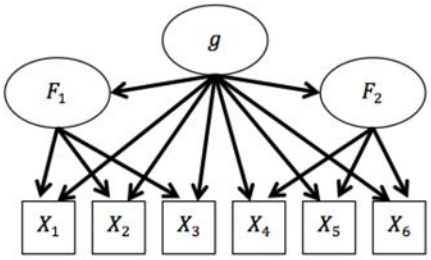


Figure 3.5. Directed acyclic graph of the second-order model with direct effects (Yung et al., 1999) or bi-factor model with conditional independence restriction (Rijmen, 2009). Note that X_1-X_6 are items, but can be replaced by tests (V_1-V_6).

The model in Figure 3.5 is not identified without further restrictions. Eliminating loadings from a general factor (g) to group factors (F_1, F_2), results in the traditional bifactor structure shown in Figure 3.3. Note that the group factors in Figure 3.3 are considered to be orthogonal wheras they are not in Figures 4 and 5 (due to their common dependence on *g*.

An alternative restriction, through the elimination of loadings from a general factor (g) to items (X_1-X_6) in Figure 3.5 also results in the second-order factor structure (Schmid & Leiman, 1957), shown in the Figure 3.4. The so-called Schmid-Leiman transformation essentially attributes the variation in the primary factors to the second-order factors. Note that these factors are necessarily correlated, due to the assumed causal structure.

3.2 Multidimensional item response theory

The assumption of unidimensionality in traditional IRT models is violated when items are measuring more than a single underlying dimension. The inferences from the

unidimensional IRT model are valid only to the extent that the unidimensionality assumption is tested and confirmed.

Multidimensional models in IRT were proposed as "categorical" variants of the factor analysis methods, and confirmatory factor analysis (CFA) in particular (McKinley & Reckase, 1983a; Reckase, 1985). MIRT is sometimes referred to as full information factor analysis (Bock et al., 1988), due to the fact that IRT models are fitted to the raw data directly rather than to summary statistics such as polychoric correlations.

CFA and IRT are both "confirmatory" methods, aimed at confirming a hypothetical structure of the data.[76] FA models for categorical data exist and in some circumstances parameters estimated from these models can be converted into IRT parameters (McLeod, Swygert, & Thissen, 2001). However, unlike IRT models, FA models were not originally developed for categorical data. Multidimensional IRT methods can prove to be quite useful in many practical situations and have attracted significant interest and contributions within the last two decades (see Ackerman, 1996).

The main estimation limitation of MIRT models is that the dimensionality of the integration during the estimation of the models increases exponentially with each dimension. Thus, it becomes increasingly difficult to obtain solutions as the number of dimensions rises. However, as was shown in Gibbons & Hedeker (1992), based on Stuart (1958), a dimension reduction technique can be employed in some circumstances due to an assumed simplicity in the item structure in the model. One common situation which result in many "dimensions" is the use of *testlets,*which will be discussed later in the chapter (Section 3.2.4).

For all of the models presented below, let items be indexed as $i = 1, 2 \ldots I$ and categories as $k = 0, 1 \ldots K$, each item having $K_i + 1$ response categories. Assume that $d = 1, 2, \ldots, D$ domain-specific dimensions (i.e., group factors) and the overall dimension (i.e., general factor) for the second-order and bifactor structures underlie the responses of $p = 1, 2, \ldots, P$ examinees. Domain-specific dimensions will be denoted as $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_d, .., \theta_D)$ and the overall dimension will be denoted as θ_o. Let response patterns be indexed as $r = 1, 2 \ldots R$. The random variable X_{pik} can be expressed such that:

$$X_{pik} = \begin{cases} 1 \text{ if response of person } p \text{ on item } i \text{ is in category } k, \\ 0 \qquad\qquad otherwise \end{cases} \quad \textbf{(1)}$$

Then, $\boldsymbol{X}_{pi} = (X_{pi1}, X_{pi2}, \ldots, X_{piK})$ is a binary vector over *K* categories and $\boldsymbol{X}_p = (\boldsymbol{X}_{p1}, \boldsymbol{X}_{p2}, .., \boldsymbol{X}_{pI})$ is a matrix indicating a response vector for person *p*.

3.2.1 Multidimensional random coefficients multinomial logit model – Multidimensional Rasch model

The Multidimensional random coefficient multinomial logit (MRCML) model (a.k.a., multidimensional Rasch model) was proposed by Adams et al., (1997) to analyze

[76] Within MIRT, there are two broad types of models in terms of a hypothesized cognitive process and the formulation of the likelihood: compensatory and non-compensatory models. Models discussed in this paper fall into the former category. An example of the noncompensatory (also called *conjunctive* or *partially noncompensatory*) model is Multicomponent Latent Trait Model (MLTM) of Embretson (1980). See DiBello et al. (2007).

the data in the multidimensional item response modeling framework. The model assumes that a set of *D* domain-specific dimensions underlie the examinee responses. The general formulation of the model (in terms of the factor structure) is "open"— the model does not make the restrictive assumptions that the bifactor and second-order factor models make. MRCML uses two matrices, namely the scoring matrix **B**,[77] in which relationships between items and dimensions are represented, and a design matrix **A**,[78] in which relationships between items and item or step parameters are represented.
Then, the MRCML model can be formulated as:

$$\mathrm{P}\left(X_{pik}=1;\mathbf{A},\mathbf{B},\boldsymbol{\xi}\middle|\boldsymbol{\theta}_p\right)=\frac{\exp\left[\mathbf{b}'_{pik}\boldsymbol{\theta}_p+\mathbf{a}'_{ik}\xi\right]}{\sum_{\mathrm{k}-1}^{\mathrm{K_i}}\exp\left[\mathbf{b}'_{pik}\boldsymbol{\theta}_p+\mathbf{a}'_{ik}\xi\right]}, \qquad \textbf{(2)}$$

in which $\boldsymbol{\theta}_p=(\theta_{p1},\theta_{p2},\ldots,\theta_{pD})$ is a Dx1 vector of dimension parameters for person p; $\mathbf{b}'_{pik}$ is a 1xD vector of person p for item i and category k representing relationship with the dimension d; $\boldsymbol{\xi}$ is a mx1 vector of item parameters, and $\mathbf{a}'_{ik}$ is a 1xm vector that represents the link between items and corresponding item or step difficulties. Item parameter vector $\boldsymbol{\xi}$ is considered fixed[79]. The vector of ability parameters $\boldsymbol{\theta}_p$ is considered random and assumed to have the multivariate normal distribution with a mean of μ and a variance-covariance matrix of Σ, both of which are fixed unknown parameters. The software ConQuest (Adams, Wu, & Wilson, 2015) implements the MRCML as its core structure. ConQuest estimates these parameters by maximizing the marginal maximum likelihood (MML) for a set of *R* response patterns, as in

$$L(\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{X})=\prod_{r=1}^{R}\int_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta},\boldsymbol{\xi})\exp[x'_r(\mathbf{B}\boldsymbol{\theta}+\mathbf{A}\boldsymbol{\xi})]\,dG(\boldsymbol{\theta};\boldsymbol{\alpha}), \qquad \textbf{(3)}$$

in which $\boldsymbol{\alpha}=(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $G(\boldsymbol{\theta};\boldsymbol{\alpha})$ is the cumulative distribution function for the multivariate normal distribution. See Adams et al. (1997) for a more detailed presentation of the model and Briggs & Wilson (2003) for an explication of the model.

Thus, using this multidimensional Rasch model, apart from item difficulties and dimension-specific person abilities[80], modeled parameters include the variance for each dimension and correlations between dimensions. Figure 3.6 shows the factor structure of the three-dimensional Rasch model. In that model, all item loadings for each dimension are constrained to unity and we denote variances and covariances by ψ and ζ respectively.

---

[77] A response of person p in category k on factor*s* of item $i$is scored b_{piks}, thus $\boldsymbol{b}_{pik}=(b_{pik1},b_{pik2},..,b_{pikD})$ representing scoring across D factors and $\boldsymbol{B}_{pi}=(\boldsymbol{b}_{pi1},\boldsymbol{b}_{pi2},..,\boldsymbol{b}_{piD})$ representing scoring matrix for item i and person p, and $\boldsymbol{B}_p=(\boldsymbol{B}_{p1},\boldsymbol{B}_{p2},\ldots,\boldsymbol{B}_{pI})$ representing scoring for person p across I items.

[78] Items are described by $\boldsymbol{\xi}=(\xi_1,\xi_2,..,\xi_m)$ vector of m item parameters. Let $\boldsymbol{a}_{ik}=(a_{ik1},a_{ik2},\ldots,a_{ikm})$ indicate design vector that describes the empirical characteristics of the response category$k$ of item$i$, $\boldsymbol{A}=(\boldsymbol{a}_{11},\boldsymbol{a}_{12},\ldots,\boldsymbol{a}_{1K_1},\boldsymbol{a}_{21},\ldots,\boldsymbol{a}_{2K_2},\ldots,\boldsymbol{a}_{IK_{\square}})$ then being the design matrix.

[79] See De Boeck (2008) for models that assume that item parameters are "random."

[80] Note that person "abilities" are random variables and not parameters when MML is used (thus person abilities are "predictions").

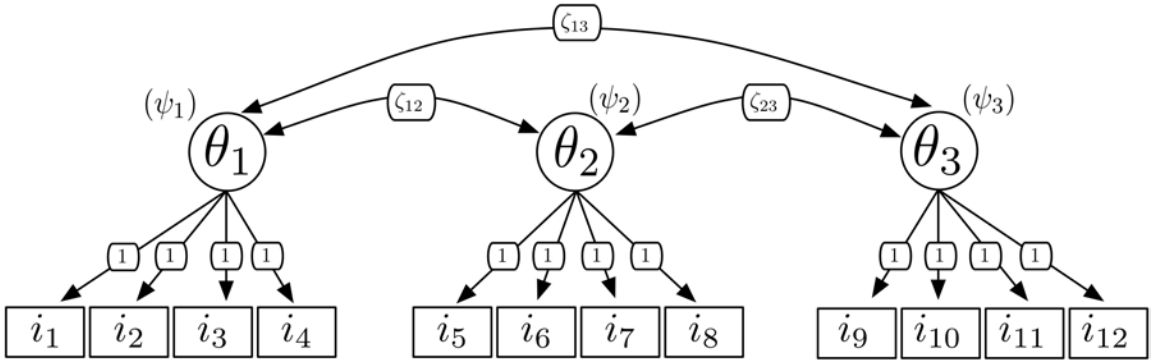


Figure 3.6. Three-dimensional Rasch model.

