PSYCHOMETRIC FOUNDATIONS FOR MODELING RELATIVE GROWTH IN A DIAGNOSTIC FRAMEWORK

by

MADELINE ALYCE SCHELLMAN

(Under the Direction of Laine Bradshaw and Matthew Madison)

ABSTRACT

Educators desire metrics to evaluate student progress over time. However, merely tracking changes in students' scores is often deemed insufficient. Stakeholders are also interested in evaluating whether individual student growth or class average growth is satisfactory or lacking compared to other students or classes. The student growth percentile (SGP; Betebenner, 2009) was devised to compare student growth to the growth of peers with similar score histories. SGPs are commonly used with assessments that provide sum or scaled scores generated from item response theory (IRT) models. However, such assessments fall short of meeting the increasing demand for reliable results that pinpoint students' specific strengths and weaknesses. This demand can be addressed through diagnostic assessments designed for use with diagnostic classification models (DCMs; Rupp et al., 2010). The recent transition of diagnostic models and methods from theoretical research to practical application makes it crucial that psychometricians provide metrics to evaluate growth within the DCM framework. Hence, this dissertation introduces the diagnostic growth percentile (DGP)—an adaptation of SGP for use with student results from DCMs—and reliability metrics for the DGP.

Specifically, to evaluate the efficacy, reliability, and validity of DGPs, I conducted a simulation study and two empirical data analyses, which illustrate the computation, interpretation, and utility of the DGPs. Results of these studies showed that DGPs have acceptable levels of reliability for various assessment conditions and are viable approaches for comparing student growth in the DCM framework. I conclude this dissertation by comparing SGPs and DGPs. In sum, DGPs overcome some issues that plague SGPs while introducing new limitations that pose issues for potential uses and interpretations. This dissertation explores one approach for using an SGP-like metric in the DCM framework and shows favorable results for the DGP metric. However, additional research is needed before DGPs can be used in practice, and, as always, anyone interested in using DGPs in practice should prepare a comprehensive validity argument to support their intended use(s) for DGPs in their specific situations.

INDEX WORDS: diagnostic classification modeling, longitudinal student classification, growth modeling, diagnostic growth percentile, reliability, polytomous attributes

PSYCHOMETRIC FOUNDATIONS FOR MODELING RELATIVE GROWTH IN A DIAGNOSTIC FRAMEWORK

by

MADELINE ALYCE SCHELLMAN

B.S., University of Georgia, 2017

M.A., University of Georgia, 2021

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2024

PSYCHOMETRIC FOUNDATIONS FOR MODELING RELATIVE GROWTH IN A DIAGNOSTIC FRAMEWORK

by

MADELINE ALYCE SCHELLMAN

Major Professor:	Laine Bradshaw Matthew Madison
Committee:	George Engelhard Ashley Harrison Shiyu Wang

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2024

DEDICATION

To my parents, Richard and Kathe Schellman, for your constant examples of unconditional love, commitment, hard work, dedication, honesty, integrity, and living life to its fullest. No day but today.

ACKNOWLEDGEMENTS

I am so fortunate to be surrounded by many dedicated, passionate, and supportive people in all aspects of my life. Without such a strong system of connections, I would not have been able to write this dissertation, so I owe thanks to many. First, I would like to thank my advisor and major professor, Laine Bradshaw. Over the years, I have learned countless things from Laine about psychometrics, of course, but also about education, teaching, mentoring, business, people, and life. I am extremely grateful for my unique graduate experience because of the many opportunities Laine gave me to apply psychometrics in the real world and support student learning every day. I truly believe that Laine's work can change education for the better. I am deeply honored to be part of her mission, and I hope to make her proud throughout my career.

I would also like to thank Matthew Madison for always believing in me more than I ever felt I deserved and for always being available for psychometric and measurement discussions. This dissertation was only possible because Matt generously shared his expertise in longitudinal diagnostic measurement with me.

I would also like to thank George Engelhard, Ashley Harrison, and Shiyu Wang for serving on my doctoral committee and providing key insights that significantly improved this work and helped me learn more about conducting psychometric research.

It is not easy to be employed full-time and write a dissertation, so I would like to express gratitude to Pearson and my amazing Navvy coworkers for supporting me in completing my PhD.

Dissertation writing can be a very time-consuming and stressful process, so it is extremely important that even during such a busy and important time in a doctoral student's life, they find ways to enjoy life and find fulfillment outside of school. Anyone who knows me even a little bit knows that I spend my free time training at my gym, Megalodon, or watching/listening to musical theater. So, I would like to thank my gym family, especially Grant, Jess, Vinnie, Casey, Bunny, Johnny, Don, Bart, Tracy, Kati, Aly, and Scott. You all believe in me so much, and your support has helped me learn that I can achieve greatness. I have applied your teachings (dig!) in the gym as well as in my professional life, so thank you all for pouring your time and energy into me. Also, thank you, Idina.

Finally, I would like to thank my incredible family: my parents, Kathe and Richard, and Kingsley, George, Harper, Sloane, Mel, Ashton, Mike, Haven, Jody, and Trinity. Your unending love and support in everything I do means the world to me. And an extra huge thank you to Trinity for all of the things she did to make my life easier and happier while I was writing this dissertation, including helping me around the house and with grocery shopping and driving so I could work in the car on the way to our adventures. Another extra huge thank you to my mom for reading this ENTIRE dissertation and letting me know where I was missing commas.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Diagnostic classification models (DCMs; Rupp et al., 2010) are constrained and confirmatory latent class models that use categorical item response data (e.g., correct/incorrect) to estimate students' statuses, which are usually binary (with labels such as *master* versus *non-master*) with respect to one or more categorical latent variables called *attributes*. This approach contrasts with confirmatory factor analysis (CFA), which uses continuous observed data to model continuous latent variables (Bramlett, 2018), and item response theory (IRT), which uses categorical observed data to model continuous latent variables (Hambleton et al., 1991).

DCMs have mostly been used in research with simulation studies, and most empirical data analyses have retrofitted DCMs to data from assessments that were designed to be used with traditional psychometric models. Empirical data from diagnostic assessments is almost nonexistent because few assessments have been developed from the ground up in the DCM framework. Additionally, the student response data from operational diagnostic assessments and diagnostic assessment systems that have been developed (e.g., Bradshaw et al., 2014; Bradshaw et al., 2017-2021; Dynamic Learning Maps Consortium, 2022) is not publicly available to DCM researchers, so DCM studies are often not able to fully investigate the properties and practicability of DCMs in practice. However, educators and practitioners have sought results that provide clear indications of students' specific strengths and areas for improvement to support

personalized student learning and growth. DCMs and longitudinal DCMs are well-suited to provide such results, but because DCMs are still relatively new models and have few true applications, many educators have not yet been able to take advantage of DCMs' ability to support student learning. Therefore, it is important that psychometric researchers fully and appropriately investigate the methods and statistical properties of DCMs for use in practice, evaluate how these methods and properties compare with those of traditional psychometric models, and apply these methods and properties to link research and practice.

My dissertation investigates methodology that is logically of interest, as DCMs are well-suited for educational needs but have not yet been widely used in practice. I begin this chapter by generally describing what DCMs are. Then, I discuss why stakeholders would want to utilize DCMs in practice, as well as describe some challenges with using DCMs in practice. I then describe the practical scenario that needs methodological investigation and that motivates the methods for this dissertation. Finally, I provide an overview of this dissertation to address the methodological questions related to the application of DCMs.

What are Diagnostic Classification Models?

Latent variables are things that we cannot see or tangibly measure, but we believe they impact how students respond to assessment items. They can be cognitive, such as skills or knowledge, or they can be affective, such as attitudes or beliefs. For example, we cannot physically measure a student's level of knowledge about algebra in the same way we can use a yardstick to measure the width of a driveway because we cannot literally see students' knowledge about algebra, and, more importantly, we do not have a universal

yardstick for algebra knowledge. The purpose of educational assessment is to serve as a sort of yardstick or, more generally, a tool for psychological measurement to estimate students' levels of whatever latent variables are of interest. Our assumptions about those latent variables and the interpretations we want to make about them dictate the type of assessment and psychometric model we need to use. If we believe that a particular latent variable has infinitely many levels (i.e., it is a continuous variable), then we need an assessment that uses a continuous variable model, like an IRT model. Generally, we refer to latent variables in the IRT framework as *abilities*. If we believe that a particular latent variable has two levels (i.e., it is a discrete or categorical variable), then we need an assessment that uses categorical variables, like a DCM. Generally, we refer to latent variables in the DCM framework as *attributes*.

Because DCMs assume that the latent variables being measured are categorical, the goal with DCMs is to estimate students' levels or *proficiency statuses* with respect to one or more attributes. As mentioned above, attributes are typically binary (*dichotomous*) and use labels such as *master* versus *non-master*. This dissertation uses only dichotomous attributes with the labels *proficient* and *non-proficient* to refer to students' proficiency statuses for the measured attributes. DCMs estimate for each student and each attribute the *posterior probability* that the student is proficient for the attribute. If the posterior probability is greater than .5, then the student is most likely proficient for the attribute. After estimating a student's proficiency statuses for all measured attributes, the DCM provides an *attribute profile*, which includes all of the student's proficiency statuses. The attribute profile provides a snapshot of each student's strengths and areas for improvement within the set of measured attributes.

To illustrate the utility of DCMs, suppose Student X and Student Y both received scores of 40% on their most recent math test, which was about using addition, subtraction, multiplication, and division. Because these students have the same total score, one might conclude that both students need the same kind of remediation. However, total scores (and overall ability scaled scores) do not give the full picture of students' specific needs. Suppose, instead, that the math teacher administered a diagnostic math assessment that uses a DCM to estimate students' proficiency statuses for four attributes: addition, subtraction, multiplication, and division. Suppose the diagnostic assessment showed that Student X is proficient for addition, but not proficient for subtraction, multiplication, and division, and Student Y is proficient for subtraction, but not proficient for addition, multiplication, and division. Though Student X and Student Y have the same total score on their classroom test, they have different attribute profiles from the diagnostic assessment, so they need different learning support. Teachers can save time and effort by using their students' attribute profiles to tailor their differentiated instruction plans so students receive exactly the support that they need—no more, no less (i.e., students do not need remediation for attributes for which they are already proficient, and they do need remediation for attributes for which they are not already proficient). In sum, DCM attribute profiles provide efficient, statistically-driven, and actionable information to inform personalized learning plans.

I discuss further details about DCMs and provide specific DCM parameterizations in Chapters 2 and 3 of this dissertation.

Why Stakeholders Would Want to Use Diagnostic Classification Models in Practice

The previous section introduced DCMs conceptually and highlighted a few components of the DCM framework. This section dives deeper into some of the key reasons that DCMs are attractive for use in educational assessment. The four key benefits of using DCM-based assessments in education that I discuss in this section include (1) DCMs provide actionable multidimensional diagnoses, (2) DCMs allow for the use of short assessments, (3) standard setting is not necessary with DCMs, and (4) DCM results add to the cognitive theory literature related to the attributes.

DCMs Provide Actionable Multidimensional Diagnoses

Because DCMs are latent class models, they allow for multidimensionality between and within items and typically measure latent variables that are more fine-grained than the latent variables measured in traditional assessments that utilize IRT (Rupp & Templin, 2008; Rupp et al., 2010). For example, an IRT-based assessment might measure students' levels of geometry ability, but a DCM-based assessment might diagnose students' proficiency statuses for latent variables within geometry, such as special triangles, circles, and transformations. Students' diagnoses for these fine-grained attributes provide specific information that can help to tailor instruction to fit their specific needs.

The latent variable in an IRT-based assessment could also be fine-grained like a DCM attribute; however, even with a fine-grained latent variable, IRT-based assessments would still require many items to have enough data to make confident, precise estimates of student ability along a continuum (more details below). Additionally, based on my experience with item and assessment development, as the latent variable grain size gets

smaller, the difficulty of developing a large pool of items increases, so it would be difficult to develop IRT-based assessments for some fine-grained latent variables. Alternatively, DCMs are well-suited to measure fine-grained latent variables because they do not require as many items as do IRT-based assessments (see the next section for more details). The choice of the attribute grain size and modeling framework depends on the desired interpretations of the assessment results and should be central to the assessment development process.

Although IRT largely requires and assumes unidimensionality, multidimensional IRT (MIRT) models have been developed to accommodate multidimensionality between and within items as well. However, in most applications with MIRT models, items tend to be simple structure (Rupp et al., 2010) even though MIRT models do not restrict items to be simple structure (Reckase, 1997). Even in situations where MIRT items are complex, the confirmatory MIRT model typically excludes interaction parameters and includes only the main effect parameters for each measured latent variable (Templin & Bradshaw, 2013). DCMs, on the other hand, lend themselves more readily than MIRT models to complex structure items because they can, and typically do, include all higher-order interaction terms. Although some DCMs add restrictions on item parameters (e.g., constrain main effects to be zero; constrain the maximum interaction term to be zero), general DCMs allow all possible item parameters to be estimated freely. The studies I conducted for this dissertation use only general DCMs that do not place any extra restrictions on item parameters.

DCMs Allow for the Use of Short Assessments

In measurement, the amount of data available for fitting psychometric models comes from the number of items per latent variable and the number of students who respond to the items. *Item data* refers to the number of items included in an assessment. IRT models require more item data than DCMs because IRT models aim to precisely estimate students' locations on a continuous variable with as little error as possible. DCM-based assessments require fewer items because attributes are categorical. Thus, DCMs only need enough item data to confidently place students into one of two groups (typically two—there could be more than two groups, but likely not many more than five groups) rather than an infinite number of possible outcomes, as with a continuous variable.

Not only is student parameter estimation more efficient with DCMs than with IRT, but DCM-based assessments have been shown to be as reliable, or more reliable, than their much longer IRT counterparts. Templin and Bradshaw (2013) fit several DCMs and several IRT models to empirical data from an end-of-grade reading assessment that had 73 items. They showed that a DCM with a single binary attribute achieved the same level of reliability of student estimates as the two-parameter logistic (2PL) IRT model with only 14 (19%) of the original 73 items. Having fewer items is desirable for classrooms where instructional time is limited and reliable formative diagnostic assessments are needed to support day-to-day decisions. The more items an assessment has, the more time students spend testing, which has become a greater concern in today's classrooms as

> Today's state tests [which are typically IRT-based assessments] often take several hours and are spread across several days. In Bibb County, Georgia, for example, a state or national exam is given to elementary, middle or high-school students in 70 of 180 school days. In Maryland, state testing occurs across 55 days of the school year; in Texas, 51 days (71 if you count field testing); in Michigan it's 50 days. Except for Bibb County, none of these numbers include other tests students may take such as the National Assessment of Educational Progress (NAEP), SAT or ACT college admission tests, Advance Placement (AP) exams or other commercially available standardized tests the local school districts may choose to administer. (Madaus et al., 2009, p. 2).

The point is that students spend a lot of time taking assessments. We must impose constraints on the number of items we can realistically expect students to answer during any given assessment because time is limited, and students can get fatigued with long assessments. Testing time and fatigue are considerations that could lead educators to prefer shorter assessments if the shorter assessments can provide the information needed—the intended outcomes and uses of an assessment drive the decision about what assessment type to use.

We can compare characteristics of DCM-based and IRT-based assessments and models in research to learn how these models behave, but it would be practically inappropriate to take a situation that uses an IRT-based assessment (e.g., college entrance exams) and switch it with a shorter, DCM-based assessment without considering the purpose of the assessment. Because IRT- and DCM-based assessments provide different types of results, their results cannot be interpreted the same way or used for the same

purposes. For example, an IRT-based college entrance exam might be used to estimate students' math ability in a way that allows for rank-ordering students to determine which students have greater or lesser math ability. A shorter DCM-based assessment with the same (or greater) reliability would be inappropriate for this context because its results would not allow for the same types of student comparisons, as DCMs do not allow for rank-ordering students outside of their proficiency statuses and profiles.

Standard Setting is not Necessary with DCMs

In the DCM framework, if the marginal probability that a student is proficient for an attribute is greater than .5, then the student is diagnosed as proficient for the attribute. That DCM classifications are driven by statistics sets DCMs apart from MIRT models and reduces error in student classifications "by matching the statistical model with the purpose of the test" (Rupp et al., 2010, p. 47). In general, DCMs provide direct classifications for any setting by providing different weights for each assessment item to yield optimal student classifications based on their response patterns (Harrison et al., 2017). This characteristic of DCMs would be particularly useful for applications in which setting cut scores on a scale to achieve classifications could be considered somewhat subjective, such as the use of standard setting to fulfill the federal requirement for end-of-year student achievement level classifications.

The federal government requires that summative assessment results classify students into categories (Every Student Succeeds Act, 2015). For example, the Georgia Board of Education reports students as belonging to the Beginning, Developing, Proficient, or Distinguished group for each subject area (Georgia Department of Education, n.d.). DCM results directly classify students, whereas IRT results must be

converted from scaled scores to classifications. DCM classifications are statistically driven because they come from the likelihood that each student is proficient for each attribute. Therefore, DCM classifications are more objective because they do not require human decisions outside of the assessment design and construction process about how to classify students based on a scaled score.

Classification based on IRT scaled scores, however, requires a procedure called *standard setting* in which people who are experts on the measured latent trait make decisions about how to divide the scale to classify students into categories that satisfy the federal requirement for student classification. Standard setting is a costly endeavor in terms of time and human resources and is somewhat subjective. However, a common criticism of DCM's ability to do away with standard setting is that there is less oversight of the model results. Practitioners must weigh the costs and benefits of employing any standard setting procedure.

DCM Results Add to the Cognitive Theory Literature Related to the Attributes

DCMs empirically evaluate latent trait structures by providing (1) an estimate (called a *base rate*) of the prevalence of each attribute in the target population, (2) estimates of the correlations between each pair of attributes, and (3) information about the relationships between the items and the attributes via item parameters (Bradshaw, 2011; Bradshaw, 2016; Rupp et al., 2010). For example, very high attribute correlations might indicate that a pair of attributes are not distinct and should be merged into one attribute (e.g., Bradshaw et al., 2014). The interpretations that can be made from (1)-(3), when used to add to or adjust the items or the cognitive models that support the assessment (Bradshaw, 2016), can support student learning by adding to the knowledge

base related to a learning outcome which then improves curriculum, instructional planning, and decision making (Jurich & Bradshaw, 2014).

Challenges of Using Diagnostic Classification Models in Practice

When selecting any psychometric framework to use for a specific application, it is important to not only consider the benefits of using the framework but also its limitations. Although DCMs offer some valuable characteristics, they also have some challenges and limitations for their use in practice. The three key challenges of using DCM-based assessments in education that I discuss in this section include (1) the loss of precision in exchange for a smaller grain size, (2) stakeholders' unfamiliarity with DCMs, and (3) the limited estimation software for DCMs.

Loss of Precision in Exchange for a Smaller Grain Size

One of the benefits of DCMs can also be seen as a downside: Relative to an IRT context, DCMs trade precision in estimation for latent variables that are smaller in grain size (Rupp et al., 2010). From a theoretical perspective, student classifications from DCMs are less precise than student ability estimates from IRT because of the nature of the latent variables being distinctive and ordered but not having equal intervals or an absolute zero. "Measurement precision is typically defined by the presence or absence of the following four characteristics, which are ordered in terms of precision: (1) distinctiveness, (2) magnitude, (3) equal intervals, and (4) absolute zero," (Azen & Walker, 2011, p. 2). Because IRT's latent ability variables are represented on a scale without a meaningful zero-point, IRT uses an interval level of measurement, which is more precise than the ordered level of measurement that DCMs use. Note that IRT could estimate a student's ability to be zero because zero is in the middle of the ability scale,

but it is not a true zero because it does not mean that students have no ability. Categorical latent variables (as in DCM-based assessments) are less precise than continuous variables (as in IRT-based assessments) because they are coarser outcomes: two proficiency groups for a given attribute rather than infinite ability levels.

Another precision-related concern in the DCM framework is that all students with a posterior probability greater than .5 are diagnosed as proficient for the attribute, but "no further distinction is made among [students'] ability" (Bradshaw, 2016, p. 317). DCM's categorical attributes do not allow for student comparisons within groups. It is assumed that all students who are classified as proficient for a given attribute have an identical level of the attribute(s) and all students who are classified as non-proficient for a given attribute have an identical level of the attribute(s), even though it is likely that the students within the proficiency groups have different underlying abilities. Therefore, a student with a posterior probability of .39 has the same estimated proficiency status as a student with a posterior probability of .06, and the teacher's instructional decision will likely be the same for both students if the teacher plans to use students' proficiency statuses to group them for differentiated instruction, which is often the type of educational decision that DCMs are intended to inform at the classroom level (Rupp et al., 2010). The choice to use a DCM means that the desired interpretation does not require comparing students within proficiency groups.

Stakeholders' Unfamiliarity with DCMs

Educators and students need diagnostic results from classroom assessments. However, DCM use in practice is in its infancy. Few assessments have been developed from the ground up in the DCM framework (e.g., Bradshaw et al., 2014; Bradshaw et al.,

2017-2021), so students and educators are not familiar with DCM-based assessments. Diagnostic student reports themselves are unfamiliar and could be prone to misinterpretation (e.g., Jang, 2005). Successful use of DCMs in practice will require training for users at all levels so they can make appropriate use and interpretations of student reports (Bradshaw, 2016). Without such training, teachers and other stakeholders cannot take advantage of actionable, personalized, and timely information provided by the attribute profiles that result from DCMs.

Such training would be beneficial outside of the educational context, too, where DCMs are being implemented in other disciplines, such as listening comprehension (Dong et al., 2021), gambling disorder (Templin & Henson, 2006), and autism (Harrison et al., 2017). In the context of psychological disorders, a DCM-based assessment would provide clinical psychologists (who have training for interpreting DCM results) with statistically-driven and reliable classifications to support their expert opinions about course of treatment for clients.

Limited Estimation Software

Compared with the amount of software available to estimate IRT models, the software available to estimate DCMs is limited. Mplus (Muthén & Muthén, 1998-2017) and R (R Core Team, 2023) packages such as mirt (Chalmers, 2012), cdm (George et al., 2016), and gdina (Ma & de la Torre, 2020) are the primary programs available to estimate DCMs.

Mplus is a powerful software that is not free, but it is well-vetted for many statistical analyses. Mplus functionality must be somewhat “rigged” to function properly for DCMs, as it was not originally designed to accommodate DCM estimation. Templin

and Hoffman (2013) explain in detail how to use existing Mplus functionalities to make it possible to estimate DCMs. The Mplus script for estimating DCMs is complex and prone to errors (Rupp et al., 2010; Templin & Hoffman, 2013). The code required to estimate DCMs in Mplus includes many lines, especially as the number of items and the number of attributes measured by each item increases. SAS and R macros have been developed to generate the Mplus code for DCMs, but any manual adjustments that need to be made must be handled carefully and require meticulous attention. Some DCM applications require manually editing the Mplus code generated by the macros. For example, to estimate a DCM that measures misconceptions instead of skills, a modification of the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model must be made to create the bug-DINO model (Schellman, 2021).

R is a free, open-source program, which means that the packages and procedures are generally not as well-vetted as Mplus. Therefore, R users must proceed with caution when using packages written by others. However, R has the benefit over Mplus in that it has packages that were specifically designed to accommodate DCM analyses, so R does not have to be "rigged" to estimate DCMs. This is why the code for estimating DCMs in R is simpler than the code for estimating DCMs in Mplus: The R code used for DCM estimation is a simple, one-line code for many applications, while the Mplus code is long and complex for even the simplest applications.

This dissertation utilizes a complex DCM called the polytomous transition DCM (PTDCM; Madison et al., 2021) that estimates student classifications for multiple attributes with multiple proficiency statuses over multiple testing occasions—I describe this model in more detail in Chapter 2. Currently, no package in R supports the

estimation of this particular DCM, so I utilized Mplus for all analyses in this dissertation. However, the Mplus code is long and must be manually constructed because the SAS and R macros mentioned above do not generate code for the PTDCM. Additionally, as with any model, the chance of successful convergence for the PTDCM decreases as the assessment design gets more complex (i.e., increase the number of attributes, the number of proficiency statuses, and/or the number of testing occasions), but even for simple uses of the PTDCM (e.g., three attributes, each with three proficiency statuses measured at two testing occasions) estimation is time-consuming, and the convergence rate is low. See Chapter 4 for my discussion about estimation time and convergence rates for my simulation study with the PTDCM.

Mplus and R are both command line interfaces, requiring programming knowledge and specific knowledge about DCM parameterization and estimation. No point-and-click program yet exists for DCM estimation. Thus, the choice of software is a trade-off in terms of cost, programming complexity, and trustworthiness of results.

Comparing Student Growth

As DCMs become more popular in research and in practice, it is important to ensure that research supports specific methodologies and practices related to applying DCMs. Specifically, I focus this section on one practical scenario using DCMs: comparing student growth.

Stakeholders in grade-school education are increasingly concerned with measuring student growth from year to year or from pre-test to post-test. However, it is argued that simply knowing how students' scores change over time—known as *gains* (Castellano & Ho, 2013a)—is not enough information. Stakeholders are also interested in

evaluating whether their students' individual growth or a class's average growth is good or bad. They want to know how students' growth compares to other students' growth. The *student growth percentile* (SGP; Betebenner, 2009) was developed as a metric for comparing student growth to the growth of other students who have similar score histories (i.e., similar scores in the previous school year; similar scores at the pre-test). SGPs are typically used with assessments that provide sum scores or scaled scores generated from IRT models. However, such assessments cannot satisfy the growing demand for reliable results that highlight students' individual and specific strengths and weaknesses. This demand can be satisfied through diagnostic assessments—assessments designed for use with DCMs. The DCM framework is beginning to break out of the realm of research and is now being used in practice (e.g., Bradshaw et al., 2014; Bradshaw et al., 2017-2021), so it is important to evaluate the use of metrics for modeling growth in the DCM framework. Therefore, this dissertation introduces the *diagnostic growth percentile* (DGP)—a metric that was adapted from the SGP for use with student results from DCMs and evaluates its student classification reliability and utility.

Chapter 2 gives a thorough description of SGPs and their limitations. In fact, Sireci and colleagues' 2016 article on SGPs encourages practitioners to stop using the SGP because of its many limitations. So, you might be wondering why we are discussing a diagnostic version of a metric that researchers want people to stop using. This dissertation is meant to be an exploration of the DGP and to consider its potential for use before educators have a need for it. SGPs were in use before they were fully researched, and the research that was conducted after implementing SGPs showed some key issues and concerns. This dissertation is seeking to switch the order of events for the diagnostic

framework. I aim to thoroughly investigate the metric before anyone tries to use it in practice. Specifically, this dissertation seeks to answer two primary research questions and a few sub-research questions:

1. How do we conceptualize growth percentiles in a DCM framework?
    a. What are adjustments to DGPs to account for uncertainty in student classification from the DCM?
    b. How can DGPs be adjusted to consider penalties for forgetting?
    c. What are the implications of DGP interpretability for different adjusted DGPs?
    d. How can DGP reliability be conceptualized and computed?
2. How reliable are DGPs under different DCM assessment conditions?
    a. How do the reliability metrics for DGPs perform under different assessment conditions?

In Chapter 2, I describe the key limitations of the SGP. In Chapter 6, I address these limitations in the context of DGPs to answer three primary discussion questions:

1. Which SGP limitations remain with DGPs?
2. Which SGP limitations are lessened when using DGPs?
3. Do DGPs have any new limitations that SGPs do not have?

Overview of This Dissertation

Chapter 2 provides a literature review that describes how SGPs are used and calculated and some of the issues related to their use. In Chapter 2, I also describe and provide the parameterizations for a general DCM and three extensions of the general model. In Chapter 3, I introduce the basic DGP, which can be situated in the framework

for growth models using the critical questions proposed by Castellano and Ho (2013b). After introducing the basic DGP, I introduce three approaches for adjusting the DGP to accommodate uncertainty in student classifications from the DCM. Additionally, as with any method in measurement, it is critical that the DGP has reliability metrics that can be computed to evaluate the student classification and DGP consistency. Therefore, Chapter 3 also includes a description of the metrics that can be used for measuring DGP reliability, which I adapted from the reliability metrics for a general longitudinal DCM (Madison, 2019; Schellman & Madison, in press). In Chapter 4, I explain the design and results of the simulation study I conducted to investigate how DGP reliability is impacted by various assessment conditions that are common in DCM literature. In Chapter 5, I explain three empirical data analyses I conducted to illustrate a practical use of the DGPs metrics and their reliabilities. Finally, in Chapter 6, I discuss the significance of this dissertation, including how it adds to the literature and provides tools for researchers who are seeking to use DCMs in practice in educational settings.

CHAPTER 2

LITERATURE REVIEW

In this chapter, I describe SGPs including how to compute them, how to interpret them, what makes them attractive, and why they should not be used in practice. Next, I describe DCMs including the parameterizations for a general DCM as well as several DCMs that extend the general DCM to accommodate multiple testing occasions or polytomous attributes, or both.

Student Growth Percentiles

The SGP is a metric that was developed to compare a student's growth over time with the growth of other students who have the same score histories. Thus, rather than reporting only the magnitude of a student's growth (e.g., Student X's score increased by 33 points), which could be difficult to interpret in isolation, especially in the absence of vertical scaling, SGPs allow for a "normative quantification of student growth" (Betebenner, 2009, p. 44). More specifically, the SGP uses a distribution that is conditional on the student's previous score(s) to locate their current score, which provides the student's SGP (Monroe & Cai, 2015).