The advantage of the flexible factor structure (e.g., MRCML) over the other two restricted structures presented next (bifactor and second-order structures) is that items do not have restrictions on the number of factors on which it can load. In other words, items can freely load on any number of items as long as conditions for the identification of the model detailed in Volodin & Adams (1995) are satisfied.

3.2.2 Full information bifactor model

Gibbons & Hedeker (1992) (see also Gibbons et al.,2007) adapted the structure originally proposed by Holzinger & Swineford (1937) to the IRT family of models, for dichotomous and polytomous models respectively. The full information bifactor model changes the traditional MIRT model with simple structure items by adding a general factor that all indicators are supposed to load on. Thus, each item is an indicator of an overall dimension and one of the *D* domain-specific dimensions. Dimensions are assumed to be uncorrelated[81] and normally distributed. See also Cai et al. (2011) for a formulation of the bifactor structure with various dichotomous and polytomous MIRT models.

The main advantage of the bifactor structure in terms of estimation, as opposed to less restricted MIRT models, is the reduction in the dimensionality of integration during the estimation[82]. When items have a simple structure, the covariance matrix is determined by the loadings of the domain-specific dimensions, and the probability of the particular pattern ($\boldsymbol{X}_p$) can be evaluated in terms of one-dimensional integrals (Stuart, 1958), as in:

$$P(\boldsymbol{X}_p|\boldsymbol{\theta}) = \prod_{d=1}^{D} \int_{R(d)} \prod_{i=1}^{I} \prod_{k=1}^{K_i} \Phi_{ik}^{w_{id}u_{pik}}(\theta_d) g(\theta_d) d\theta_d \,, \qquad \textbf{(4)}$$

in which $w_{id} = 1$ when item i has a non-zero loading on dimension d, and 0 otherwise. The above equation can be generalized to the bifactor model (Gibbons & Hedeker, 1992) by adding the overall dimension (θ_o) and thus requiring only a series of two-dimensional integrations as in:

$$P(\boldsymbol{X}_p|\theta_o, \boldsymbol{\theta}) = \int_{R(\theta_o)} \left[\prod_{d=1}^{D} \int_{R(\theta_d)} \left(\prod_{i=1}^{I} \prod_{k=1}^{K_i} \Phi_{ik}^{w_{id}X_{pik}}(\theta_o, \theta_d) \right) g(\theta_d) d\theta_d \right] g(\theta_o) d\theta_o. \qquad \textbf{(5)}$$

[81] See Paek, Yon, Wilson & Kang (2009) for the case with correlated domain-specific dimensions.

[82] However, not many estimation routines exploit this advantage.

Thus, for the bifactor model, the dimensionality of the integration depends on the hierarchy of the factors and not on the number of factors. Note, however, that this advantage of the bifactor model (over the MIRT model) applies to the estimation aspect of the model only, and might be void of any use if the substantive theory on which the items are based is not consistent with a bifactor structure.

Yung et al. (1999) (see also Rijmen 2009) showed how second-order and bifactor models are identical. Rijmen (2009) showed that second-order factor model is "nested" within the bifactor model (is a special case of the bifactor model).

The bifactor model is sometimes referred as hierarchical model (see for instance McLeod et al., 2001). However, this should not be confused with the set of models aimed for handling the nested (i.e., multilevel) data, to which the bifactor model can be extended straightforwardly.

3.2.2.1 Rasch testlet model and possible extensions

The Rasch testlet model presented in Wang & Wilson (2005) is a special case of the bifactor model. Its underlying structure is depicted in the Figure 3.7 below. Note that all of the item loadings for both overall and domain-specific dimensions are constrained to unity.

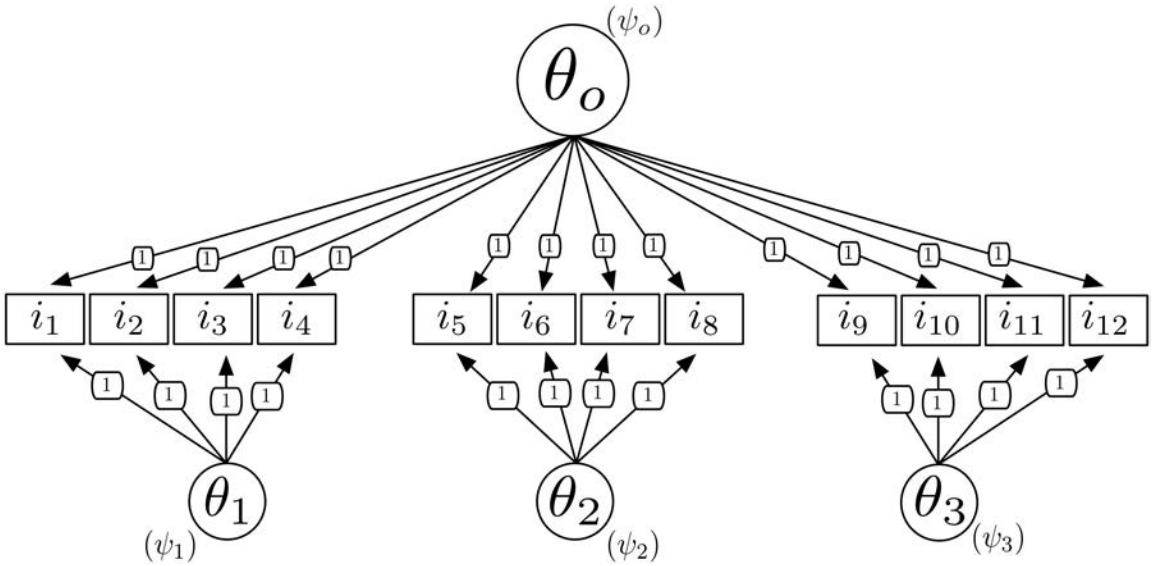


Figure 3.7. Rasch testlet (bifactor) model with three group factors. In this example, we estimate four variances ($\psi_o, \psi_1, \psi_2, \psi_3$) and 12 item intercepts (i.e., item difficulties).

The Rasch testlet model can be also presented using the parameterization shown in Figure 3.8 below. In this parameterization, we still estimate 12 item intercepts. Instead of estimating four variances, however, we estimate four sets of loadings ($\alpha_o, \alpha_1, \alpha_2, \alpha_3$) and fix all four variances to unity.

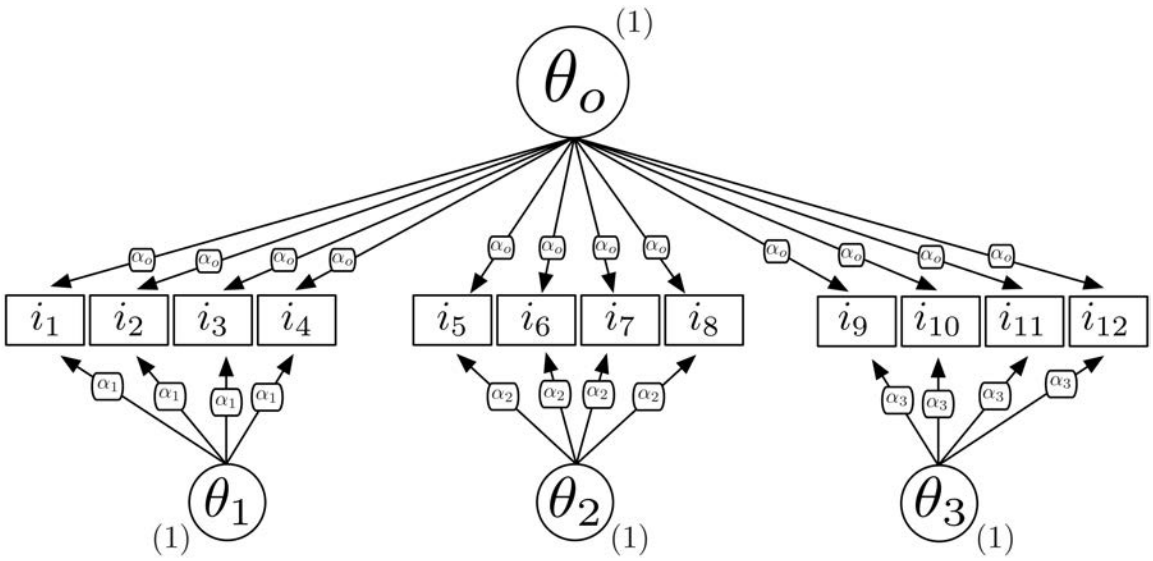


Figure 3.8. Rasch testlet (bifactor) model with alternative parameterization.

One intuitive extension, then, would be to allow loadings of items on overall dimension to vary across domains. The underlying structure of such extension is presented in the Figure 3.9 below.