SGPs have two primary advantages (Sireci et al., 2016). First, they give students credit for growth not reflected in their overall *achievement level*. In this dissertation, the phrase "achievement level" is reserved for references to summative assessment reporting categories, while the phrase "proficiency status" is reserved for references to student classifications in the DCM framework. For example, consider a state that uses four

achievement levels: beginning, developing, proficient, and distinguished. A student might fall into the proficient level in fourth grade and in fifth grade, even if they had positive gains in their scores between grades. The gain-score growth is not reflected in the student's achievement levels across grades because the growth was not enough for them to move from the proficient category to the distinguished category. SGPs are a measure of growth that can supplement students' achievement level classifications.

Second, SGPs do not require vertical scaling from grade to grade (Betebenner, 2009; Sireci et al., 2016), making them a flexible metric for evaluating growth. Because SGPs do not consider the magnitude of growth, they are invariant to the relative scales of the assessments used in the SGP computation (Betebenner, 2009). In fact, because vertical scaling is not a requirement for SGPs, SGPs allow for the inclusion of assessments that measure different latent variables over time. So, you could use SGPs if you are interested in quantifying growth for students who took a science assessment last year and a history assessment this year, though, with different latent variables over time, "growth" might not be the most appropriate term (Betebenner, 2009).

Additionally, SGPs overcome a limitation of growth-to-standard, or trajectory, models, in which student growth is predicted for future testing occasions based on the assumption that students will show the same amount of growth that they showed in the past (Castellano & Ho, 2013a). In growth-to-standard models, schools with many students who are already high-performing will show greater proportions of students who are projected to be high-performing in the future, while schools with students who are not already high-performing will show lower proportions of students who are projected to be high-performing, and the conclusion would be that the already high-performing schools

are of greater quality than are the schools with fewer high-performing students (Betebenner, 2009). SGPs allow for a separation of students' present achievement and their growth, which cannot be separated in growth-to-standard models.

Despite these key advantages and the general attraction to a metric that compares students' growth with that of students with similar score histories, the SGP, as it was originally introduced, requires complicated statistics and is often calculated incorrectly. The use of SGPs is further complicated because of misunderstandings about what SGPs are and how to use and interpret them. Additionally, the measurement community has voiced concerns about the premature use of SGPs in practice, as they were implemented before being thoroughly researched as required by the *Standards of Educational and Psychological Testing* (AERA, APA, NCME, 2014). The following sections explain these complexities and complications.

Calculation of SGPs

This section describes the two primary methods of computing a student's SGP: the *quantile regression approach* and the *cohort approach*. Betebenner (2009) introduced SGPs for the first time and implemented the quantile regression approach. However, the computation of SGPs via the quantile regression approach is complex because the statistics needed to compute the conditional density distribution and its associated regression lines are unfamiliar to many people (Castellano & Ho, 2013a). Therefore, users often implement the cohort approach, either as an intentional simplification of the SGP metric or due to misunderstandings of how to compute the SGP as it was originally intended. For both approaches, the SGP takes on whole number values between 0 and

100, and students' scores are unrelated to their SGPs because students with any score can obtain any SGP value, depending on their academic peers (Castellano & Ho, 2013a).

Quantile Regression Approach

The original approach for computing SGPs is conceptually intuitive: It involves estimating a density function for students' current scores (i.e., scores at testing occasion T), conditional on students' previous scores (i.e., scores at testing occasions 1, 2, …, $T-1$), but the computation is tricky. A student's SGP is the percentile of their current score within the conditional density function, which indicates the likelihood that the student's current score is observed given the student's previous scores. In other words, the SGP is 100 times the probability of observing the current score given the past score(s) (Betebenner, 2009).

Quantile regression is required for estimating the conditional density function of SGPs. Quantiles are values on the scale of the data that split the data into equal groups. For example, the median of a data set is a quantile because it splits the data into two equal groups; quartiles are quantiles that split the data into four equal groups; and percentiles are quantiles that split the data into 100 equal groups (Koenker, 2005). Note that some people argue that quantiles and percentiles have the same meaning but are represented as decimals and percentages, respectively (Ford, 2015). Quantiles can be represented mathematically by $0 < \tau < 1$. Then, the 2-quantile (median) is given by $\tau = .5$. The 4-quantile (quartiles) is given by $\tau = .25$. The 100-quantile (percentiles) is given by $\tau = .01$.

Quantile regression is best understood by first considering linear regression. In linear regression, the goal is to estimate an expected mean of a dependent variable

conditioned on one or more independent variables. In linear regression, we estimate the mean regression line. What if we, instead, want to estimate the median (the $\tau = .5$ quantile)? Then, we would estimate the median regression line. What if we want to estimate the quartiles (the $\tau = .25$, $\tau = .5$, and $\tau = .75$ quantiles)? Then, we would estimate three quartile regression lines. When applying quantile regression in the context of SGPs, 100 quantile regression lines (for the $\tau = .01$, $\tau = .02$, …, $\tau = .98$, $\tau = .99$ quantiles) are computed (Betebenner et al., 2022). Quantile regression is beneficial for use with SGPs because it does not require assumptions about the distribution of the outcome variable (the current scores), and it is resistant to outliers (Castellano & Ho, 2013a). Thus, quantile regression can be used in place of linear regression for situations in which the assumptions required for linear regression are violated, which happens if the data is nonlinear, heteroscedastic, non-independent, or non-normal (Castellano & Ho, 2013b).

Quantile regression estimates the quantiles of the current scores given all previous scores and other predictor variables (Castellano & Ho, 2013b). The predictor variable(s) may be continuous or categorical, but the dependent variable (the current score) must be continuous. Quantile regression uses B-spline cubic basis functions to "smooth irregularities found in the multivariate assessment data" (Betebenner, 2009, p. 46). In this smoothing, the SGPs for groups of academic peers are determined by using information not just from the current academic peer group but also from other academic peer groups that have similar score histories to those of the current group (Castellano & Ho, 2013a). This borrowing of information from nearby peer groups mitigates the issue that sample sizes for academic peer groups decrease as the number of previous assessments increases.

A further discussion of B-spline computation is beyond the scope of this dissertation. Here, it is sufficient to understand (1) that the B-spline approach to quantile regression yields "99 nonlinear regression curves … for each level of … the prior year's test score" (Wells & Sireci, 2020, p. 351), (2) the regression curves are flexible to data irregularities and do not cross, which is critical for determining students' SGPs, (3) R packages are available for running quantile regression in general (e.g., the quantreg package; Koenker, 2022) and for computing SGPs via the quantile regression approach (i.e., the SGP package; Betebenner et al., 2022), and (4) a student's SGP is given by the value of quantile regression curve closest to the student's current score (Wells & Sireci, 2020) or as the average of the values for the two quantile regression curves in between which the student's current score lies (Castellano & Ho, 2013b).

To illustrate how the conditional quantile regression lines can be used to compute a student's SGP using the midpoint approach described in (4) above, consider Figure 1, which shows the B-spline conditional regression deciles (the $\tau = .1$, $\tau = .2$, …, $\tau = .8$, and $\tau = .9$ quantiles) for some students' scaled scores from Grades 5 and 6. Figure 1 is copied from Betebenner's Figure 3 (Betebenner, 2009, p. 47), which also includes a linear version of the conditional regression deciles to show how the B-spline version better fits the data. In Figure 1, I added a purple dot to show a student with a Grade 5 scaled score of 400 and a Grade 6 scaled score of about 460. This student's bivariate score lies between the $\tau = .8$ and $\tau = .9$ quantile (specifically, decile) regression curves. Therefore, this student's SGP is 85 because 85 equals 100 times the midpoint between these two values for the quantile regression curves: $100\left(\frac{.9+.8}{2}\right) = 85$.

Cohort Approach

The quantile regression approach is complicated, and users often struggle to implement it correctly. In a review of materials used by states such as Georgia, Michigan, Mississippi, and Virginia (Michigan Department of Education, n.d.) to explain SGPs to the public, SGP computation is described using an approach that differs from Betebenner's (2009) original approach. In such materials, SGPs are described as the percentiles that are obtained from ranking a cohort of students with similar past scores by their current scores, which is equivalent to ranking them by the differences (growth) between their current scores and past scores.

To illustrate how this computation differs from Betebenner's approach, let us consider a situation in which students in a given state have taken an end-of-year test for each of the fourth, fifth, and sixth grades, and we are interested in Student X's SGP. Student X scored 434 in fourth grade, 467 in fifth grade, and 490 in sixth grade. We know that Student X's score grew by 33 points between fourth and fifth grade and by 23 points between fifth and sixth grade. However, we do not know how this growth compares with the growth of other similar students.

First, we must identify Student X's cohort based on their previous scores. Student X's cohort includes all sixth-grade (Student X's current grade) students in the state who had a fourth-grade score of 434 and a fifth-grade score of 467. This cohort is used to compute Student X's (and the other students in the cohort's) sixth-grade SGP. Student X's cohort can be rank-ordered by their sixth-grade scores (or, equivalently, the differences between their sixth- and fifth-grade scores). Student X scored 490 in sixth grade, which is a difference of +23 points from fifth grade. Suppose this score of 490 (or

a difference of 23) points puts Student X at the 87th percentile for growth within their cohort. Then, Student X's SGP is 87, which can be interpreted to mean that Student X showed growth greater than or equal to the growth of 87% of the students in the cohort.

Thus, the cohort approach is literally the percentile of a student's current score, given their past scores, and it is often used as a heuristic to help explain, in general, how the true SGP is computed (Castellano & Ho, 2013a). However, the cohort approach yields very small groups of students, especially as the number of previous assessments increases, and can yield imprecise percentile ranks (Castellano & Ho, 2013a), while the quantile regression approach does not require a literal cohort at all. The quantile regression and cohort approaches yield different interpretations that are often confused and present several other issues related to the use of SGPs.

Limitations and Issues Related to SGPs

The previous section introduced the key limitations of both approaches to SGP computation: the quantile regression approach is complicated, and the cohort approach is inaccurate in terms of how the SGP was originally intended. This section elaborates on some limitations and issues related to SGP use via the quantile regression and cohort approaches.

SGP Computation is Misunderstood

As mentioned above, the cohort approach is easy to describe and compute, which is likely why it is so popular in practice. However, it is not the metric that was originally proposed by Betebenner (2009), and there are several key issues with its use. The cohort approach can be considered a theoretical overview of SGPs (Wells & Sireci, 2020), but cohorts were not part of the original SGP computation. Quantile regression does not

require cohorts and does not compute simple percentiles based on rank ordering. The "cohort" in the quantile regression is derived from the model (Wells & Sireci, 2020). Therefore, states that create materials such as those described above are inaccurately describing SGPs (Sireci et al., 2016). Part of this confusion could be that the word "percentile" is included in the name (Castellano & Ho, 2013b). SGPs based on the quantile regression approach are not percentiles as we typically understand and use them because they do not involve rank-ordering students and dividing them into equal groups. SGPs are, instead, estimates of the probability of observing a student's score pattern (Sireci et al., 2016), which is a percentile based on a conditional density function.

Proper Use of SGPs is Misunderstood

Stakeholders often do not know how to use SGPs. First, although SGPs should not be given value judgements such as "good growth" or "poor growth" without using standard setting procedures with content experts (Betebenner, 2009), some stakeholders may attempt to categorize SGPs to indicate, for example, low, moderate, or high growth without implementing standard setting procedures for the SGP. Second, because SGP computation and interpretations are inconsistent, users often make incorrect instructional decisions (Sireci et al., 2016). For example, consider Student X, who scored 490 on the sixth-grade end-of-year test and has an SGP of 87. Student Y scored 495 on the sixth-grade end-of-year test but has an SGP of 27. Some school administrators would incorrectly conclude that Student Y needs more remediation than Student X because Student Y's SGP was less than Student X's SGP, even though Student Y has the greater scaled score (Sireci et al., 2016).

Measurement Concerns Related to SGP Use

Research has shown that SGPs are not reliable and have large margins of error (Sireci et al., 2016). However, SGPs are being used to make important decisions at the student, teacher, and school levels. It is concerning that SGPs were put into practice without researchers first conducting thorough studies related to the reliability of the metric and the validity of the intended interpretations, which seem to vary widely. The lack of evidence to support the use of SGPs at the individual and group levels violates the *Standards* (Sireci et al., 2016).

Finally, SGPs step away from criterion-referenced measurement and toward norm-referenced comparisons of students (Sireci et al., 2016). Criterion-referenced assessments are designed to report students' statuses with respect to the latent variables that are measured by the assessments, while norm-referenced assessments are designed to report students' statuses with respect to their peers. Most assessments today are designed to be criterion-referenced and report achievement levels (e.g., beginning, developing, proficient, and distinguished) rather than comparisons across students. However, SGPs use the results from criterion-referenced assessments to generate norm-referenced results, violating the intended use of the assessments (Sireci et al., 2016). This issue is of particular concern with diagnostic measurement, which is discussed below.

Interpretations of SGPs

Interpreting SGPs depends on how they are calculated. Even then, interpretations seem to vary. Let us use an SGP of 58 to show how interpretations vary. Under the quantile regression approach, a student's SGP of 58 can be interpreted in two ways: (1) There is a 58% chance of observing the student's current score given their past scores

(Betebenner, 2009), and (2) The likelihood estimate of the student's particular score pattern is 58 (Clauser et al., 2016). For the cohort approach, the literature includes several different interpretations of an individual student's SGP of 58: (3) A student with an SGP of 58 indicates "that this student's current achievement is higher than [58%] of students who share the same score history … SGPs can add to our understanding of how well students are doing, and how they are progressing" (Monroe et al., 2014, p. 2), and (4) An SGP of 58 indicates the "percentile rank of [the] student's current score relative to those students at the same grade level who share the same prior score(s)" (Shang et al., 2015, p. 2). Regardless of whether the quantile regression or cohort approach is used to compute SGPs, the literature agrees about these two facets of SGP interpretation: (5) SGPs should not be interpreted as "low" or "high" growth without standard setting procedures (Betebenner, 2009), and (6) "SGPs do not represent growth in terms of how much students have learned in a given subject area" (Sireci et al., 2016, p. 3).

SGPs can be aggregated to a group level (classroom, school, district, etc.) by averaging (or otherwise aggregating) individual students' SGPs across a specific group. Aggregated SGPs are often interpreted as a measure of teacher, school, or district effectiveness (Monroe et al., 2014). However, it is argued that this interpretation is an inappropriate use of SGPs because the standard error of aggregated SGPs is so large that it covers nearly all of the SGP scale, meaning that almost any value on the scale could be the true value (Sireci et al., 2016; Wells & Sireci, 2020).

Diagnostic Classification Models

As mentioned above, DCMs are psychometric tools that model one or more fine-grained categorical latent variables using categorical item response data (i.e., responses to

assessment items that are scored as correct or incorrect). This contrasts with the popular IRT models, which model coarse-grained continuous latent variables using categorical item response data (Hambleton et al., 1991). The categorical latent variables (known as *attributes* in the DCM framework) that are modeled by DCMs are typically binary and indicate whether students are proficient or non-proficient. Below, I discuss how to model polytomous rather than dichotomous attributes, but for this section, I am considering only dichotomous attributes. Table 1 summarizes the notation and indices used throughout this dissertation to refer to the various components of the DCM framework. This chapter uses many examples to illustrate these components.

Student-level results from DCMs are proficiency statuses—one for each attribute measured by the assessment rather than sum scores or scaled scores. Consider a diagnostic assessment that measures three attributes. Suppose that the DCM indicates that Student X is likely proficient for Attribute 1, non-proficient for Attribute 2, and proficient for Attribute 3. Student X's proficiency statuses for these three attributes can be represented as an *attribute profile,* which is the vector [101], where 1's indicate proficiency and 0s indicate non-proficiency for the corresponding attributes. In general, a diagnostic assessment that measures A dichotomous attributes has 2^A unique attribute profiles which characterize the C latent classes ($C = 2^A$) and are denoted $\boldsymbol{\alpha_c}$ for the cth class. For the example three-attribute diagnostic assessment, the $2^3 = 8$ attribute profiles are $\boldsymbol{\alpha_1} = [000]$, $\boldsymbol{\alpha_2} = [001]$, $\boldsymbol{\alpha_3} = [010]$, $\boldsymbol{\alpha_4} = [011]$, $\boldsymbol{\alpha_5} = [100]$, $\boldsymbol{\alpha_6} = [101]$, $\boldsymbol{\alpha_7} = [110]$, and $\boldsymbol{\alpha_8} = [111]$. The individual proficiency statuses for attribute a within the attribute profile for latent class c are denoted α_{ca}. For example, Latent Class 4 has the

profile $\boldsymbol{\alpha_4} = [011]$, so the individual proficiency statuses are $\alpha_{4,1} = 0$, $\alpha_{4,2} = 1$, and $\alpha_{4,3} = 1$.

Because the profiles clearly show the attributes for which a student is proficient and the attributes for which the student could use additional instruction and support, the DCM framework is well-suited to satisfy the demand for specific and actionable student information in a formative assessment process (Black & Wiliam, 1998).

Similar to how the attribute profiles show the attributes for which a student is proficient, *q-vectors* show which attributes are measured by the items in the diagnostic assessment. The q-vector for item i is an A-length vector of the form $[q_{i1}, \ldots, q_{ia}, \ldots, q_{iA}]$, where $q_{ia} = 1$ if item i measures the ath attribute, or $q_{ia} = 0$ if item i does not measure the ath attribute. The q-vectors can be compiled into an $I \times A$ matrix called a *Q-matrix*, where I is the number of items in the diagnostic assessment. The hypothesized item-attribute alignment represented in the Q-matrix is specified before any DCM analysis is conducted. The DCM analysis provides results that support or refute the Q-matrix specifications.

The DCM family contains many models that are different in how they parameterize the item responses because of the assumptions that users or developers choose to make about the item-attribute relationships. The next few sections describe and provide the parameterizations for a general DCM—the *log-linear cognitive diagnosis model* (LCDM; Henson et al., 2009)—and three extensions of the LCDM: the *transition DCM* (TDCM; Madison & Bradshaw, 2018), the *polytomous DCM* (PDCM; Bao, 2019), and the PTDCM (Madison et al., 2021).

The Log-Linear Cognitive Diagnosis Model (LCDM)

The LCDM is a general DCM that can yield most other DCMs when different sets of parameters are constrained. The full latent class model is given by Equation 1.

$$P(\boldsymbol{X_r} = \boldsymbol{x_r}) = \sum_{c=1}^{C} \upsilon_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1 - x_{ir}} \tag{1}$$

where $\boldsymbol{x_r}$ is the observed item response pattern for student r and $r \in \{1, \ldots, N\}$, $C = 2^A$ for dichotomous attributes, υ_c is the probability of a student belonging to class c, π_{ic} is the probability of a correct response (known as the *item response probability*; IRP) to item i for a student in latent class c, and x_{ir} is student r's observed scored response to item i. The part of the model included in the summation ($\sum\cdot$) makes up the *structural component*, and the υ_c parameters follow the Bernoulli distribution. The part of the model included in the product ($\prod\cdot$) makes up the *measurement component.* The expression $\pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1 - x_{ir}}$ simplifies to π_{ic} or $1 - \pi_{ic}$ for a correct and incorrect response to item i, respectively, and gives the probability or likelihood of observing the item response x_{ir}. Therefore, the product of the item-level likelihoods gives the joint likelihood of observing the entire response pattern, $\boldsymbol{x_r}$, given membership to each latent class (Rupp et al., 2010).

Rupp et al. (2010) give the LCDM equation for computing the IRP to item i for a student in latent class c with the attribute profile $\boldsymbol{\alpha_c}$ as

$$\pi_{ic} = P(x_i = 1 | \boldsymbol{\alpha}_c) = \frac{\exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i})\right)}{1 + \exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i})\right)} \tag{2}$$

The *kernel* of Equation 2, $\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i})$, shows the log odds of a correct response for a student in latent class c with the latent profile $\boldsymbol{\alpha}_c$. The kernel can be expanded to show the individual item parameters:

$$\text{Kernel}_i = \lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1(a)} \alpha_{ca} q_{ia} + \sum_{a=1}^{A} \sum_{a' \neq a} \lambda_{i,2(a,a')} \alpha_{ca} \alpha_{ca'} q_{ia} q_{ia'} + \cdots \tag{3}$$

$\lambda_{i,0}$, $\lambda_{i,1(a)}$, and $\lambda_{i,2(a,a')}$ indicate the item parameters for item i, α_{ca} and $\alpha_{ca'}$ indicate the proficiency statuses for attributes a and a', respectively, for students in latent class c, and q_{ia} and $q_{ia'}$ indicate the Q-matrix entries for item i and attributes a and a', respectively. In this expansion, one can see how the LCDM models items responses in a manner similar to a latent variable analysis of variable (ANOVA) model (Bradshaw, 2016).

To illustrate the LCDM, let us first consider an example: Suppose that Item 2 on the example three-attribute diagnostic assessment measures only Attribute 3. The probability that a student with the attribute profile $\boldsymbol{\alpha}_c$ will provide a correct response to Item 2 is

$$\begin{aligned} P(x_2 = 1 | \boldsymbol{\alpha}_c) &= \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1(a)} \alpha_{ca} q_{ia}\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1(a)} \alpha_{ca} q_{ia}\right)} \\ &= \frac{\exp\left(\lambda_{2,0} + \lambda_{2,1(3)} \alpha_{c3} q_{23}\right)}{1 + \exp\left(\lambda_{2,0} + \lambda_{2,1(3)} \alpha_{c3} q_{23}\right)} \end{aligned} \tag{4}$$

The $\lambda_{i,0}$ is the intercept parameter for Item 2, and it indicates the log odds that a student who is non-proficient for any of the attributes measured by Item 2 (i.e., a student who is non-proficient for Attribute 3) will provide a correct response.

The $\lambda_{2,1(3)}$ is the main effect parameter for Item 2 and Attribute 3. This main effect parameter indicates the increase in the log odds of a correct response that a student

gets for being proficient for Attribute 3. The proficiency status, α_{c3}, shows 0 or 1 depending on the attribute profile for students in latent class c for Attribute 3. The Q-matrix entry for Item 2 and Attribute 3, q_{23}, is 1 because Item 2 measures Attribute 3. Therefore, Equation 4 can be simplified in two ways depending on the proficiency status: If students in class c are non-proficient for Attribute 3 ($\alpha_{c3} = 0$), then their IRP for Item 2 is

$$P(x_2 = 1|\alpha_{c3} = 0) = \frac{\exp(\lambda_{2,0})}{1 + \exp(\lambda_{2,0})} \quad (5)$$

If students in class c are proficient for Attribute 3 ($\alpha_{c3} = 1$), then their IRP for Item 2 is

$$P(x_2 = 1|\alpha_{c3} = 1) = \frac{\exp(\lambda_{2,0} + \lambda_{2,1(3)})}{1 + \exp(\lambda_{2,0} + \lambda_{2,1(3)})} \quad (6)$$

In the LCDM, the IRP increases as the number of attribute proficiencies increases, so the main effects are positive, and $P(x_2 = 1|\alpha_{c3} = 0) < P(x_2 = 1|\alpha_{c3} = 1)$.

The other items on the example diagnostic assessment can be similarly parameterized, but as the number of attributes measured by an item increases, the number of terms in the LCDM kernel increases. Items that measure two attributes have IRPs given, in general, by

$$P(x_i = 1|\boldsymbol{\alpha}_c) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1(a)}\alpha_{ca}q_{ia} + \lambda_{i,1(a')}\alpha_{ca'}q_{ia'} + \lambda_{i,2(a,a')}\alpha_{ca}\alpha_{ca'}q_{ia}q_{ia'}\right)}{1 + \exp(\lambda_{i,0} + \lambda_{i,1(a)}\alpha_{ca}q_{ia} + \lambda_{i,1(a')}\alpha_{ca'}q_{ia'} + \lambda_{i,2(a,a')}\alpha_{ca}\alpha_{ca'}q_{ia}q_{ia'})} \quad (7)$$

In this kernel, the first term, $\lambda_{i,0}$, is the intercept, the second term includes the main effect for attribute a, $\lambda_{i,1(a)}$, the third term includes the main effect for attribute a', $\lambda_{i,1(a')}$, and the fourth term includes an interaction effect for attributes a and a', $\lambda_{i,2(a,a')}$. A two-way interaction parameter indicates the change in the log odds of a correct response that a student gets for being proficient for both of the corresponding attributes. The IRP

equation for items that measure two attributes can be simplified in four ways, depending on the attribute proficiency statuses: If students in class c are non-proficient for attributes a and a' ($\alpha_{ca} = 0, \alpha_{ca'} = 0$), then their IRP for item i is

$$P(x_i = 1|\alpha_{ca} = 0, \alpha_{ca'} = 0) = \frac{\exp(\lambda_{i,0})}{1 + \exp(\lambda_{i,0})} \quad (8)$$

If students in class c are non-proficient for attribute a and proficient for attribute a' ($\alpha_{ca} = 0, \alpha_{ca'} = 1$), then their IRP for item i is

$$P(x_i = 1|\alpha_{ca} = 0, \alpha_{ca'} = 1) = \frac{\exp(\lambda_{i,0}) + \lambda_{i,1(a')}}{1 + \exp(\lambda_{i,0}) + \lambda_{i,1(a')}} \quad (9)$$

If students in class c are proficient for attribute a and non-proficient for attribute a' ($\alpha_{ca} = 1, \alpha_{ca'} = 0$), then their IRP for item i is

$$P(x_i = 1|\alpha_{ca} = 1, \alpha_{ca'} = 0) = \frac{\exp(\lambda_{i,0}) + \lambda_{i,1(a)}}{1 + \exp(\lambda_{i,0}) + \lambda_{i,1(a)}} \quad (10)$$

If students in class c are proficient for attributes a and a' ($\alpha_{ca} = 1, \alpha_{ca'} = 1$), then their IRP for item i is

$$P(x_i = 1|\alpha_{ca} = 1, \alpha_{ca'} = 1) = \frac{\exp(\lambda_{i,0}) + \lambda_{i,1(a)} + \lambda_{i,1(a')} + \lambda_{i,2(a,a')}}{1 + \exp(\lambda_{i,0}) + \lambda_{i,1(a)} + \lambda_{i,1(a')} + \lambda_{i,2(a,a')}} \quad (11)$$

An item that measures three or more attributes can be parameterized similarly, where the higher order interaction terms are added as the number of attributes measured increases. Higher dimensional item parameters can be interpreted like the two-way interaction parameter: the change in the log odds of a correct response for being proficient for all corresponding attributes. Thus, the LCDM is a saturated model that includes all possible item parameters given the attributes that each item measures.

The Transition DCM (TDCM)

The TDCM is a longitudinal extension of the LCDM to model student attribute proficiency over time (Madison & Bradshaw, 2018). The TDCM is given by Equation 12

$$P(\boldsymbol{X_r} = \boldsymbol{x_r}) = \sum_{c_1=1}^{C} \sum_{c_2=1}^{C} \dots \sum_{c_T=1}^{C} \upsilon_{c_1} \tau_{c_2|c_1} \tau_{c_3|c_2} \dots \tau_{c_T|c_{T-1}} \prod_{t=1}^{T} \prod_{i=1}^{I} \pi_{ic_t}^{x_{irt}} \left(1 - \pi_{ic_t}\right)^{1-x_{irt}} \quad (12)$$

where T is the total number of testing occasions, $C = 2^{AT}$ for dichotomous attributes, υ_{c_t} is the probability of a student belonging to class c at testing occasion t, $\tau_{c_T|c_{T-1}}$ is the probability of a student transitioning from one attribute proficiency status to another between testing occasions $T-1$ and T, π_{ic_t} is the IRP to item i for a student in latent class c at testing occasion t, which uses the same LCDM parameterization presented above, and x_{irt} is student r's observed scored response to item i at testing occasion t (Madison & Bradshaw, 2018). Note that the LCDM is a special case of the TDCM when $T = 1$. Additionally, note that tau has a different meaning in the TDCM framework than it had in the SGP framework.

To better illustrate the transition probabilities, consider a situation with two testing occasions ($T = 2$). With two testing occasions, Equation 12 simplifies to

$$P(\boldsymbol{X_r} = \boldsymbol{x_r}) = \sum_{c_1=1}^{2^A} \sum_{c_2=1}^{2^A} \upsilon_{c_1} \tau_{c_2|c_1} \prod_{t=1}^{2} \prod_{i=1}^{I} \pi_{ic_t}^{x_{irt}} \left(1 - \pi_{ic_t}\right)^{1-x_{irt}} \quad (12)$$

Students can be classified as proficient or non-proficient for a given attribute at the first testing occasion. They can then be classified as proficient or non-proficient the attribute at the second testing occasion. Thus, students could have four different attribute proficiency patterns across the two testing occasions: proficient first and proficient second, proficient first and non-proficient second, non-proficient first and proficient

second, or non-proficient first and non-proficient second. Each of these patterns (or *transitions*) has a probability, $\tau_{c_2|c_1}$, associated with it. Let C' indicate the total number of attribute-level transitions, which is computed as $C' = 2^T$ when only dichotomous attributes are used. In other words, C' denotes the number of attribute-level transitions, which is not always the same as the number of latent classes because attribute-level transitions are often subsets of latent classes, so it is important to keep the index for latent classes (C) and the index for attribute-level transitions (C') distinct. I chose to use C' to denote the number of attribute-level transitions because when only one attribute is being measured $C' = C$ because the latent classes are the same as the attribute-level transitions. For the TDCM with only dichotomous attributes, C' is identical for all attributes.