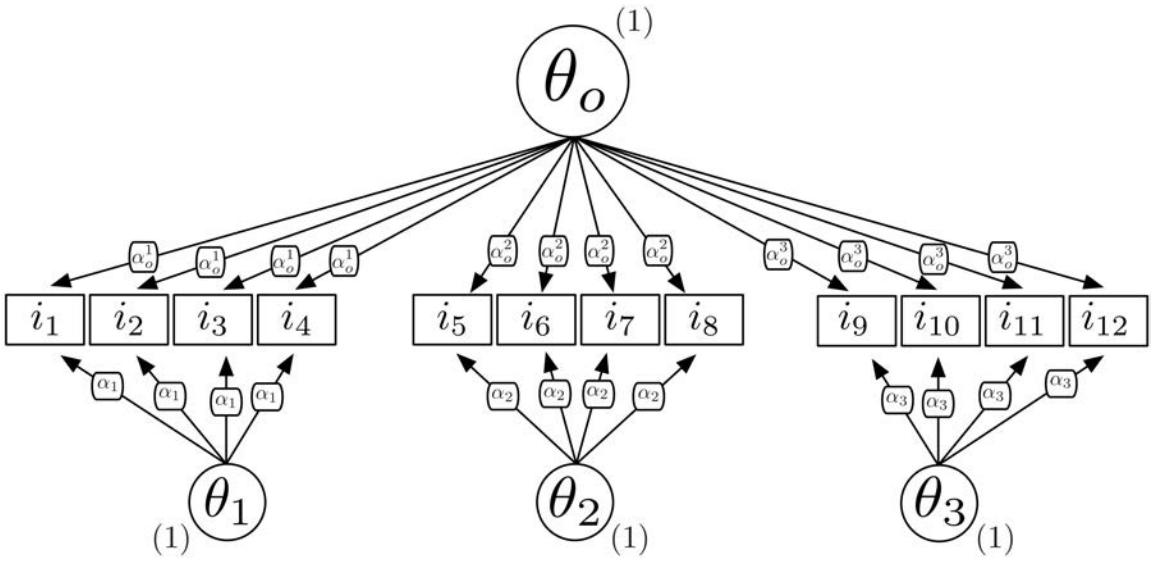


Figure 3.9. Extended Rasch testlet (bifactor) model with alternative parameterization.

Note that instead of estimating a single set of loadings (α_o) on overall dimension for all items, we estimate three sets of loadings ($\alpha_o^1, \alpha_o^2, \alpha_o^3$).

The extended bifactor Rasch model is closely related to the second-order Rasch model. Estimates for item intercepts and the number of parameters are identical in the two, as will be shown in the demonstration example below.

3.2.3 A Second-order Rasch model

In the second-order factor model, domain-specific (first-order) dimensions (i.e., F_1 and F_2 in Figure 3.5) serve as indicators. This formulation attempts to aggregate multiple domain-specific dimensions using correlations between these dimensions. Domain-specific dimensions are expressed as linear functions of the overall dimension, that is,

$$\theta_d = \lambda_d \theta_o + \epsilon_d, \quad \textbf{(6)}$$

in which λ_d indicates the (latent) regression coefficient of regressing the domain-specific dimension (θ_d) on the overall dimension (θ_o). The error term ϵ_d represents the part of θ_d that is unique, and all ϵ_d are assumed to be independent of one another and of θ_o. The structure of the second-order model is similar to the testlet model presented in Bradlow, Wainer, & Wang, 1999 (see for instance Wainer, Bradlow & Wang, 2007).

Note that the second-order model does not allow within-item multidimensionality (such models are not identified). Also note that the model with only two first-order factors is not identified without additional restrictions. This is due to the fact that when only two group factors are modeled, the estimates of the regression coefficients are not unique. This is because different pairs of regression coefficients can result in the same correlation coefficient between two domain-specific dimensions. One particular restriction, then, is to constrain the regression coefficients to be equal[83].

In the second-order model, an a priori specified number of correlated domain-specific dimensions are estimated first, and the overall dimension is estimated as the second-order factor. Unlike in the bifactor model, the overall dimension in the second-order model does not have a direct "effect" on respondent's performance and the performance of respondents is accounted for solely by the domain-specific dimensions. Thus, the structure of the bifactor model can be considered as a "top-down" structure while the structure of the second-order model can be considered as a "bottom-up" structure (Jensen, 1998).

In the second-order Rasch model, the dimensions shown in Figure 3.6 are considered first-order dimensions. These first-order dimensions, in turn, serve as indicators of the second-order (overall) dimension. Instead of three covariance estimates, however, we obtain three regression estimates—loadings of the component dimensions on the overall dimension. This model is shown in the Figure 3.10 below. In Figure 3.10, regression estimates (arrows from θ_o to $\theta_1, \theta_2, \theta_3$) are denoted as γ and variances are denoted as ψ. In order to estimate these γ's (loadings of the component dimensions on the overall dimension) we constrain the variance of the overall score to unity.[84]

[83] In this case, the overall dimension we obtain from the second-order Rasch model is identical to the average of two dimensions we obtain from the two-dimensional Rasch model. However, we in addition obtain model-based standard errors for the overall dimension when using the second-order Rasch model.

[84] Alternatively, we could constrain one of these regression coefficients (e.g., γ_1) to unity and free the variance of the overall score. This may be preferable for interpretation under certain circumstances.

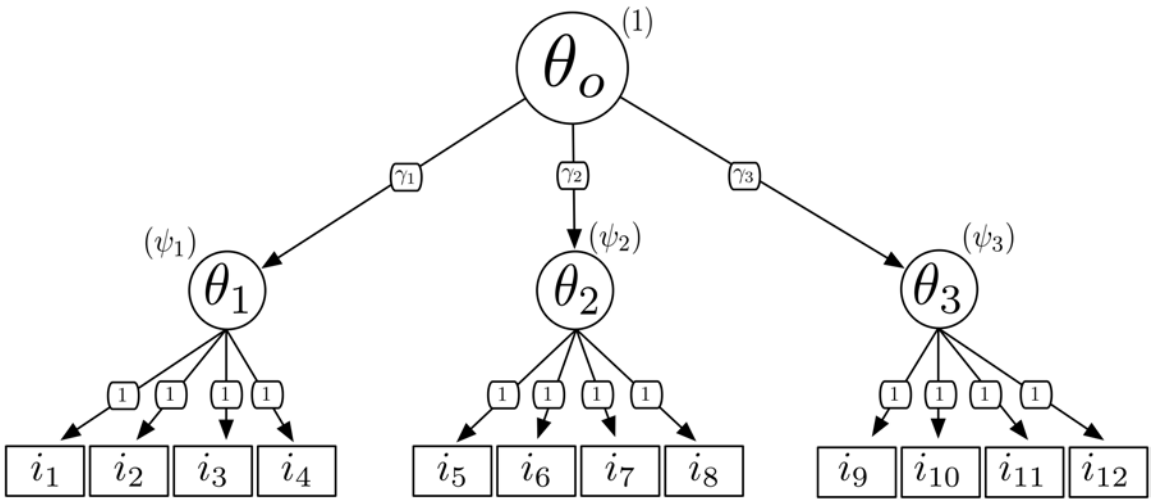


Figure 3.10. Structure of the second-order Rasch model

However, the model in the Figure 3.10, in its current form may be difficult to estimate when unique variances in one or more of the domain-specific dimensions is low. Before I show an alternative formulation, however, note that it is possible to reparameterize the multidimensional Rasch model (shown in Figure 3.6 above) as the model shown in Figure 3.11 below. In this (Figure 3.11) formulation of the multidimensional Rasch model, we estimate item loadings for each dimension $(\alpha_1, \alpha_2, \alpha_3)$—which are identical for items within dimensions—and instead constrain the variances of dimensions to unity. Variances in Figure 3.6 can be easily reproduced using loadings in Figure 3.11 by taking the squares.

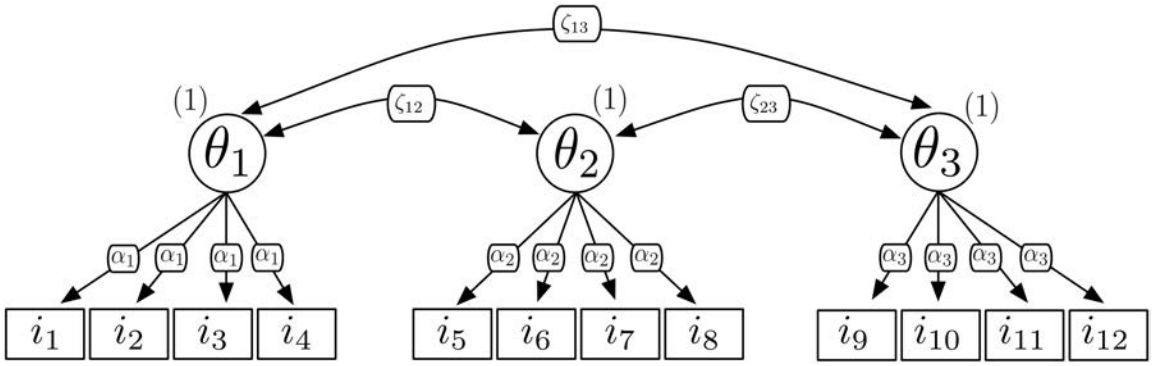


Figure 3.11. Alternative parameterization of the multidimensional Rasch model.

The alternative parameterization of the second-order Rasch model follows the same logic of the parameterization shown in Figure 3.11. Such parameterization of the second-order model (shown in Figure 3.12) avoids difficulties in estimation of the model when some of the domain-specific dimensions are correlated very highly (i.e., high loadings on the overall dimension) and thus result in very small unique domain-specific variances. When these variances are close to zero, the software may need to fix them to zero to avoid negative variance.

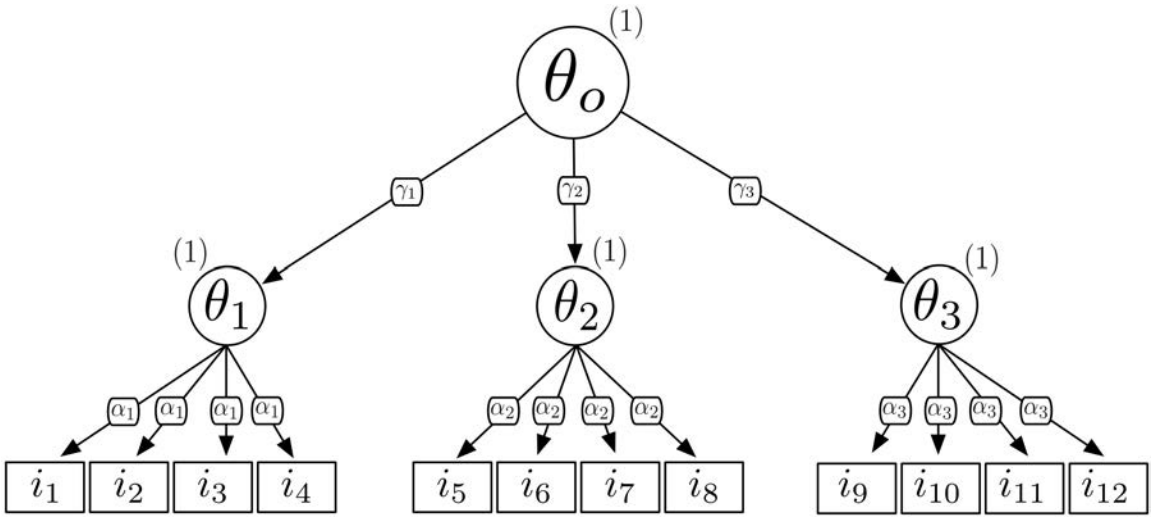


Figure 3.12. Second-order Rasch model with alternative parameterization.

When there are three domain-specific dimensions, the number of parameters in the second-order Rasch model is identical to the number of parameters we would obtain from the multidimensional Rasch model.

Bifactor and second-order models presented above can be used to determine the so called "testlet effect" when the response data contains testlets[85] . The most common approaches of modeling the response data with testlets are discussed next.

3.2.4 Testlets

Items are locally independent of each other if, once we know the respondent and item locations, there is no more information needed to calculate their joint probability. This assumption can be violated when several items have a relationship over and above what would be indicated by their respective difficulties, and the respondents' abilities. One such case appears when several items relate to the same stimulus material, such as in a paragraph comprehension test. In this case, understanding or misunderstanding the paragraph can improve or worsen the performance on all items in the set, but not on other items in the test.

Thus, responses to items belonging to the same testlet are conditionally dependent, a point first made by Rosenbaum (1988). Note that the motivation to account for such dependence is somewhat different from accounting for dependence due to the multidimensionality discussed above. In the case of multidimensionality, the dependencies between items due to the group factors are of the primary interest and the necessity to account for the dependence stems not only from the statistical requirement of the model but also from the scientific interest. In the cases with testlets, in which statistically similar issues arise (i.e., a residual dependency is likely, and needs to be accounted for), the dependencies within testlets are regarded as nuisances (or random components representing person-testlet interaction), and thus the necessity for accounting

[85] Group of items that share some stimulus (stem or content) in common are known as *item bundles* (Rosenbaum, 1988; Wilson & Adams, 1995) or testlet (Wainer & Kiely, 1987; Bradlow et al., 1999). Testlets are used in many educational contexts (see Wainer et al., 2007).

for such dependencies is primarily a statistical issue, rather than a substantive scientific issue.

One simple approach for handling dependency due to the testlet structure is the so-called "sum-score" method. This approach essentially considers the testlet as a single polytomous item (Wilson & Adams, 1995). As a result, the conditional independence between item bundles (polytomous "items") can be assumed. Note that in this approach, responses to individual items within the bundle become "invisible". Also, complications in the interpretation might arise if the items within testlets are polytomous themselves. For discussion of these issues see Wilson, (1988) and Wilson & Adams, (1995).

An alternative method to account for testlets is by introducing a random effect for each testlet (Bradlow et al., 1999; Scott & Ip, 2002; Wainer et al., 2007) using the bifactor and second-order factor models we described above. As a result, items within the particular testlet will be regarded as independent conditional on the testlet-specific random effect. A slightly modified version of the testlet model (Bradlow et al., 1999) described in Tuerlincks & De Boeck (2004) is presented below.[86]

Assume the simplest case in which the test with I items contains a single testlet with I^* items within the testlet ($1<I^*<I$). Let an item predictor $Z_{i(I+1)}$ be defined as

$$Z_{i(I+1)} = \left\{ \begin{matrix} 1 & \text{if item } i \text{ belongs to testlet} \\ 0 & \text{otherwise} \end{matrix} \right\}. \quad \textbf{(7)}$$

Then, the response probability of the person p on the item i can be expressed as

$$P\left(X_{pi} = x_{pi}\right) = \frac{\exp\left(\theta_{p0}+Z_{i(I+1)}\theta_{p1}-\beta_i\right)}{1+\exp\left(\theta_{p0}+Z_{i(I+1)}\theta_{p1}-\beta_i\right)}, \quad \textbf{(8)}$$

in which θ_{p1} is the person-specific testlet effect, generally assumed to be normally distributed with the mean of zero and the variance of σ_{θ_1}. The covariance between the random intercept (θ_{p0}) and the testlet effect (θ_{p1}) is generally assumed to be zero. See Rijmen (2009) for the example of the use of bifactor and second-order models.

3.2.5 Interpretation of factors in a bifactor model[87]

The overall dimension in the bifactor model ("g" in Figure 3.3) represents common variance shared by all of the items in the model. Domain-specific dimensions (F_1 and F_2 in Figure 3.3) represent the variance that is unique to the items loading on the particular domain. Thus, the overall dimension is independent of domain-specific dimensions by definition. However, group factors may be allowed to be correlated if such a model is identified (Little, 2013)—see for instance Paek, Yon, Wilson & Kang (2009) for the case with Rasch testlet model.

Recall that in the 2-PL bifactor model, if the loadings of items on the overall dimension are substantively larger than loadings on domain-specific dimensions, we can conclude that items are unidimensional. In the Rasch framework, since all loadings are

[86] Following Tuerlincks & De Boeck (2004), unlike Bradlow et al. (1999), slope parameters (discriminations) are assumed to be known and logit link used instead of the probit link.

[87] See Briggs & Wilson (2003), Adams et al. (1999), and Reckase (2009) for the interpretation of dimensions in the MIRT models.

fixed to unity, the criteria of judging the dependence induced by domain-specific dimensions (or, testlets) is based instead on the variances of the domain-specific factors (see for instance Wang & Wilson, 2005).

The most important distinction of the bifactor model from the multidimensional IRT model is that dimensions are considered orthogonal from each other. This implies that domain-specific dimensions are defined as being independent to the overall dimension. For instance, in a mathematics test measuring geometry, algebra, and probability, the overall dimension obtained from the bifactor model represents the mathematics ability that is orthogonal to the unique geometry, algebra, and probability dimensions. Put another way, the geometry, algebra, and probability dimensions represent the part of those domains that are orthogonal to the overall mathematics dimension. Therefore, MIRT or UIRT models are more appropriate for such settings. In turn, the bifactor model is ideal when domain-specific dimensions are considered nuisance dimensions and need to be accounted away as in the case with testlets.

3.2.6 Interpretation of dimensions in the second-order model

Interpretation of overall dimension in the second-order model is similar to the interpretation in the bifactor model. The main difference is in how the second-order factor (overall dimension) is extracted. In the second-order model, the overall dimension is extracted from the common variation among first-order factors. Thus, items are involved indirectly (unlike bifactor model, in which items are involved directly).

Covariance between lower-order factors, however, can be estimated post-estimation by multiplying loadings of the relevant factors with each other and multiplying the result with the variance of the second-order factor (which is constrained to unity for identification, as discussed above).

Domain-specific dimensions in the second-order model are implicitly assumed to be correlated. However, this correlation is attributed to the overall dimension (second-order factor). In other words, it is assumed that the correlation between first-order factors is due to the second-order factor. Thus, although the overall dimension is interpreted similarly in the second-order and bifactor models, interpretation of domain-specific dimensions in the second-order model is different from the bifactor model and is similar to the interpretation of dimensions in the MIRT model.

3.3 Demo dataset: ADM assessment

Details of the ADM assessment were presented in Chapter 1 of this dissertation. For demonstration purposes, I use items from the ADM Post-test 2013 instrument measuring only three of the dimensions: DAD (Data Display), COS (Conceptions of Statistics), and INI (Informal Inference). I analyze this response data using (1) three-dimensional Rasch; (2) second-order Rasch; (3) bifactor Rasch; and (4) extended bifactor Rasch models.

There are a total of 19 polytomous items, with five, eight, and six items measuring DAD, COS, and INI dimensions respectively. Item and step fit statistics are all within the accepted lower and upper bounds of 3/4 and 4/3 (Adams & Khoo, 1996) when

analyzed using three-dimensional Rasch model (using adjacent-category logit link[88]). EAP/PV reliabilities are 0.85, 0.79, and 0.86 for DAD, COS, and INI dimensions respectively. (Note, this is not surprising, as these items have already been selected partly on the basis that they are fitting the model reasonably well.)

Second-order and bifactor Rasch models were obtained using Mplus (Muthen & Muthen, 2011), which has only cumulative logit link option in the latest version (for the modeling of polytomous responses). Therefore, the three-dimensional Rasch model was also estimated using the cumulative logit link.

Table 3.1 below shows variances and correlations between three dimensions obtained using the cumulative logit link from the three-dimensional Rasch model. For comparison, Table 3.2 shows variances and correlations between dimensions using the adjacent-category logit link. Notice that correlations from the two models are not very different, although the variances are quite different. The difference in the two variances can be attributed to the difference in the variances of item parameters between the two models. In particular, the variance of step parameters in the cumulative Rasch model is higher than the variance of parameters in the partial credit model. Recall that in the cumulative Rasch model, item parameters are following a strict ordering (see Chapter 2 of this dissertation and Adams, Wu, & Wilson, 2012, for a detailed explanation). In other words, step parameters in the cumulative Rasch model are strictly increasing. As a result, within each item, step parameters in the cumulative Rasch model in general will tend to have higher variance than step parameters in the partial credit model, especially for items with higher number of response categories.[89]

Table 3.1. Correlations between domains and variance for each domain using the cumulative Rasch model

	DAD	COS	INI
DAD			
COS	0.88		
INI	0.99	0.91	
variance	1.79 (0.15)	1.35 (0.12)	2.01 (0.15)

Table 3.2. Correlations between domains and variance for each domain using the partial credit model

	DAD	COS	INI
DAD			
COS	0.84		
INI	0.99	0.87	
variance	0.28 (0.03)	0.49 (0.05)	0.59 (0.05)

[88] See Chapters 1 and 2 of this dissertation for more on adjacent-category and cumulative logit link functions.