Measurement invariance is a theoretical requirement for measurement models that establishes item-invariant measurement of students and student-invariant measurement of items (Engelhard, 2013). In other words, estimating student's proficiency statuses is independent of the specific items used in the diagnostic assessment, and estimating item parameters is independent of the specific students used for calibration. For longitudinal studies, measurement invariance holds when item parameters from different testing occasions are equal, and most longitudinal DCM studies assume that invariance holds over time to keep the meaning of the proficiency statuses consistent over time (Madison & Bradshaw, 2018). DCMs have theoretical invariance properties (de la Torre & Lee, 2010; Bradshaw & Madison, 2016). Thus, student classifications do not depend on the particular items on the students' assessments, and item parameters do not depend on the specific sample of students who saw the items. However, these properties are theoretical and only hold with adequate model fit. When assuming measurement invariance,

Equations 12 and 13 change slightly: the IRP, π_{ic_t}, changes to π_{ic} because each student's IRP is no longer dependent on which testing occasion, t, is being considered.

In this dissertation, to simplify the language, let "T1" represent the first testing occasion, let "T2" represent the second testing occasion, and so on.

The Polytomous DCM (PDCM)

Up to this point, I have discussed only dichotomous attributes that have binary statuses such as proficient and non-proficient. The PDCM is a polytomous extension of the LCDM (Bao, 2019). The PDCM allows for modeling attributes that have any number of proficiency statuses to indicate a categorical degree of proficiency. Let l_a denote the total number of proficiency statuses for attribute a. For diagnostic assessments with polytomous attributes, the total number of latent classes is given by $C = \prod_{a=1}^{A} l_a$.

For example, a trichotomous attribute ($l_a = 3$) might have proficiency status labels such as *beginning*, *developing*, and *proficient*. Consider a diagnostic assessment that measures one attribute with three proficiency statuses ($l_1 = 3$) and one attribute with four proficiency statuses ($l_2 = 4$). The $C = \prod_{a=1}^{A} l_a = l_1 \times l_2 = 3 \times 4 = 12$ attribute profiles for this assessment are $\boldsymbol{\alpha_1} = [00]$, $\boldsymbol{\alpha_2} = [01]$, $\boldsymbol{\alpha_3} = [02]$, $\boldsymbol{\alpha_4} = [03]$, $\boldsymbol{\alpha_5} = [10]$, $\boldsymbol{\alpha_6} = [11]$, $\boldsymbol{\alpha_7} = [12]$, $\boldsymbol{\alpha_8} = [13]$, $\boldsymbol{\alpha_9} = [20]$, $\boldsymbol{\alpha_{10}} = [21]$, $\boldsymbol{\alpha_{11}} = [22]$, and $\boldsymbol{\alpha_{12}} = [23]$, where the first attribute has the proficiency statuses 0, 1, and 2 ($\alpha_{c1} \in \{0,1,2\}$) and the second attribute has the proficiency statuses 0, 1, 2, and 3 ($\alpha_{c2} \in \{0,1,2,3\}$). For item i that measures a polytomous attribute a, students who have a greater proficiency status for attribute a have a greater understanding of the attribute and, thus, have a greater (or equivalent) IRP for item i than students who have a lower proficiency status for attribute a (Bao, 2019).

The rest of this section describes PDCM concepts using a diagnostic assessment that measures one attribute with five proficiency statuses ($l_1 = 5$ and $\alpha_{c1} \in \{0,1,2,3,4\}$) with the proficiency status labels (1) *beginning*, (2) *developing*, (3) *proficient*, (4) *distinguished*, and (5) *advanced*. The five attribute profiles and latent classes for this diagnostic assessment are $\boldsymbol{\alpha_1} = [0]$, $\boldsymbol{\alpha_2} = [1]$, $\boldsymbol{\alpha_3} = [2]$, $\boldsymbol{\alpha_4} = [3]$, and $\boldsymbol{\alpha_5} = [4]$. Table 2 shows the alignment of these proficiency status labels, latent classes, and attribute profiles, though this terminology can be used interchangeably.

The parameterization of the PDCM requires dummy coding (Bao, 2019). Table 2 shows the dummy coding for an attribute a that has five proficiency statuses. Notice that α_{ca}^2 cannot equal 1 unless α_{ca}^1 also equals 1, and α_{ca}^3 cannot equal 1 unless α_{ca}^2 and α_{ca}^1 also equal 1, and so on. Additionally, when $\alpha_{ca}^1 = \alpha_{ca}^2 = \alpha_{ca}^3 = 1$, for example, the student has reached the developing, proficient, and distinguished proficiency statuses.

The IRP for an item i that measures an attribute with five proficiency statuses is given by

$$\pi_{ic} = P(x_i = 1|\boldsymbol{\alpha}_c) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1(1)}^1 \alpha_{ca}^1 + \lambda_{i,1(1)}^2 \alpha_{ca}^2 + \lambda_{i,1(1)}^3 \alpha_{ca}^3 + \lambda_{i,1(1)}^4 \alpha_{ca}^4\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1(1)}^1 \alpha_{ca}^1 + \lambda_{i,1(1)}^2 \alpha_{ca}^2 + \lambda_{i,1(1)}^3 \alpha_{ca}^3 + \lambda_{i,1(1)}^4 \alpha_{ca}^4\right)} \quad (13)$$

If the diagnostic assessment contains only one attribute with five proficiency statuses, then Equation 14 simplifies to

$$P(x_i = 1|\boldsymbol{\alpha}_c) = \frac{\exp\left(\lambda_{i,0} + \lambda_i^1 \alpha_{ca}^1 + \lambda_i^2 \alpha_{ca}^2 + \lambda_i^3 \alpha_{ca}^3 + \lambda_i^4 \alpha_{ca}^4\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_i^1 \alpha_{ca}^1 + \lambda_i^2 \alpha_{ca}^2 + \lambda_i^3 \alpha_{ca}^3 + \lambda_i^4 \alpha_{ca}^4\right)} \quad (14)$$

where α_{ca}^1, α_{ca}^2, α_{ca}^3, and α_{ca}^4 are the dummy variables from Table 2. The intercept parameter for item i, $\lambda_{i,0}$, gives the log odds of a correct response for students at the beginning proficiency status ($\alpha_{ca}^1 = \alpha_{ca}^2 = \alpha_{ca}^3 = \alpha_{ca}^4 = 0$). The parameter λ_i^1 is the main effect for item i for the developing proficiency status and represents the increase in the

IRP when a student reaches the developing proficiency status ($\alpha_{ca}^{1} = 1, \alpha_{ca}^{2} = \alpha_{ca}^{3} = \alpha_{ca}^{4} = 0$). The parameter λ_{i}^{2} is the main effect for item i for the proficient proficiency status and represents the increase in the IRP when a student reaches the proficient proficiency status ($\alpha_{ca}^{1} = \alpha_{ca}^{2} = 1, \alpha_{ca}^{3} = \alpha_{ca}^{4} = 0$). The parameter λ_{i}^{3} is the main effect for item i for the distinguished proficiency status and represents the increase in the IRP when a student reaches the distinguished proficiency status ($\alpha_{ca}^{1} = \alpha_{ca}^{2} = \alpha_{ca}^{3} = 1, \alpha_{ca}^{4} = 0$). Finally, the parameter λ_{i}^{4} is the main effect for item i for the advanced proficiency status and represents the increase in the IRP when a student reaches the advanced proficiency status ($\alpha_{ca}^{1} = \alpha_{ca}^{2} = 1, \alpha_{ca}^{3} = \alpha_{ca}^{4} = 1$). Equation 15 has no interaction parameters because it only measures one attribute. See Bao (2019) for the parameterization of an item with more than one polytomous attribute with varying proficiency statuses across the attributes. Table 3 shows how the log odds of a correct response for an attribute with five proficiency statuses (the kernel in Equation 15) can be simplified according to the proficiency statuses.

As mentioned above, the structural component of the LCDM follows a Bernoulli distribution, but the structural model of the PDCM follows a categorical distribution because the attributes are polytomous. For the attribute with five proficiency statuses, there are five probabilities—one for each proficiency status for the attribute—that give the probability that the proficiency status equals each of the five proficiency statuses. These five probabilities must sum to one for each attribute. Note that the LCDM is a special case of the PDCM when $l_a = 2$ for all attributes.

The Polytomous Transition DCM (PTDCM)

The PTDCM combines the TDCM and the PDCM to get a model that has more fine-grained estimates of student growth because the transitions between testing occasions reflect multiple proficiency statuses for the polytomous attributes (Madison et al., 2021). This model is achieved by specifying the PDCM as the measurement model in the TDCM.

In the PTDCM framework, the total number of latent classes is computed as

$$C = \prod_{t=1}^{T} \prod_{a=1}^{A} l_{at} \tag{15}$$

where l_{at} is the number of proficiency statuses for attribute a at testing occasion t. The total number of transitions for attribute a across all testing occasions T is computed as

$$C'_a = \prod_{t=1}^{T} l_{at} \tag{16}$$

Although l_a does not need to be the same for all attributes, for this dissertation, it is assumed that the number of proficiency statuses for each attribute does not change over time. In this case, the computation for the total number of latent classes can be simplified to $C = \prod_{a=1}^{A} l_a^T$ and the number of transitions for attribute a can be simplified to $C'_a = l_a^T$.

Table 4 shows a summary of how to denote and compute the total number of latent classes and attribute-level transitions for different types of attributes, different numbers of testing occasions, and restrictions on the number of proficiency statuses for each attribute over time. Now that we have established background information for the SGP and for DCMs, the next chapter introduces the basic DGP and its extensions, as well as metrics for measuring the reliability of DGPs.

CHAPTER 3

DIAGNOSTIC GROWTH PERCENTILE

This chapter introduces the basic DGP, adjusted DGPs, and reliability metrics for student classification and DGPs that come from assessments that use the PTDCM as the measurement model. In this chapter, I aim to address the first research question and its sub-research questions presented in Chapter 1:

1. How do we conceptualize growth percentiles in a DCM framework?
    a. What are adjustments to DGPs to account for uncertainty in student classification from the DCM?
    b. How can DGPs be adjusted to consider penalties for forgetting?
    c. What are the implications of DGP interpretability for different adjusted DGPs?
    d. How can DGP reliability be conceptualized and computed?

In Chapters 4 and 5, I explain the studies I conducted to evaluate the DGP and its reliability. Specifically, in this chapter, I first introduce the DGP conceptually and situate it within the DCM framework by discussing the student-level results that are obtained from DCMs. Then, I explain the limitations in attempting to apply the quantile regression approach for computing SGPs to the diagnostic framework. Next, I introduce the basic DGP with examples that use different numbers of attributes, testing occasions, and proficiency statuses. To facilitate the introduction of the DGP, I also address the six critical questions for growth models posed by Castellano and Ho (2013a) and include a

discussion about the aggregation of DGPs across attributes and across students. Next, I introduce three adjusted DGP metrics to accommodate uncertainty in student classifications and penalties for forgetting or regressing over time. Then, I explain how I extended the reliability metrics from the TDCM framework (Madison, 2019; Schellman & Madison, in press) to the PTDCM framework and applied them to DGPs.

Building Blocks for the Diagnostic Growth Percentile

This section first describes the DCM-based student-level results that can be used to compute a DGP. Here, I explain why the quantile regression approach is not appropriate in the DCM framework, which is why the DGP metrics proposed in this study utilize the cohort approach, but the cohort approach is more appropriate in the DCM framework, anyway. This section introduces two DGP metrics, which are adapted from the cohort approach for computing SGPs: The first DGP metric (the *basic DGP*) directly mimics the cohort approach for computing SGPs, and the other DGP metric (the *adjusted DGP*) adjusts the basic DGP to account for uncertainty in the DCM classifications. The adjusted DGP can be computed in three different ways depending on how the users want to incorporate uncertainty and whether they want to penalize students who have a chance of being in a transition that shows forgetting (e.g., moving from proficient at T1 to non-proficient at T2). I introduce these metrics using the framework and critical questions from Castellano and Ho (2013a) and compare them with SGPs and categorical growth models. Finally, I present reliability metrics for the cohort-based DGPs.

Student-Level Results from DCMs

In the DCM framework, student results include three key components: class-level posterior probabilities, most likely classes (MLCs), and marginal posterior probabilities. The class-level and marginal posterior probabilities are continuous variables that range from 0 to 1. The MLC is a categorical variable. This section describes each of these components for the LCDM, TDCM, PDCM, and PTDCM. The next section describes how these components can be used to compute DGPs.

Student-Level Results from the LCDM

Consider a diagnostic assessment that measures two dichotomous attributes at one testing occasion, and we are interested in Student Z's results.

First, DCMs provide class-level posterior probabilities—probabilities that each student will be in each latent class given their observed response patterns. This diagnostic assessment has four latent classes that are characterized by the attribute profiles $\boldsymbol{\alpha_1} = [00]$, $\boldsymbol{\alpha_2} = [01]$, $\boldsymbol{\alpha_3} = [10]$, and $\boldsymbol{\alpha_4} = [11]$ where the first number shows the proficiency status for Attribute 1, and the second number shows the proficiency status for Attribute 2. The model will estimate four class-level posterior probabilities for Student Z.

Second, for each student, the greatest class-level posterior probability corresponds with the student's MLC. The attribute profile that corresponds with the MLC shows the student's strengths and weaknesses that can be used for differentiated classroom instruction. The most likely attribute profile is the key component that makes DCMs desirable in practice.

Third, the class-level posterior probabilities can be aggregated to determine the marginal probability of attribute proficiency for each attribute. Suppose we want to know

the probability that Student Z is proficient for Attribute 1. Then, we need to add Student Z's posterior probabilities for Classes 3 and 4 (the latent classes with the attribute profiles [10] and [11]—the latent classes that have proficiency for Attribute 1).