[89] For instance, when variances (from the cumulative Rasch and partial credit models) of only first step parameters (of all 19 items) are compared, the variance of the step parameters from the partial credit model is slightly higher. However, variability of step parameters in the cumulative Rasch model is higher when all step parameters are investigated together. This is due to higher variability of step parameters within each item in cumulative Rasch model.

Table 3.3 below shows fit statistics for four of the models using 30 quadrature nodes for each dimension as implemented in Mplus 7 (Muthen & Muthen, 2011). It also shows fit information for the unidimensional Rasch model.

Table 3.3. Model fit statistics (using cumulative logit link)

MODEL	deviance	AIC	BIC	# of parameters
three-dimensional Rasch	30792.3	30948.3	31313.5	78
second-order Rasch	30798.3	30954.3	31319.5	78
testlet-Rasch	30840.4	30992.4	31431.7	76
extended bifactor Rasch	30795.6	30951.6	31402.5	78
unidimensional Rasch	30858.6	31004.7	31346.5	73

When the model is a special case of the less restricted model (such as unidimensional Rasch model being nested in multidimensional Rasch model), the difference in deviances is assumed to have a chi-square distribution with the difference in the number of parameters as degrees of freedom. Thus, we can statistically test if the less restricted model fits the data significantly better than the simpler (i.e., more restricted) model. When the models are not nested, this likelihood ratio test cannot be used. Instead, model fit indices such as *Akaike Information Criterion* (AIC; Akaike, 1974) and *Bayesian Information Criterion* (BIC; Schwarz, 1978) can be used to compare the models in terms of fit. Using AIC and BIC[90] shown in Table 3.3, we can conclude that the three-dimensional Rasch model is the best fitting model when applied on this data.

Estimates from the three-dimensional Rasch model are shown in Figure 3.13 below. Correlations between dimensions are indicated on double-headed arrows between circles. In this model, each item measures only one dimension with items 1-5 measuring DAD dimension, and items 6-13 and 14-19 measuring COS and INI dimension respectively.

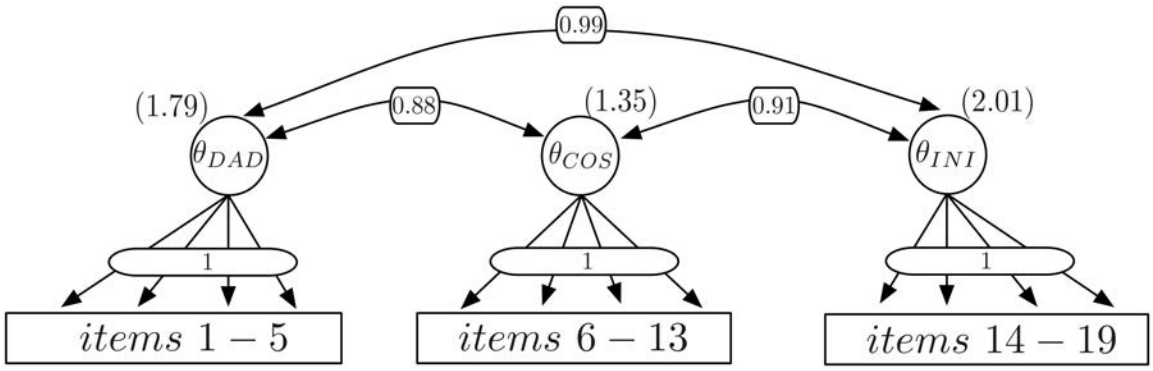


Figure 3.13. Results from the three-dimensional Rasch model.

The correlation between DAD and INI dimensions is very high and could well be modeled as a single dimension. However, this is not the primary focus of this example and therefore will not be investigated further.

[90] Smaller is "better"

Figure 3.14 below shows estimates obtained using the second-order Rasch model (with alternative parameterization). Note that after extracting the overall dimension, loadings of DAD and INI items on respective domain-specific dimensions are very small. Correlation of the DAD EAP scores obtained from the MIRT model and the second-order Rasch model is high at 0.999 (similarly for the COS and INI dimensions).

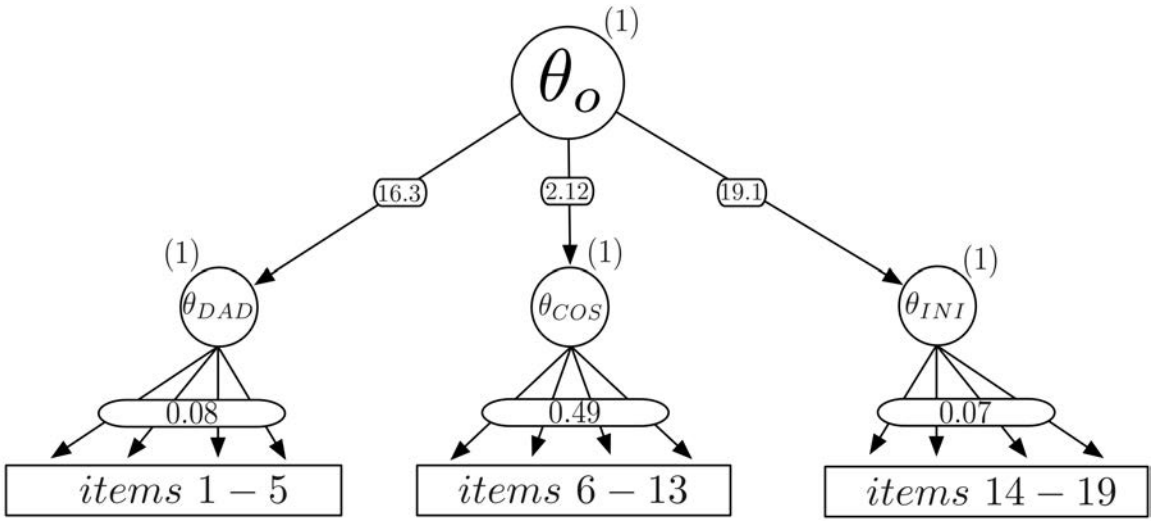


Figure 3.14. Results from the second-order Rasch model.

Figure 3.15 below shows estimates obtained from the Rasch testlet model with the alternative parameterization. Notice that the number of parameters in this model is less than the number of parameters in the two previous models shown above.

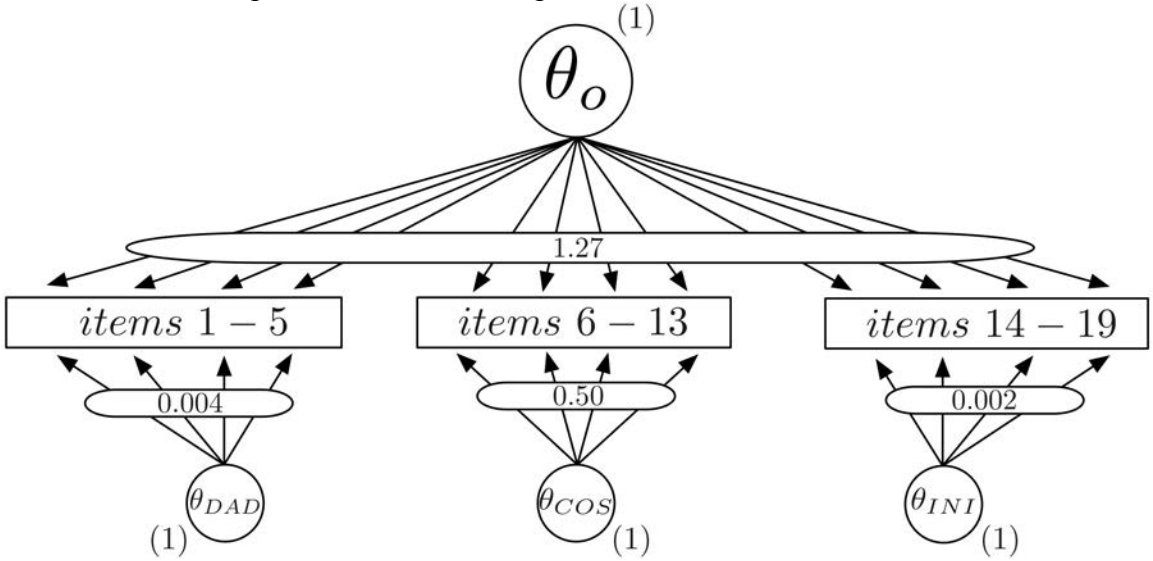


Figure 3.15. Results from the Rasch testlet model with alternative parameterization.

Figure 3.16 below shows estimates obtained from the extended Rasch bifactor model (with alternative parameterization). One can easily approximate the three sets of loadings on the overall dimension in the extended bifactor Rasch model (Figure 3.16) using estimates from the second-order Rasch model (Figure 3.14). For instance, the product of 0.08 and 16.3 (=1.304) in Figure 3.14, which is quite close to the loading of

the DAD items on the overall dimension in Figure 3.16 (1.33), and which is also true for the other two subdimensions. The correlation of EAP scores among the overall dimensions from these two models is 1.00.