Student-Level Results from the TDCM

Now consider a diagnostic assessment that measures two dichotomous attributes at two testing occasions. The 16 TDCM classes for this example diagnostic assessment are the *transitions* between the LCDM latent classes are denoted $\boldsymbol{\alpha_1} = [0000]$, $\boldsymbol{\alpha_2} = [0001]$, $\boldsymbol{\alpha_3} = [0010]$, $\boldsymbol{\alpha_4} = [0011]$, $\boldsymbol{\alpha_5} = [0100]$, $\boldsymbol{\alpha_6} = [0101]$, $\boldsymbol{\alpha_7} = [0110]$, $\boldsymbol{\alpha_8} = [0111]$, $\boldsymbol{\alpha_9} = [1000]$, $\boldsymbol{\alpha_{10}} = [1001]$, $\boldsymbol{\alpha_{11}} = [1010]$, $\boldsymbol{\alpha_{12}} = [1011]$, $\boldsymbol{\alpha_{13}} = [1100]$, $\boldsymbol{\alpha_{14}} = [1101]$, $\boldsymbol{\alpha_{15}} = [1110]$, and $\boldsymbol{\alpha_{16}} = [1111]$ where the first number indicates the proficiency status for Attribute 1 at T1, the second number indicates the proficiency status for Attribute 2 at T1, the third number indicates the proficiency status for Attribute 1 at T2, and the fourth number indicates the proficiency status for Attribute 2 at T2. The first two numbers together indicate the MLC at T1. The last two numbers together indicate the MLC at T2. In general, the transitions show the full-crossing of the attribute profiles at each testing occasion. The transition with the greatest posterior probability is the student's most likely transition across both testing occasions. One can aggregate posterior probabilities across classes and/or across testing occasions to compute marginal attribute proficiency probabilities.

Student-Level Results from the PDCM

Now consider a diagnostic assessment that measures one polytomous attribute with three proficiency statuses ($\alpha_{c1} \in \{0,1,2\}$) and one polytomous attribute with four

proficiency statuses ($\alpha_{c2} \in \{0,1,2,3\}$) at one testing occasion, and we are interested in Student Z's results.

The PDCM estimates posterior probabilities of each student having each of the 12 attribute profiles $\boldsymbol{\alpha_1} = [00]$, $\boldsymbol{\alpha_2} = [01]$, $\boldsymbol{\alpha_3} = [02]$, $\boldsymbol{\alpha_4} = [03]$, $\boldsymbol{\alpha_5} = [10]$, $\boldsymbol{\alpha_6} = [11]$, $\boldsymbol{\alpha_7} = [12]$, $\boldsymbol{\alpha_8} = [13]$, $\boldsymbol{\alpha_9} = [20]$, $\boldsymbol{\alpha_{10}} = [21]$, $\boldsymbol{\alpha_{11}} = [22]$, and $\boldsymbol{\alpha_{12}} = [23]$ where the first number shows the proficiency status for the attribute with three proficiency statuses, and the second number shows the proficiency status for the attribute with four proficiency statuses. The latent class with the greatest posterior probability is the student's MLC.

The class-level posterior probabilities can be summed to compute the marginal probability of a student reaching a certain proficiency status for a given attribute. For example, suppose we are interested in the probability that Student Z has reached the second proficiency status for the first attribute ($\alpha_{c1} = 1$). Then, we need to sum the class-level posterior probabilities for the latent classes with the profiles [10], [11], [12], and [13] (latent classes 5, 6, 7, and 8—the classes with 1 as the first number in the attribute profile).

Student-Level Results from the PTDCM

Finally, consider a diagnostic assessment that measures one polytomous attribute with five proficiency statuses ($\alpha_{ca} \in \{0,1,2,3,4\}$) at two testing occasions. The PTDCM estimates posterior probabilities of each student having each of the 25 attribute profiles/transitions shown in Table 5.

The transition that has the greatest posterior probability is the student's most likely transition.

Aggregating posterior probabilities in the PTDCM framework can be tricky because one must consider which proficiency status and which testing occasion they want to consider.

Now that we have established the student-level variables in the DCM framework, we can consider how to use them to obtain diagnostic versions of the SGP. For the next section, I am considering longitudinal data with two testing occasions.

Quantile Regression Approach for DGPs

The quantile regression approach for computing DGPs can be applied when either the PDCM is estimated separately for each testing occasion or when the PTDCM is estimated for both testing occasions simultaneously. Quantile regression requires that the dependent variable is continuous, but the independent variable(s) can be continuous and/or categorical. For both modeling approaches, the only continuous DCM-based student result is a posterior probability. Therefore, if I adapt the quantile regression approach for the DCM framework, the dependent variable would have to be a posterior probability associated with T2. The independent variable(s) can be a posterior probability associated with T1 and/or the MLC from T1.

The quantile regression approach has one primary advantage: flexibility in variable selection, and three primary disadvantages: complexity in calculation, difficulty in interpreting DGPs given the flexibility in variable selection, and the questionable appropriateness of using posterior probabilities for DGP calculation.

Flexibility

As mentioned above, because the PDCM and PTDCM estimate a posterior probability for each class and marginal posterior probabilities can be computed in many

ways, we have many options for how to use posterior probabilities as independent and dependent variables.

For example, suppose that we estimate the PDCM separately for each testing occasion, and it is of interest to estimate DGPs with respect to Level 5 of the attribute. Then, the dependent variable would be the posterior probability for Class 5 ($\boldsymbol{\alpha_5} = [4]$) at T2. The independent variable could be the most likely attribute proficiency status from T1, the posterior probability for Class 5 from T1, or any other posterior probability from T1. The choice of which independent variable to use depends on the desired interpretations of the DGPs. Note that one could choose to focus on any of the proficiency statuses instead of Class 5.

Suppose instead that we estimate the PDCM separately for each testing occasion, and it is of interest to estimate DGPs with respect to the *overall* proficiency of the attribute. The attribute has five proficiency statuses, so how does one determine how to sum the posterior probabilities to obtain a likelihood of overall proficiency? One option might be to sum the posterior probabilities for Class 4 (with $\boldsymbol{\alpha_4} = [3]$ and the label *distinguished*) and Class 5. Another option might be to estimate the LCDM for a single dichotomous attribute at each testing occasion. The LCDM will yield a marginal probability of overall attribute proficiency, but this approach would dichotomize the attribute with five proficiency statuses, which may be inappropriate for the attribute and result in an unacceptable loss of information.

Complex Computation

As described above, the computation of quantile regression-based SGPs is complex. Fortunately, one does not need to do this computation manually or from scratch

because the SGP package in R (Betebenner et al., 2022) is available to estimate SGPs given student-level results for at least two testing occasions. This package can be used to compute quantile regression-based DGPs using the chosen independent and dependent variables, as described in the previous section.

However, many of the functions in the package, including the function that computes confidence intervals for the SGP estimates, depend on past results and specifications from states that use SGPs in practice. These state-specific results and specifications were derived from scaled scores on state assessments; therefore, many of the package's functions are unavailable or inappropriate for applications in diagnostic assessment.

Additionally, the package fails to estimate quantile regression-based DGPs for some combinations of independent and dependent variables. One factor that might impact whether the package is successfully able to estimate DGPs is the variability of the data. For example, if a posterior probability includes many probabilities that are close to zero, the data might not have enough variability to estimate the DGPs. With a large number of latent classes, it is possible that many posterior probabilities would be near zero depending on the sample size and distribution of proficiency statuses across the sample.

Thus, the flexibility in variable selection is limited by the package's current structure and by the variability of the data.

Difficult to Interpret

The flexibility in variable selection comes at a cost in terms of interpretability. Consider the example where we estimate the PDCM separately for each testing occasion, and it is of interest to estimate DGPs with respect to Class 5. We can choose the posterior

probability of Class 5 from T1 as the independent variable and the posterior probability of Class 5 from T2 as the dependent variable. Then, the DGP indicates the likelihood of observing the current Class 5 posterior probability given the previous Class 5 posterior probability, but what does that mean when we think about growth? A student's posterior probability for Class 5 is not their only result from the PDCM (i.e., they also have posterior probabilities for each other latent class), and it is possible that the student's MLC is not Class 5, so it may not be appropriate to use their Class 5 posterior probability as one of the variables in the DGP.

Having multiple choices for the independent and dependent variables for each student adds a layer of complication in determining the intended use of the DGP and how to interpret it.

Questionable Appropriateness

One must consider whether it is appropriate to use posterior probabilities to estimate DGPs. Posterior probabilities are not the student results that DCMs are designed to report and produce. Students should not be ranked by their posterior probabilities, and even though the quantile regression-based DGP does not directly rank-order students, using posterior probabilities for a normative metric such as the DGP is an inappropriate application.

The Basic Diagnostic Growth Percentile

Because of the disadvantages of the quantile regression-based DGP metric described above and because cohort-based SGPs are often used in practice (Michigan Department of Education, n.d.), this chapter presents the cohort-based approach for computing a DGP. The transitions resulting from longitudinal DCMs provide cohorts as

part of model estimation, so the cohort approach is an intuitive choice because it aligns with the DCM framework.

The basic DGP is computed for each attribute in the assessment. A student's cohort for attribute a in the DCM framework is the group of students who have the same attribute a proficiency diagnoses at all testing occasions. A student's basic DGP for a single attribute is given by the cumulative proportion of students from the previous latent class for the attribute who are at least in the current latent class for the attribute. The basic DGP is equivalent for all students within a cohort. Although the SGP is typically expressed on the 0 to 100 scale, the 0 to 1 scale is most appropriate for the DGP because it aligns more closely with typical DCM literature.

The Basic DGP for a Dichotomous Attribute

For example, consider a dichotomous attribute measured at two testing occasions. In this situation, there are four cohorts: (1) those students who were non-proficient at T1 and non-proficient at T2 (students with the transition [00] for the attribute where the first value indicates the proficiency status at T1, and the second value indicates the proficiency status at T2), (2) those students who were non-proficient at T1 and proficient at T2 (students with the transition [01] for attribute a), (3) those students who were proficient at T1 and non-proficient at T2 (students with the transition [10] for attribute a), and (4) those students who were proficient at T1 and proficient at T2 (students with the transition [11] for attribute a). Suppose that 60% of students were non-proficient and 40% of the students were proficient at T1. Further, suppose that at T2, 46% of the students who were non-proficient at T1 are still non-proficient at T2 (transition [00]), while the remaining 54% of the students who were non-proficient at T1 are now

proficient at T2 (transition [01]), and 6% of the students who were proficient at T1 are now non-proficient at T2 (transition [10]), while the remaining 94% of the students who were proficient at T1 are still proficient at T2 (transition [11]).

Then all students with the transition [00] have a basic DGP of 46 because they showed growth greater than or equal to (they are all "equal to") 46% of the students who were non-proficient at T1. All students with the transition [01] have a basic DGP of 100 because they showed growth greater than or equal to 100% of the students who were non-proficient at T1. All students with the transition [10] have a basic DGP of 6 because they showed "growth" (the transition [10] shows forgetting, so calling this "growth" is counterintuitive) greater than or equal to 6% of the students who were proficient at T1. All students with the transition [11] have a basic DGP of 100 because they showed growth greater than or equal to 100% of the students who were proficient at T1.

More generally, for a dichotomous attribute a with proficiency status labels of "non-proficient" (indicated by a 0 in the transition profile) and "proficient" (indicated by a 1 in the transition profile), there are four cohorts/transitions, denoted by $c' \in \{[00], [01], [10], [11]\}$ or, equivalently, $c' \in \{1, 2, 3, 4\}$. I chose to use c' to denote attribute-level transitions because they are not always the same as the latent classes—they are often subsets of latent classes, so it is important to keep the index for latent classes (c) and the index for attribute-level transitions (c') distinct. Each of these cohorts has its own basic DGP, denoted $D_{B_{a,c'}}$, where "D" represents "DGP" and "B" represents "basic". For a dichotomous attribute, the basic DGP for attribute a for the transition [00], $D_{B_{a,00}}$, is:

$$D_{B_{a,00}} = \frac{p_{a,00}}{p_{a,00} + p_{a,01}} \tag{17}$$

where α_1 is the proficiency status at T1 and α_2 is the proficiency status at T2. $p_{a,c'}$ is the proportion of all students who have the transition c' for attribute a. $p_{a,c'}$ is also known as the *attribute transition base rate* or *transition probability*, and it will be important later when I discuss reliability for DGPs. In other words, the sum of $p_{a,c'}$ across all values of c' is 1 because all students have a transition for attribute a. For this example of a dichotomous attribute measured at two testing occasions, $p_{a,1} + p_{a,2} + p_{a,3} + p_{a,4} = p_{a,00} + p_{a,01} + p_{a,10} + p_{a,11} = 1$.

$D_{B_{a,00}}$ can be interpreted as the proportion of students who were non-proficient at T1 who are at least non-proficient at T2. The "at least" part of the interpretation does not mean much for the [00] transition because 0 is the lowest proficiency status. "At least" makes more sense for the interpretation of DGPs for transitions with T2 proficiency statuses that are greater than the lowest proficiency status. For example, the basic DGP for Attribute a for the cohort [01], $D_{B_{a,01}}$, is:

$$D_{B_{a,01}} = \frac{p_{a,00} + p_{a,01}}{p_{a,00} + p_{a,01}} \tag{18}$$

Then, $D_{B_{a,01}}$ can be interpreted as the proportion of students who were non-proficient at T1 who are at least proficient at T2. Here "at least" means a little more than it did for the [00] transition because a diagnosis of proficient at T2 is not the lowest proficiency status.

Likewise, the other two basic DGPs for Attribute a are:

$$D_{B_{a,10}} = \frac{p_{a,10}}{p_{a,10} + p_{a,11}} \tag{19}$$

$$D_{B_{a,11}} = \frac{p_{a,10} + p_{a,11}}{p_{a,10} + p_{a,11}} \tag{20}$$

These DGPs can be interpreted similarly as above. We can see from this dichotomous attribute example that when a student is diagnosed as having the greatest proficiency status at T2, their basic DGP is always 1 no matter what their proficiency status was at T1.

The Basic DGP for a Trichotomous Attribute

Now, let us consider a trichotomous attribute a with proficiency status labels of "non-proficient" (indicated by a 0 in the transition profile), "partially proficient" (indicated by a 1 in the transition profile), and "proficient" (indicated by a 2 in the transition profile). For Attribute a, there are nine cohorts/transitions, denoted by $c' \in \{[00], [01], [02], [10], [11], [12], [20], [21], [22]\}$. Each of these cohorts has its own basic DGP, denoted $D_{B_{a,c'}}$, which are given in Table 6. In the trichotomous case, we can see that "at least" has more meaning than it did in the dichotomous case because the proficiency status has more levels. Again, we can see from the trichotomous attribute example that when a student is diagnosed as having the greatest proficiency status at T2, their basic DGP is still always 1 no matter what their proficiency status was at T1 because they have achieved the maximum amount of growth in attribute proficiency. The basic DGP for attributes with any number of proficiency statuses can be computed similarly.