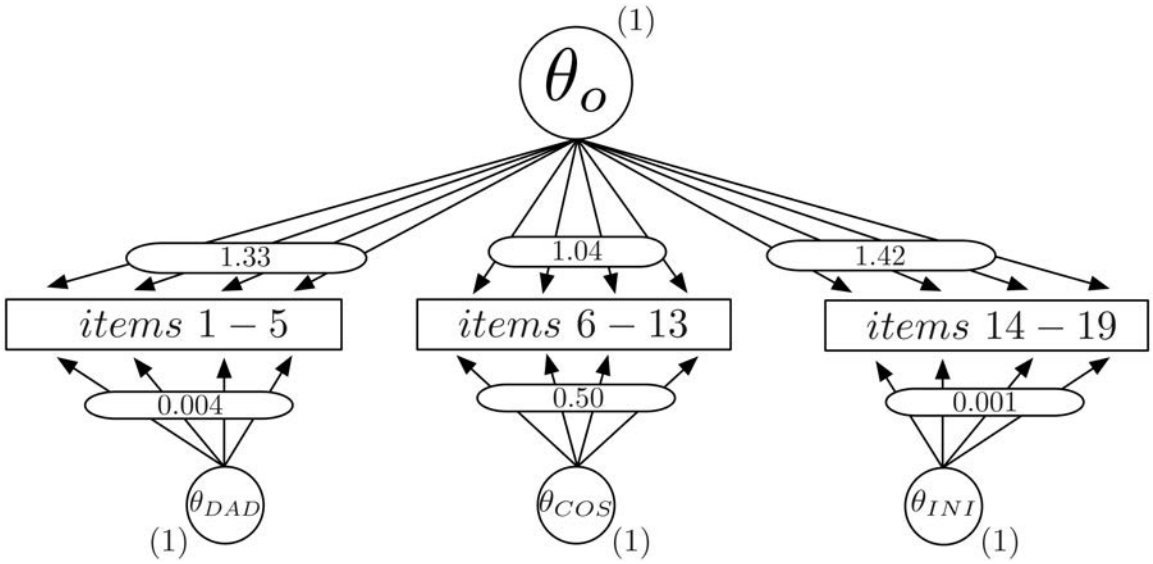


Figure 3.16. Results from the extended Rasch testlet model.

The correlations provided by the three-dimensional Rasch model (Table 3.1) range from 0.88 to 0.99 and indicate that dimensions are highly correlated. In particular, we found that the correlation between DAD and COS dimension is 0.88, the correlation between the COS and INI dimensions is 0.91, and the correlation between DAD and INI is 0.99. High correlations in this dataset are not surprising since all of these domains are parts of a single curriculum.

The second-order Rasch model provides regression coefficients instead of correlations. We can observe that the loadings of the DAD and INI dimensions (on the overall dimension) are very high (when compared to that of COS). After the common variance between these two first-order dimensions is extracted, the remaining dimension-specific variation is minimal, as can be seen from loadings of DAD and INI items on their respective dimensions (Figure 3.14). Had we not used the alternative parameterization in the estimation of the second-order Rasch model, the variance for DAD and INI dimensions would have been estimated approximately at $0.08^2 = 0.0064$ and $0.07^2 = 0.0049$ respectively, which would most likely be difficult to estimate accurately.

The Rasch testlet model, shown in Figure 3.15, was found to have the worst fit among the four models. Judging by BIC alone, unidimensional Rasch—the simplest model, which assumes that all items measure the same construct—is preferred to the Rasch testlet model when applied on this demonstration data. Findings related to the loadings of DAD, COS, and INI items on their respective dimensions are very similar to that of the second-order Rasch model.

Recall that Rasch testlet model assumes that items from all the three dimensions load equally on the overall dimension. The extended bifactor Rasch model (Figure 3.16) relaxes this assumption. I found that items from the INI dimensions load the highest on

the overall dimension. This finding is very similar to the observation from the second-order Rasch model, where the INI dimension loads the highest on the overall dimension.

3.4 Discussion

The three-dimensional Rasch model was found to be the best fitting model when judged by both AIC and BIC. This model is the most flexible in terms of the structure (i.e., both within- and between-item multidimensionality are allowed) and does not assume any hierarchy in constructs. This model is ideal when the main purpose is to estimate respondents' location in these constructs by exploiting the covariance between constructs.

One the other hand, the second-order Rasch model may be preferred if the main focus of the analysis is to estimate respondents' location on the overall and domain-specific dimensions simultaneously while assuming that the two are linearly related. However, there are several limitation to this model: the model requires that only between-item multidimensionality be included, and also there is an assumption that the overall dimension "causes" domain-specific dimensions.

Bifactor models, in turn, are useful when domain-specific dimensions are nuisances that nevertheless need to be accounted for. They can lead to difficult interpretations of the domain-specific dimensions if the conditioning on the main dimension is not borne in mind. The extended bifactor Rasch model relaxes the assumption of the Rasch testlet model that all items from all domains load equally and provides a more flexible approach.

Limitations of the study. One limitation of the current chapter is that I didn't address how reliability is affected by accounting for the dimensionality differently (and failing to account for the structure of the dimensionality properly). Reliability is one important consideration when opting for multidimensional models (instead of unidimensional models). For instance, the reliability of EAP scores of domain-specific dimensions tend to be higher when the multidimensional Rasch model is used, instead of modeling each dimension separately (Briggs & Wilson, 2003).

Secondly, I didn't discuss extensions of these models to (1) the nested data; and (2) regression on manifest variables (i.e., latent regression). Extensions to such cases should be straightforward and can be implemented readily using programs such as Mplus (Muthen & Muthen, 2011) and Latent Gold (Vermunt & Magidson, 2013).

All the models I discuss in this chapter assume that latent variables are continuous. These latent variables can also be assumed categorical or ordinal (e.g., bifactor model with ordinal latent variables). See Cho, Cohen, & Kim (2014) for one possible extension to such a model.

References

Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

Ackerman, T. (1996). Developments in Multidimensional Item Response Theory. *Applied Psychological Measurement*, *20*(4), 309–310.

Adams R. J.,& Khoo, S-T. (1996). *Quest* [computer program]. Melbourne, Australia: ACER Press.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547-573.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest 4*. Melbourne: Australian Council for Educational Research.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley & Sons.

Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scandinavian Journal of Statistics*, *20*(1), 63–71.

Agresti, A. (2012). *Analysis of Ordinal Categorical Data* (2nd edition). Hoboken, NJ: Wiley & Sons.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken, NJ: Wiley & Sons.

Agresti, A., & Lang, J. B. (1993). A Proportional Odds Model with Subject-Specific Effects for Repeated Ordered Categorical Responses. *Biometrika*, *80*(3), 527–534.

Aitchison, J., & Silvey, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, *44*(1/2), 131–140.

Akaike, H.(1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61–98.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

Allison, P. D. (2009). *Fixed Effects Regression Models*. SAGE Publications.

Allison, P. D., & Christakis, N. A. (1994). Logit models for sets of ranked items. *Sociological Methodology*, *24*, 199–228.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology*, *26*(6), 1323–1333.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*(2), 283–301.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*(1), 3–16.

Anderson, C. J., & Yu, H.-T. (2007). Log-Multiplicative Association Models as Item Response Models. *Psychometrika*, *72*(1), 5–23.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch Model that dispels any "Threshold Disorder Controversy." *Educational Psychological Measurement, 73*, 78–124.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics*. Princeton University Press.

Arah, O. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology, 5*, 5.

Armstrong, B. G., & Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, *129*(1), 191–204.

Arnau, R. C., Meagher, M. W., Norris, M. P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *20*(2), 112–119.

Baetschmann, G. (2012). Identification and estimation of thresholds in the fixed effects ordered logit model. *Economics Letters*, *115*(3), 416–418.

Bartolucci, F., Colombi, R. and Forcina, A. (2007). An Extended Class of Marginal Link Functions for Modelling Contingency Tables by Equality and Inequality Constraints. *Statistica Sinica, 17,* 691–711.

Bennett, R. E. (1993). Toward Intelligent Assessment: An Integration Of Constructed Response Testing, Artificial Intelligence, And Model-Based Measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Bereiter, C. (1963). Some persisting problems in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, pp. 3–20.

Bergsma, W., & Croon, M. A. (2005). Analyzing categorical data by marginal models. In A. Van der Ark, M. A. Croon, & K. Sijtsma (Eds), *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*. Mahwah, NJ: Erlbaum.

Bishop, Y., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Psychology*, *53*, 605–634.

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models*. Hoboken, NJ: Wiley.

Boorsboom, D., Mellenbergh, G. J., & Heerden, J. V. (2001). *Philosophy of science and psychometrics: Reflections on the theoretical status of the latent variable* (No. 20011). *Methodological Rep*. Amsterdam: Univ. Amsterdam Dept.Psychol.

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for Experimenters*, 2nd edition. New York: John Wiley & Sons.

Bradley, R. A. and Terry, M. E. (1952) Rank analysis of incomplete block designs: I, The method of paired comparisons. *Biometrika, 39*, 324–345.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Breinegaard, N., Rabe-Hesketh. S. and Skrondal, A. (2015). The transition model test for serial dependence in mixed-effects models for binary data. *Statistical Methods in Medical Research*, in press.

Breslow, N. E. & Day, N (1980). *Statistical Methods in Cancer Research. Vol I —The Analysis of Case-Control Studies.* Lyon: IARC.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment*, *25*(1), 136–145.

Bryk, A. S., & Weisberg, H. I. (1977). Use of Nonequivalent Control-Group Design When Subjects Are Growing. *Psychological Bulletin*, *84*(5), 950–962.

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Boca Raton: CRC Press.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248.

Carroll, J. B. (1957). Biquartimin Criterion for Rotation to Oblique Simple Structure in Factor Analysis. *Science, 126*(3283), 1114–1115.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine, 8*, 1075–1093.

Carroll, R. J., Gallo, P. & Gleser, L. J. (1985). Comparison of least squares and errors in variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association, 80*, 929–932.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. New York: Chapman Hall.

Cattell, R. B. (1982). The clinical use of difference scores: Some psychometric problems. *Multivariate Experimental Clinical Research, 6,* 87–98.

Chamberlain, G. (1984). Panel data. In Z. Griliches & D. Intriligator (Eds), *Handbook of econometrics* (Vol. 2). Elsevier.

Chen, Z., & Kuo, L. (2001). A Note on the Estimation of the Multinomial Logit Model With Random Effects. *The American Statistician*, *55*(2), 89–95.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2014). A mixture group bi-factor model for binary responses. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 375-395.

Choi, T., & Cole, S. R. (2004). A family of ordered logistic regression models fit by data expansion. International Journal of Epidemiology, 33(6), 1413. Choi, T., & Cole, S. R. (2004). A family of ordered logistic regression models fit by data expansion. *International Journal of Epidemiology, 33*(6), 1413.

Clayton, D.G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika, 61*, 525-531.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum.

Cliff, N., & Keats, J. A. (2003). *Ordinal Measurement in the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables* (Vol. 4). Thousand Oaks, CA: Sage Publications.

Cole, S. R., & Ananth, C. V. (2001). Regression models for unconstrained, partially or fully constrained continuation odds ratios. *International Journal of Epidemiology*, *30*(6), 1379–1382.

Cole, S. R., Allison, P. D., & Ananth, C. V. (2004). Estimation of cumulative odds ratios. *Annals of Epidemiology*, *14*(3), 172–178.

Collins, L. M. (1996). Is reliability obsolete? A commentary on "Are simple gain scores obsolete?" *Applied Psychological Measurement*, *20*(3), 289–292.

Cornfield, J. (1978) Randomization by group: A formal analysis. *American Journal of Epidemiology, 108*(2), 100–102.

Cox, D. R. (1958). *Planning of experiments*. New York: John Wiley & Sons.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68–80.

Crowder. M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*. London: Chapman & Hall.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, L. A., Anderson, K., & Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine*, *9*(12), 1501–1515.

Das, M., & van Soest, A. (1999). A panel data model for subjective information on household income growth. *Journal of Economic Behavior and Organization, 40*, 409-426.

de Finetti, B. (1964). Foresight: Its Logical Laws, Its Subjective Sources. In H. E. Kyburg, & H. E. Smokier (Eds.), *Studies in Subjective Probability*. New York: Wiley.

de Finetti, B.(1972). *Probability Induction and Statistics*. New York: Wiley.

de la Torre, J., & Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model Approach. *Applied Psychological Measurement*, *33*(8), 620–639.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412–423.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds), *Handbook of Statistics* (Vol. 26). Amsterdam: Elsevier Science B.V.

Diehr, P., Martin, D. C., Koepsell, T., & Cheadle, A. (1995a). Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Statistics in Medicine, 14*, 1491–1504.

Donner, A., Taljaard, M., & Klar, N. (2007). The merits of breaking the matches: a cautionary tale. *Statistics in Medicine, 26*, 2036–2051.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495–515.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, *23*(4), 335–353.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press.

Feinberg, S. E., and W. M. Mason. (1978). Identification and estimation of age—period—cohort models in the analysis of discrete archival data. In K. F. Schuessler (Ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.

Fisher, R. (1937). *The design of experiments*. London: MacMillan.

Fisher, R. (1951). *The design of experiments* (6 ed.). Edinburgh: Oliver & Boyd.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test Theory for A New Generation of Tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Freedman, D. A. (2010). *Statistical Models and Causal Inference*. Cambridge University Press.

Freedman, D., Pisani, R., Purves R., & Adhikari, A. (1991), *Statistics*, 2nd edition. New York: Norton.

Fullerton, A. S., & Xu, J. (2012). The proportional odds with partial proportionality constraints model for ordinal response variables. *Social Science Research*, *41*(1), 182–198.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute 15*, 246-263.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge University Press.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Gibbons, R.,D., Bock, R., Hedeker, D., Weiss, D., J., Segawa, E., Bhaumik, D., K., Kupfer, D., Frank, E., Grochocinski, V., J., Stover, A. (2007). Full-Information Item Bifactor Analysis of Graded Response Data. *Applied Psychological Measurement, 31*(1), 4–19.

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology. *European Journal of Social Psychology, 44*(7), 673-682.

Goodman, L. A. (1979). Simple models for the analysis of association in cross classifications having ordered categories. Journal of American Statistical Association,74, 537–552.

Goodman, L. A. (1983). The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics*, *39*(1), 149–160.

Goodman, L. A. (1985). *Analyzing qualitative/categorical data*. (J. Magidson). Lanham, MD: University Press of America.

Greenland S., & Robins J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations 6,* 4.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, *13*(16), 1665–1677.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407–434.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

Harman, H. H. (1967). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press.

Harris, C. W. (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Hayes, R. J., & Moulton, L. H. (2009). *Cluster randomised trials*. Boca Raton: Taylor & Francis.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika, 88,* 973–985.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, *30*(1), 239–267.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. New York: John Wiley & Sons.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association, 81*, 945–970.

Holland, P. W., & Rubin, D. B. (1982). On Lord's Paradox. *ETS Research Report Series*, *1982*(2), i–41.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Holzinger, K. J., & Swineford, F. (1939). *A Study in Factor Analysis: The Stability of a Bi-factor Solution* (Vol. 48). Chicago: University of Chicago.

Hood, S. B. (2008). *Latent variable realism in psychometrics*. (Doctoral dissertation). Indiana University.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441.

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, *1*(1), 27–35.

Imai, K. (2008). Variance Identification and Efficiency Analysis in Randomized Experiments under the Matched-Pair Design. *Statistics in Medicine, 27*(24), 4857–4873.

Imai, K., King, G., & Nall, C. (2009). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science, 24*(1), 29–53.

Imai, K., King, G., & Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A 171*, 481–502

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika*, *38*(4), 593–604.

Jensen, A. R. (1998). *The g factor: the science of mental ability*. Westport, CT: Praeger.

Jeon, M. & Rabe-Hesketh, S. (2015). An autoregressive growth model for longitudinal item analysis. *Psychometrika*, in press.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling Differential Item Functioning Using a Generalization of the Multiple-Group Bifactor Model. *Journal of Educational and Behavioral Statistics*, *38*(1), 32–60.

Jobson, J. D. (1991). *Applied Multivariate Data Analysis, Vol. I: Regression and Experimental Design*. New York: Springer-Verlag.

Kaiser, H. F. (1960). Varimax solution for primary mental abilities. *Psychometrika*, *25*(2), 153–158.

Kamata, A., & Cheong, Y. F. (2007). Multilevel Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differentiable mental abilities*. Stanford, CA: Stanford University Press.

Kelley, T. L. (1935). *Essential traits of mental life*. Cambridge: Harvard university press.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, *82*(3), 345–362.

Kenny, D. A. (2011). Change we cannot believe in. In R. M. Arkin (Ed.), *Most underappreciated*. New York: Oxford.

Kenny, D. A., & Cohen, S. H. (1980). A reexamination of selection and growth processes in the nonequivalent control group design. In K. F. Scheussler (Ed.), *Sociological methodology*. San Francisco, CA: Jossey-Bass.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Lindley, D.V., & Novick, M. R. (1981). The Role of Exchangeability in Statistical Inference, *The Annals of Statistics, 9*(1), 45–48.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre-and posttesting periods. *Review of Educational Research*, *47*(1), 121–150.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

London, L., & Wright, D. (2012). Analyzing Change between Two or More Groups. In B. P. Laursen, T. D. Little & N. A.Card (Eds), *Handbook of Developmental Research Methods*. New York : Guilford Press.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 304–305.

Lord, F. M. 1973. Lord's paradox. In S. B. Anderson, S. Ball, & R.T. Murphy (Eds), *Encyclopedia of Educational Evaluation*. San Francisco: Jossey-Bass.

Luce, R. D. (1959). *Individual Choice Behavior.* New York, NY: John Wiley & Sons.

Luecken, L. J. & Tanaka, R. (2012). Health Psychology. In J.A. Schinka & W. Velicer (Eds), *Handbook of Psychology: Volume 2: Research Methods*. New Jersey: Wiley Publications.

Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26-43.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3*, 309–327.

Martin, D.C., Diehr, P., Perrin, E.B. & Koepsell, T.D. (1993). The effect of matching on the power of randomized community intervention studies*, Statistics in Medicine, 12*, 329–338.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika, 65*, 413-418.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*(2), 109–142.

McCullagh, P., & Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

McCullagh, R. (2005). Exchangeability and Regression Models. In A. Davison, Y. Dodge, & N. Wermuth, (Eds.), *Celebrating Statistics: Papers in Honour of David Cox on his 80th Birthday*. Oxford: Oxford University Press.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed.), *Frontiers of Econometrics*, New York, NY: Academic Press.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *J. Mathematical Sociology*, *4*(1), 103–120.

McKinley, R. L., & Reckase, M. D. (1983). *An Extension of the Two-parameter Logistic Model to the Multidimensional Latent Space* (No. Research Report ONR 83-2). Iowa City, IA: The American College Testing Program.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In D. Thissen & H. Wainer (Eds), *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*(1), 91–100.

Mellenbergh, G. J., & Van den Brink, W.P. (1998). The measurement of individual change. *Psychological Methods, 3,* 470–485.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81–91.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25*(3), 271–284.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2001a). Optimal experimental designs for multilevel logistic models. *The Statistician, 50*, 17–30.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2001b). Optimal experimental designs for multilevel models with covariates. *Communications in Statistics, Theory and Methods, 30*, 2683–2697.

Moerbeek, M, van Breukelen, G., & Berger, M. P. F. (2008). Optimal designs for multilevel studies. In J. de Leeuw & E. Meijer (Eds). *Handbook of Multilevel Analysis*. New York: Springer.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Moustaki, I. (2007). Factor analysis and latent structure of categorical and metric data. In R. Cudeck & R. MacCallum, *Factor Analysis at 100: Historical Developments and Future Directions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor Models for Ordinal Variables With Covariate Effects on the Manifest and Latent Variables: A Comparison of LISREL and IRT Approaches. *Structural Equation Modeling: a Multidisciplinary Journal*, *11*(4), 487–513.

Muggeo, V., & Aiello, F. (2011). Modeling Ordinal Item Responses via Binary GLMMs and Alternative Link Functions: An Application to Measurement of a Perceived Service Quality. In M. Attanasio & V. Capursi (Eds), *Statistical Methods for the Evaluation of University Systems*. Berlin: Springer-Verlag.