The Basic DGP for More Than Two Testing Occasions

Additionally, more than two testing occasions can be incorporated—the only difference is that the basic DGP can be computed multiple different ways, depending on which testing occasion(s) you want to consider. For an attribute measured at three testing occasions, you can compute the basic DGP in three different ways: by comparing T1 and

T2, T2 and T3, or T1 and T3. In general, comparing adjacent testing occasions among multiple testing occasions yields DGPs analogous to those described above for two testing occasions. Therefore, this section considers comparisons between testing occasions that are not adjacent.

Let us consider the dichotomous attribute a again. With three testing occasions, Attribute a has eight transitions, denoted by $c' \in$ $\{[000], [001], [010], [011], [100], [101], [110], [111]\}$, where the first value indicates the proficiency status for T1, the second value indicates the proficiency status for T2, and the third value indicates the proficiency status for T3. When comparing T1 and T2, the computation is the same as what was previously described except that now $p_{a,00}$ includes all students with either of these transitions: [000] or [001], $p_{a,01}$ includes all students with either of these transitions: [010] or [011], $p_{a,10}$ includes all students with either of these transitions: [100] or [101], and $p_{a,11}$ includes all students with either of these transitions: [110] or [111]. More generically, when comparing T1 and T2, $p_{a,00}$ includes all students with any transition that matches [00*], where * indicates that the proficiency status for T3 can take any value (i.e., 0 or 1), $p_{a,01}$ includes all students with the transition [01*], $p_{a,10}$ includes all students with the transition [10*], and $p_{a,11}$ includes all students with the transition [11*]. Similarly, when comparing T2 and T3, $p_{a,00}$ includes all students with the transition [*00], $p_{a,01}$ includes all students with the transition [*01], $p_{a,10}$ includes all students with the transition [*10], and $p_{a,11}$ includes all students with the transition [*11].

When comparing T1 and T3, there are two options: (1) use the same approach as when comparing T1 and T2 or T2 and T3, or (2) use the full transition to indicate the

cohort, as opposed to using only two of the three testing occasions to indicate the cohort. For approach (1), $p_{a,00}$ includes all students with the transition [0*0], $p_{a,01}$ includes all students with the transition [0*1], $p_{a,10}$ includes all students with the transition [1*0], and $p_{a,11}$ includes all students with the transition [1*1]. For approach (2), each of the eight transitions has its own basic DGP. Table 7 shows how to compute the eight basic DGPs for the situation with a dichotomous attribute measured at three testing occasions. The difference with approach (2) is that the cohort is conditional on the proficiency statuses for T1 and T2 rather than just one or the other, as shown in Table 7. The computation can be extended to polytomous attributes and more than three testing occasions using steps analogous to those described above.

The General Form of the Basic DGP

In general, the basic DGP is computed as a quotient of sums:

$$\boldsymbol{D}_{\boldsymbol{B_a},[\alpha_1\,\alpha_2\,\ldots\,\alpha_{T-1}\,\alpha_{T'}]} = \frac{\sum_{\alpha_{T''} \le \alpha_{T'}} \boldsymbol{p}_{\boldsymbol{a},[\alpha_1\,\alpha_2\,\ldots\,\alpha_{T-1}\alpha_{T''}]}}{\sum_{\alpha_{T'''}=\boldsymbol{0}}^{\boldsymbol{l_a-1}} \boldsymbol{p}_{\boldsymbol{a},[\alpha_1\,\alpha_2\,\ldots\,\alpha_{T-1}\alpha_{T'''}]}} \tag{21}$$

where $\boldsymbol{D}_{\boldsymbol{B_a},[\alpha_1\,\alpha_2\,\ldots\,\alpha_{T-1}\alpha_{T'}]}$ is the basic DGP for attribute $\boldsymbol{a}$ for attribute transition $\boldsymbol{c'} = [\boldsymbol{\alpha_1\,\alpha_2}\ \ldots\ \boldsymbol{\alpha_{T-1}\alpha_{T'}}]$, $\boldsymbol{\alpha_t}$ is the proficiency status for attribute $\boldsymbol{a}$ for attribute transition $\boldsymbol{c'}$ at testing occasion $\boldsymbol{t}$, $\boldsymbol{p}_{\boldsymbol{a},[\alpha_1\,\alpha_2\,\ldots\,\alpha_{T-1}\alpha_{T''}]}$ is the posterior probability for attribute $\boldsymbol{a}$ for attribute transition $\boldsymbol{c''} = [\boldsymbol{\alpha_1\,\alpha_2}\ \ldots\ \boldsymbol{\alpha_{T-1}\alpha_{T''}}]$, $\boldsymbol{p}_{\boldsymbol{a},[\alpha_1\,\alpha_2\,\ldots\,\alpha_{T-1}\alpha_{T'''}]}$ is the posterior probability for attribute $\boldsymbol{a}$ for attribute transition $\boldsymbol{c'''} = [\boldsymbol{\alpha_1\,\alpha_2}\ \ldots\ \boldsymbol{\alpha_{T-1}\alpha_{T'''}}]$, and $\boldsymbol{l_a}$ is the number of proficiency statuses for attribute $\boldsymbol{a}$ (this equation assumes that $\boldsymbol{l_a}$ is fixed across testing occasions). Equation 22 uses $\boldsymbol{\alpha_{T'}}$, $\boldsymbol{\alpha_{T''}}$, and $\boldsymbol{\alpha_{T'''}}$ to denote that the proficiency statuses for the last testing occasion are not the same for some attribute

transitions $\boldsymbol{c'}$, $\boldsymbol{c''}$, and $\boldsymbol{c'''}$. The numerator requires that the $\boldsymbol{T}$th proficiency statuses for the attribute transitions included in the numerator sum ($\boldsymbol{\alpha_{T''}}$) are less than or equal to the $\boldsymbol{T}$th proficiency status for the attribute transition for the DGP of interest ($\boldsymbol{\alpha_{T'}}$) because the basic DGP is the cumulative proportion of attribute transition probabilities up to the proficiency status at Testing Occasion T, conditional on the proficiency statuses from the previous testing occasions. Equation 22 uses $\boldsymbol{l_a - 1}$ because [0] is the first proficiency status for all attributes. Additionally, $\boldsymbol{c'} \in \{\boldsymbol{1}, \ldots, \boldsymbol{C'}\}$, $\boldsymbol{c''} \in \{\boldsymbol{1}, \ldots, \boldsymbol{C'}\}$, and $\boldsymbol{c'''} \in \{\boldsymbol{1}, \ldots, \boldsymbol{C'}\}$ where $\boldsymbol{C'}$ is the total number of transitions for attribute $\boldsymbol{a}$ and

$$\mathbf{1} = \sum_{\boldsymbol{c'}=\mathbf{1}}^{\boldsymbol{C'}} \boldsymbol{p_{a,c'}} \tag{22}$$

I want to highlight a nuanced part of the notation for the basic DGP: $D_{B_a,[\alpha_1\ \alpha_2\ \ldots\ \alpha_{T-1}\ \alpha_{T'}]}$ is not subscripted with r even though it is a student-level variable because is only unique for each transition, not for each students, so it needs to be indexed only by the transition, not by specific students.

The Basic DGP for More Than One Attribute

As described above, the basic DGP is first computed separately for each student and each attribute. However, if the longitudinal diagnostic assessment measures more than one attribute, users may be interested in an overall basic DGP for each student. To aggregate the attribute-level basic DGP to obtain an overall profile-level basic DGP for the student, one can compute the average or median of their attribute-level basic DGPs.

First, each student r is assigned the attribute-level DGP that corresponds with their cohort for each attribute a, $D_{B_{r,a}}$. In general, if student r has the transition c' for attribute a,

$$D_{B_{r,a}} = D_{B_{a,c'}} \tag{23}$$

Student r's overall profile-level basic DGP for an assessment with A attributes, MD_{B_r}, is given by

$$MD_{B_r} = \frac{D_{B_{r,1}} + D_{B_{r,2}} + \cdots + D_{B_{r,A}}}{A} \tag{24}$$

Student r's median profile-level basic DGP for an assessment with A attributes, $MdnD_{B_r}$, is given by

$$MdnD_{B_r} = Mdn\left(D_{B_{r,1}}, D_{B_{r,2}}, \ldots, D_{B_{r,A}}\right) \tag{25}$$

A student's overall profile-level basic DGP of 88, for example, can be interpreted as such: Averaging over all measured attributes, the student showed growth that was greater than or equal to 88% of the students who had the same proficiency statuses on the attributes (i.e., same attribute profile) at earlier testing occasions. In other words, for the overall profile-level basic DGP, students' cohorts are made up of students who have the same attribute profiles across all previous testing occasions. The median basic DGP can be similarly interpreted.

Answers to the Critical Questions for the Basic DGP

Castellano and Ho (2013a) presented a thorough guide for classifying growth models so that all critical considerations are clearly communicated to potential users. This section describes the answers to these critical questions for the basic DGP while comparing the basic DGP with SGPs and categorical growth models as they are presented in Castellano and Ho (2013a).

Categorical growth models measure growth in terms of transitions between achievement levels (e.g., beginning, developing, proficient, and distinguished) that are

used to classify students' summative scaled scores for federal categorical reporting. These achievement levels are determined through a standard setting process in which experts consider which ranges of scaled scores most align with students who are typical of each achievement level group. In the categorical growth framework, growth is described as movement between achievement levels over time. In this way, the basic DGP is similar to categorical growth models. The basic DGP and the categorical growth model both utilize transition matrices, which show the proportion of students who transition from one category (or latent class, in the DCM framework) to another. However, the basic DGP provides an SGP-like metric based on a conditional status rather than a raw gain—in the case of categorical growth models, the gain is the change in the achievement level rather than the change in a scaled score (Castellano & Ho, 2013a).

Now that I have established the basic concepts of the categorical growth model, I can address the six critical questions for the basic DGP.

Question 1

"What primary interpretation does the growth model best support?" (Castellano & Ho, 2013a, p. 18). To address the first critical question, like the SGP, the basic DGP best supports a description of growth because it allows for interpretations related to relative growth. The basic DGP is not primarily focused on predicting future growth or attempting to figure out what caused the growth, as in value-added models. Alternatively, the categorical growth model is often used to project growth and determine whether students are generally on track to proficiency (Castellano & Ho, 2013a).

The appropriate interpretation for each attribute-level basic DGP with a value of X is that the student showed growth greater than or equal to X% of the students who had the same proficiency status for the attribute at previous testing occasions.

Question 2

"What is the statistical foundation underlying the growth model?" (Castellano & Ho, 2013a, p. 20). From a DCM perspective, the statistical foundation for the DGP is the PTDCM because the DGP is directly derived from the transitions yielded from the PTDCM. From the perspective of the Castellano and Ho (2013a) framework and to align the DGP with the SGP, the basic DGP has a statistical foundation of a conditional status model because the basic DGP is a metric for comparing students' growth with that of their expected growth given their past performance and in relation to their academic peers. The basic DGP is conditional on students' past proficiency status(es). The categorical growth model, on the other hand, is a gain-based model that depends heavily on the specification of achievement levels across grades, which implies a vertical scale, though a vertical scale is not explicitly required. Thus, the categorical growth model does not give a relative measure for students' growth.

Question 3

"What are the required data features for this growth model?" (Castellano & Ho, 2013a, p. 23). To address this third critical question, this section discusses vertical scales, standard setting, the number of proficiency statuses, the sample size requirements, and the longitudinal requirements.

Although vertical scaling is not required for the basic DGP (vertical scales are also not required for the SGP or categorical growth models), it is recommended that for

sensible interpretation of the basic DGP, the same attributes are measured at all testing occasions because "growth" is difficult to interpret when different attributes are measured at different testing occasions. However, it is not required that all testing occasions utilize the exact same assessment items because students' cohorts can be based on any previous attribute profiles using any item and any attributes. For example, a pre-test and a post-test need not have the same items, but the student classifications resulting from each testing occasion can be used to compute students' DGPs.

Because the basic DGP is in the DCM framework, standard setting and cut scores are not required because the attribute proficiency statuses need not be aligned with an underlying continuum of ability or proficiency (Castellano & Ho, 2013a). The DCM uses a statistical approach to "cut" the probability scale to determine which students are more or less likely to be proficient for each attribute. This removal of the somewhat subjective standard setting process is one desirable feature of DCMs.

Just like with the categorical growth model (Castellano & Ho, 2013a), the number of proficiency statuses an attribute has can greatly change the interpretation and utility of the basic DGP. Increasing the number of proficiency statuses for an attribute can add to the interpretive value and utility of the DGPs, but I do not recommend using more than five or seven proficiency statuses because after seven categories, interpretation of the results is tricky.

The basic DGP requires estimation of the TDCM for dichotomous attributes or the PTDCM for polytomous attributes. The TDCM and PTDCM are saturated DCMs that include many parameters and, therefore, have sample size requirements to achieve stable parameter estimation. In comparison to the SGP, which recommends at least 5,000

students (Castellano & Ho, 2013a), the TDCM and PDCM are flexible models that can be estimated with as few as 1,000 students (Bao, 2019; Madison & Bradshaw, 2018). The PTDCM is a more complex model than the TDCM and PTDCM, so it can be expected that the sample size requirements for the PTDCM will be at least as large as either the TDCM or the PDCM.

The basic DGP requires TDCM or PTDCM data from at least two testing occasions, but it can support any number of testing occasions. However, as the number of testing occasions increases, the number of ways to compute the basic DGP increases and could become unwieldy. Those who choose to compute the basic DGP for multiple testing occasions must carefully consider which comparison(s) is(are) most important for their situation.

Question 4

"What kinds of group-level interpretations can this growth model support?" (Castellano & Ho, 2013a, p. 26). Similar to how one can compute the average or median of the attribute-level basic DGPs to obtain an overall profile-level basic DGP for each student, one can compute the average or median basic DGP across all students who are in the same class, school, district, or state. The basic DGP can be aggregated for each attribute (i.e., the mean/median basic DGP for attribute a across all students) separately or for all attributes together (i.e., the mean/median overall basic DGP for all attributes across all students).

The mean profile-level basic DGP across N students, MD_B, is given by

$$MD_B = \frac{\sum_{r=1}^{N} MD_{B_r}}{N} \tag{26}$$

The median profile-level basic DGP across N students, $MdnD_B$, is given by

$$MdnD_B = Mdn\left(MD_{B_1}, MD_{B_2}, \ldots, MD_{B_N}\right) \quad (27)$$

The mean basic DGP across N students for attribute a, MD_{B_a}, is given by

$$MD_{B_a} = \frac{\sum_{r=1}^{N} D_{B_{r,a}}}{N} \quad (28)$$

The median basic DGP across N students for attribute a, $MdnD_{B_a}$, is given by

$$MdnD_{B_a} = Mdn\left(D_{B_{1,a}}, D_{B_{2,a}}, \ldots, D_{B_{N,a}}\right) \quad (29)$$

The interpretation for the mean basic DGP (attribute-level or overall) of Y is that, on average, students at School A showed growth that was greater than or equal to Y% of the students who had the same proficiency statuses at previous testing occasions.