Mukherjee, B., Ahn, J., Liu, I., Rathouz, P. J., & Sánchez, B. N. (2008). Fitting stratified proportional odds models by amalgamating conditional likelihoods. *Statistics in Medicine*, *27*(24), 4950–4971.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models.* New York: Routledge.

Nerlove, M., & Press, S. J. (1973). *Univariate and Multivariate Log-linear and Logistic Models*. Santa Monica, California: Rand Corporation.

Newell, S., Sanson-Fisher, R. W., Girgis, A., & Ackland, S. (1999). The physical and psycho-social experiences of patients attending an outpatient medical oncology department: a cross-sectional study. *European Journal of Cancer Care*, *8*(2), 73–82.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.

Novick, M.R. (1983). The centrality of Lord's paradox and exchangeability for all statistical inference. *Principals of modern psychological measurement*. Hillsdale, NJ: Earlbaum.

O'Connell, A. A. (2006). *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks, CA: Sage Publications, Inc.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Paek, I., Yon, H., Wilson, M., & Kang, T. (2009). Random parameter structure and the testlet model: Extension of the Rasch Testlet model. *Journal of Applied Measurement*, *10*, 394–407.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd edition). New York: Cambridge University Press.

Pearl, J. (2014). Lord's Paradox Revisited–(Oh Lord! Kumbaya!). *Technical Report*.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(6), 559–572.

Pearson, K. (1930), *The Life, Letters and Labors of Francis Galton*. Cambridge University Press.

Pfanzagl, J. (1993). On the consistency of conditional maximum likelihood estimators. *Annals of the Institute of Statistical Mathematics*, *45*(4), 703–719.

Porter, A. C., & Raudenbush, S.W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34*(4), 383–392.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata* (3rd ed., Vol. 1). College Station, TX: Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*, 301–323.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: Mesa Press.

Raudenbush, S. W., & Liu, X. (2000). Statistical Power and Optimal Design for Multisite Randomized Trials. *Psychological Methods, 5*(2), 199–213.

Raudenbush, S.W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model for self-reported criminal behavior. *Sociological Methodology, 33*(1), 169–211.

Rausch, J. R., Maxwell, S.E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology, 32*(3), 467–486.

Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement*, *9*(4), 401–412.

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

Reichardt, C. S. (1979), The Statistical Analysis of Data from the Nonequivalent Control Group Design. In T. D. Cook & D. T. Campbell (Eds.), *The Design and Analysis Issues in Field Settings.* Chicago: Rand-McNally.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The Role of the Bifactor Model in Resolving Dimensionality Issues in Health Outcomes Measures. *Quality of Life Research*, *16*, 19–31.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566.

Rijmen, F. (2009). *Three Multidimensional Models for Testlet-Based Tests: Formal Relations and an Empirical Comparison* (No. ETS Research Report No. RR-09-37). Princeton, NJ: ETS.

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, *47*(3), 361–372.

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *4*, 59–74.

Rijmen, F., Jeon, M., Rabe-Hesketh, S. & von Davier, M. (2014). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *Journal of Educational and Behavioral Statistics. 39,* 245–256.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability the difference score in the measurement of change. *Journal of Educational Measurement*, *20*(4), 335–343.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90*, 726–748.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53*(3), 349–359.

Rotnitzky, A. & Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics, 50*, 1163–1170.

Rubin, D. (2006). *Matched sampling for causal effects*. Cambridge, England: Cambridge University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology, 66*, 688–701.

Rubin, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*, 1–26.

Rubin, D. B., Stuart, E. A., & Zanutto, E.L. (2004) A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Ryan, P. T. (2013). *Sample Size Determination and Power*. Hoboken, NJ: John Wiley.

Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* New York: Owl Books.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt.2).

Samejima, F. (1997). Graded response model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61.

Schwartz, A. R. (2012). *The Development and Psychometric Modeling of an Embedded Assessment for a Data Modeling and Statistical Reasoning Learning Progression*. (Unpublished doctoral dissertation). University of California, Berkeley.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Scott, S. L., & Ip, E. H. (2002). Empirical Bayes and item-clustering effects in a latent variable hierarchical model: a case study from the National Assessment of

Educational Progress. *Journal of the American Statistical Association*, *97*(458), 409–419.

Sen, A. & Srivastava, M. (1990). *Regression analysis, theory, methods and applications*. Springer: Berlin.

Senn, S. J. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine, 25*, 4334–4344.

Senn, S. J. (2007). *Statistical issues in drug development*, 2nd edition. Chichester, UK: John Wiley & Sons.

Sijtsma, K., & Hemker, B.T. (2000). A taxonomy for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics, 25*, 391-415.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory.* Thousand Oaks, CA: Sage Publications, Inc.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling : Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd edition. London: Sage Publications.

Snow, J. (1855). *On the Mode of Communication of Cholera* (2nd ed.). London: John Churchill.

Snow, J., Frost, W. H., & Richardson, B. W. (1965). *Snow on cholera: Being a Reprint of Two Papers*. New York: Hafner Pub. Co.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–292.

Spearman, C. C. (1927). *The abilities of man; their nature and measurement*. New York: The Macmillan Company.

Stuart, A. (1958). Equally correlated variates and the multinormal integral. *Journal of the Royal Statistical Society, Series B, 20*, 373-378.

Ten Have, T. R., & Uttal, D. H. (1994). Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Applied Statistics, 43*, 371-384.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment. *Journal of Educational Psychology, 51*(6), 309–17.

Thorndike, R. L. (1966). *The concepts of over- and underachievement*. New York: Columbia University, Teachers College, Bureau of Publications.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*, 406–427.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: The University of Chicago press.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*(1), 23–30.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds), *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Tuerlinckx, F., & Wang, W.-C. (2004). Models for polytomous data. In M. Wilson & P. De Boeck (Eds), *Explanatory Item Response Models*. New York: Springer.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55.

Tutz, G. (1991). Sequential models in ordinal regression. *Computational Statistics & Data Analysis, 11,* 275–295.

Tutz, G. & Hennevogl, W. (1996). Random Effects in Ordinal Regression Models. *Computational Statistics and Data Analysis, 22*, 537–557.

van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922.

van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2005). *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*. Psychology Press.

van der Linden, W. (1992). Fundamental Measurement and the Fundamentals of Rasch Measurement. In M. Wilson (Ed.) *Objective Measurement: Theory into Practice Vol. 2*. Norwood, NJ: Ablex Publishing Corp.

van der Linden, W. & Hambleton, R. K. (1997). *Handbook of item response theory.* New York: Springer-Verlag.

Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y., & Fontaine, J. (2008). The factor structure of the Beck Depression Inventory-II: an evaluation. *Assessment*, *15*(2), 177–187.

Verhelst, N. D. and Eggen, T. J. H. M. (1989). Psychometrische en statistische aspecten van peilingsonderzoek [*Psychometric and statistical aspects of assessment research*.] (PPON-rapport, 4) Arnhem: Cito.

Vermunt, J. K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax.* Belmont, Massachussetts: Statistical Innovations Inc.

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.

Volodin, N. A., & Adams, R. J. (1995). Identifying and estimating a D-dimensional item response model. *Eighth International Objective Measurement Workshop, University of California, Berkeley*.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*(1), 147–151.

Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences. *The American Statistician*, *58*(2), 117–123.

Wainer, H., & Brown, L. M. (2007). Three Statistical Paradoxes in the Interpretation of Group Differences: Illustrated with Medical School Admission and Licensing Data. In C. R. Rao & S. Sinharay (Eds), *Handbook of Statistics* (Vol. 26). Elsevier.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

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

Wang, W., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds), *Objective Measurement: Theory into Practice* (Vol. 4). Greenwich, CT: Ablex Publishing.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement*, *26*(1), 109–128.

Weinberg, S. L., & Goldberg, K. P. (1990). *Statistics for the behavioral sciences*. Cambridge: Cambridge University Press.

Weisberg, H. I. (1979). Statistical adjustments and uncontrolled studies. *Psychological Bulletin, 86*, 1149–1164.

Weisberg, H. I. (2010). *Bias and causation: Models and judgment for valid comparisons.* Hoboken, NJ: Wiley.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479–494.

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P. W., & Peck, K. (2009). *A multisite cluster randomized trial of the effects of CompassLearning Odyssey® Math on the math achievement of selected grade 4 students in the Mid-Atlantic region* (NCEE 2009-4068). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Willett, J. B. (1997). Measuring Change: What Individual Growth Modeling Buys You. In E. Amsel & K. A. Renninger (Eds)*, Change and Development: Issues of Theory, Method, and Application.* Mahwah, NJ: Lawrence Erlbaum Associates.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Wilson, M. (1988). Detecting and Interpreting Local Item Dependence Using a Family of Rasch Models. *Applied Psychological Measurement*, *12*(4), 353–364.

Wilson, M. (1992). The Ordered artition Model: An Extension of the Partial Credit Model. *Applied Psychological Measurement*, *16*(4), 309–325.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M., & Adams, R. J. (1993). Marginal Maximum Likelihood Estimation for the Ordered Partition Model. *Journal of Educational Statistics*, *18*(1), 69–90.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181–198.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. *MESA Press Chicago*.

Wright, D. B. (2006). Comparing groups in a before–after design: When t-test and ANCOVA produce different results. *British Journal of Educational Psychology, 76*, 663–675.

Wu, M., & Ware, J. H. (1979). On the use of repeated measurements in regression analysis with dichotomous responses. *Biometrics*, *35*(2), 513–521.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*(3), 339–360.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128.

Zigmond, A. and Snaith, R. (1983). The hospital anxiety and depression scale. *Acta Psychiatrika Scandinavica, 67*, 361–370.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*(2), 149–154.

Zwinderman AH (1991). A Generalized Rasch Model for Manifest Predictors. *Psychometrika, 56*, 589–600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds). *Handbook of modern item response theory*, New York: Springer.