For reference, in practice, SGPs are most often aggregated via the median rather than the mean SGP because percentiles generally should not be averaged (Betebenner, 2009), but some practitioners use the average with favorable results (Castellano & Ho, 2013a, 2013b).

Question 5

"How does the growth model set standards for expected or adequate growth?" (Castellano & Ho, 2013a, p. 26). DCMs are better situated to model transitions between categorical student classifications than IRT models that use standard setting to obtain achievement levels and then use categorical growth modeling to describe student growth. In this way, the DGP is superior to the categorical growth models because it does not require standard setting and is not plagued with the potential issues of subjectivity that occur with standard setting. However, DCMs are occasionally criticized for not requiring standard setting because, without standard setting, there is less oversight into how the model is classifying students.

Although the TDCM and PTDCM do not require standard setting to make classifications, just like with SGPs and categorical growth models, expert judgement is needed to determine which values or ranges of the basic DGP can be considered "good," "adequate," "poor," or any other subjective qualification for the relative growth. Expert judgement is also needed to interpret the proficiency statuses for each attribute to determine appropriate status labels given the target population and the nature of the attributes.

Question 6

"What are the common misinterpretations of this growth model and possible unintended consequences of its use in accountability systems?" (Castellano & Ho, 2013a, p. 27). The primary misinterpretation of the basic DGP that I can foresee is that it might be misinterpreted as a percentile because the word "percentile" is in the name of the metric. This confusion with nomenclature has plagued SGPs as well, but because the DGP is adapted from the SGP, using the same nomenclature promotes continuity, which would be lost if I adopted another name.

The Adjusted Diagnostic Growth Percentiles

The basic DGP utilizes the raw proportions of students who are classified as most likely belonging to each transition. However, as with any measurement model, DCM classifications contain a degree of uncertainty such that a student's MLC might not be their true class. As described previously, longitudinal DCMs estimate the probability that each student belongs to each transition. Unless the greatest probability (the probability for the student's MLC) is 1, there is a non-zero chance that the student's true transition is different from their most likely transition. For example, consider one dichotomous

attribute measured at two testing occasions. Suppose Student X has a maximum posterior probability of .85 for the transition [01] and posterior probabilities of .05 for each of the other transitions [00], [10], and [11], and Student Y has a maximum posterior probability of .31 for the transition [01] and posterior probabilities of .23 for each of the other transitions [00], [10], and [11]. Both Student X and Student Y would have basic DGPs equal to 1 because they have the maximum proficiency status at the last testing occasion, but it does not make sense intuitively for these students to have the same DGPs because the model is much more certain about Student X's classification than Student Y's classification, and it is possible that Student Y's true transition is something other than [01].

To account for the uncertainty in student classifications, I propose an adjusted DGP that incorporates the student's individual posterior probabilities and the basic DGPs for all transitions. For each student r and each attribute a, the *adjusted DGP with complete weighting* (CW DGP), denoted $D_{CW_{r,a}}$, is computed as

$$D_{CW_{r,a}} = \sum_{c'=1}^{l_a^T} \alpha_{c'r} * D_{B_{a,c'}} \tag{30}$$

where l_a^T is the total number of transitions given by l_a proficiency statuses and T testing occasions for attribute a, $\alpha_{c'r}$ borrows notation from Rupp and colleagues (2010, p. 239) to indicate the marginal posterior probability that student r is in the attribute-level transition c' for attribute a, and $D_{B_{a,c'}}$ is the basic DGP for attribute a and with the attribute-level transition c'. The posterior probability is marginal because, depending on the number of attributes measured over time, the necessary posterior probability for the CW DGP might require summing the class-level posterior probabilities across the classes

that include the attribute-level transition. Thus, a student's attribute-level CW DGP is the sum of products of the student's marginal posterior probabilities for each transition and the basic DGP for each corresponding transition. Because of the relative weighting, the CW DGP has more variability than the basic DGP, which is restricted to only a few points on the 0 to 1 scale. To illustrate how to interpret the CW DGP, consider Student X, who has a CW DGP value of 78.4. We can interpret this value using a statement like, "Student X showed growth greater than or equal to 78.4% of the students who have the same classification histories when considering all of Student X's possible score histories."

Let us consider another situation with one dichotomous attribute measured at two testing occasions: Suppose Student A and Student B both have maximum posterior probabilities of .60 for the transition [01]. However, Student A's posterior probabilities for the other transitions are .10 for [00], .20 for [10], and .10 for [11], and Student B's posterior probabilities for the other transitions are .175 for [00], .05 for [10], and .175 for [11]. Student A's chance of having the forgetting transition [10] is much larger than Student B's chance of having the forgetting transition. The basic DGP and CW DGP do not directly penalize students for having non-zero probabilities of belonging to a transition that includes forgetting (e.g., transitions in which the students' proficiency statuses decrease between testing occasions). Categorical growth models use a value table to allow for weighting the transitions between the achievement levels, and this weighting process allows users to penalize students who have transitions that show forgetting or who regress from one testing occasion to the next (Castellano & Ho, 2013a). The DGP can similarly be adjusted to account for forgetting. I propose two approaches: (1) the

adjusted DGP with complete weighting and a penalty for forgetting (CWP DGP) and (2) the adjusted DGP with partial weighting (PW DGP).

The CWP DGP, denoted $D_{CWP_{r,a}}$, is computed the same way as the CW DGP, except that the terms in the sum that correspond with the transitions that include forgetting are each multiplied by a specific coefficient to further weight the DGP. Thus,

$$D_{CWP_{r,a}} = \sum_{c'=1}^{l_a^T} f * \alpha_{c'r} * D_{B_{a,c'}} \tag{31}$$

where

$$f = \begin{cases} -\dfrac{z}{T-1} \text{ if } c' = [\alpha_1\alpha_2 \dots \alpha_T] \text{ where } \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_T \text{ is not true} \\ 1 \text{ if } c' = [\alpha_1\alpha_2 \dots \alpha_T] \text{ where } \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_T \text{ is true} \end{cases} \tag{32}$$

where α_1 is the proficiency status for Attribute a at T1, α_2 is the proficiency status for Attribute a at T2, α_T is the proficiency status for Attribute a at testing occasion T, and z is the number of adjacent testing occasions t and $t-1$ within the transition $[\alpha_1\alpha_2 \dots \alpha_T]$ for which $\alpha_{t-1} > \alpha_t$. When $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_T$ is not true, the coefficient f ranges from $-1 \leq f < 0$ and $f = -1$ when the transition shows forgetting for each pair of adjacent testing occasions. For example, for an attribute with five proficiency statuses (i.e., $\alpha_{ac} \in \{0,1,2,3,4\}$) measured at three testing occasions, transitions such as [210], [310], and [421] show forgetting for each pair of adjacent testing occasions because the proficiency statuses within the transition vectors are strictly decreasing over time. The coefficient is between -1 and 0 when the transition shows forgetting for at least one pair but not all pairs of adjacent testing occasions. For example, for an attribute with five proficiency statuses measured at three testing occasions, transitions such as [213], [021], and [322] show forgetting for some, but not all, pairs of adjacent testing occasions because the

proficiency statuses within the transition vectors are not strictly decreasing over time but have some cases of decreasing.

The PW DGP, denoted $D_{PW_{r,a}}$, is computed the same way as the CWP DGP, except that the terms in the sum that correspond with the transitions that include forgetting are each multiplied by a new coefficient that gives partial credit for transitions that include forgetting. Thus,

$$D_{PW_{r,a}} = \sum_{c'=1}^{l_a^T} f' * \alpha_{c'r} * D_{B_{a,c'}} \tag{33}$$

where

$$f' = \begin{cases} \dfrac{T-1-z}{T-1} \text{ if } c' = [\alpha_1 \alpha_2 \ldots \alpha_T] \text{ where } \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_T \text{ is not true} \\ 1 \text{ if } c' = [\alpha_1 \alpha_2 \ldots \alpha_T] \text{ where } \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_T \text{ is true} \end{cases} \tag{34}$$

To illustrate how to interpret the CWP DGP and the PW DGP, consider Student X, who has a CWP DGP (or a PW DGP) value of 55.9. We can interpret this value using a statement like, “Student X showed growth greater than or equal to 55.9% of the students who had the same classification histories when considering all of Student X’s possible score histories and with a penalty for all possible occurrences of forgetting.”

The empirical data analyses in this study illustrate how the basic and adjusted DGPs compare. The answers to Castellano and Ho’s (2013a) six critical questions to the basic DGP also apply for the adjusted DGP metrics.

Reliability of Diagnostic Growth Percentiles

It is important to have a measure of reliability for DGPs so we know how stable the estimates are across testing occasions and across levels of the attribute (Bao, 2019). For the rest of this dissertation, when I refer to “reliability,” I am referring to the

reliability of student classification from DCMs, which is also the reliability of DGP estimates. The literature includes a method for measuring the reliability of quantile regression-based SGPs (Monroe & Cai, 2015) that could, in theory, be adapted to the quantile regression-based DGPs. However, the SGP reliability metric requires a measure of the standard error of the SGPs, and as mentioned above, the SGP package in R cannot compute the standard error (via confidence intervals) without having specific information about the state, which is based on scaled scores and cannot be used in the DCM framework. The cohort-based DGPs are directly derived from the latent class transition probabilities. Therefore, reliability metrics for the latent class transition probabilities under the PTDCM can be applied as reliability metrics for the cohort-based DGP.

To give a quick snapshot of the history of reliability metrics for the DCMs discussed in this chapter, Templin and Bradshaw (2013) developed a test-retest-type metric for measuring reliability with the LCDM. Madison (2019) extended the Templin and Bradshaw (2013) metric to use dichotomous attributes at multiple testing occasions (i.e., the TDCM). Bao (2019) extended the Templin and Bradshaw (2013) metric to use polytomous attributes at a single testing occasion (i.e., the PDCM). Madison and colleagues (2021) extended the Madison (2019) metric to use polytomous attributes at multiple testing occasions (i.e., the PTDCM). Johnson and Sinharay (2020) developed three additional DCM reliability metrics (point biserial, parallel forms, and information gain) for dichotomous attributes at a single testing occasion. Schellman and Madison (in press) extended the three Johnson and Sinharay (2020) metrics to the TDCM and developed weighted and unweighted versions of each.

This dissertation extends the Schellman and Madison (in press) metrics to the PTDCM setting. Therefore, the cohort-based DGP has five different reliability metrics. Recall that M represents the number of latent classes in the PTDCM framework (the number of classes across all attributes, proficiency statuses, and testing occasions). In this chapter, I provide details for the computation of all PTDCM reliability metrics. In my simulation study (Chapter 4), I used three of the developed metrics. In my empirical data analyses (Chapter 5), I used all PTDCM reliability metrics.

The five PTDCM reliability metrics presented in this chapter utilize four pieces of information called *transition components* (Schellman & Madison, in press) from the PTDCM. Transition Component 1 includes the longitudinal latent class-level *base rates* (the proportion of students who are in the latent class), which is a C-length vector of proportions given by the PTDCM. The class-level base rate for class c is denoted as p_c where $c = 1, 2, \dots, C$ and $C = \prod_{a=1}^{A} l_a^T$.

Transition Component 2 includes the attribute transition base rates. Recall that attribute transition base rates were used in the computation of the DGPs. When computing the attribute transition base rates, the class-level base rates must be summed across the classes that contain each transition. The attribute transition base rate for transition c' and attribute a is denoted as $p_{a,c'}$ where $c' \in \{1, \dots, C_a'\}$, $a \in \{1, \dots, A\}$, and $C_a' = l_a^T$.

Transition Component 3 includes the longitudinal latent class-level posterior probabilities. Each student has a C-length vector of posterior probabilities given by the PTDCM. Borrowing notation from Schellman and Madison (in press), each class-level posterior probability is denoted as $\Pr\{\boldsymbol{\alpha_C} = \boldsymbol{\alpha_c} | \boldsymbol{X_r}\}$ where $\boldsymbol{\alpha_C}$ is the attribute profile for

the student's true attribute profile, $\boldsymbol{\alpha_c}$ is the attribute profile for class c, and $\boldsymbol{X_r}$ is the vector of student r's observed item responses.

Transition Component 4 includes the attribute transition posterior probabilities. When computing the attribute transition posterior probabilities, the class-level posterior probabilities must be summed across the classes that contain each transition. The attribute transition posterior probability for transition c' and attribute a is denoted as $\hat{\alpha}_{c'_{ra}}(\boldsymbol{X_r})$, where $\hat{\alpha}_{c'_{ra}}$ is the estimated marginal probability of student r having transition c'. For simplicity, Transition Component 4 is denoted as $\hat{\alpha}_{c'_{ra}}$ in this dissertation.

Transition Components 1 and 3 are identical for the TDCM and PTDCM—the only difference is that the nature of the latent classes changes. If the diagnostic assessment measures only one attribute, then Transition Components 1 and 2 are identical, Transition Components 3 and 4 are identical, and the classes are the attribute level transitions ($c = c'$). If the assessment measures more than one attribute, then ($c \neq c'$) and Transition Components 2 and 4 require summing the class-level base rates and posterior probabilities, respectively, across the classes that have transition c' for attribute a. For example, suppose a diagnostic assessment measures one dichotomous attribute and one trichotomous attribute at two testing occasions. Then, the transitions for the dichotomous attribute are $c'_1 = \{[00], [01], [10], [11]\}$, and the transitions for the trichotomous attribute are $c'_2 = \{[00], [01], [02], [10], [11], [12], [20], [21], [22]\}$. The 36 latent classes are the 36 combinations of the attribute transitions: $c = \{[0000], [0001], [0002], [0010], [0011], \ldots, [1211], [1212]\}$, where the first and third numbers in each class show the transition for the dichotomous attribute and the second and fourth numbers in each class show the transition for the trichotomous attribute.

Suppose we are computing the reliability for the trichotomous attribute. Computing the base rate and posterior probability for a given transition for the trichotomous attribute (e.g., transition $c' = [12]$) requires summing the base rates and posterior probabilities of the latent classes that include the [12] transition for the trichotomous attribute (i.e., the latent classes [0102], [0112], [1102], [1112]—these are the classes with the profile [*1*2]).

The theoretical motivations behind the specifications of the equations are beyond the scope of this dissertation. Here, I present only the equations adapted from previous works to fit the PTDCM and leave out explanations about how the original equations came to be.

Polychoric Reliability Metric for the PTDCM

The polychoric reliability metric for the PTDCM was first used by Madison and Bao (2018) and Madison and colleagues (2021). These paper presentations did not cover the details of the metric or its computation, so I present those details here. The polychoric reliability metric for the PTDCM utilizes Transition Component 4. First, one must change posterior probability values of 0 and 1 to .00001 and .99999, respectively. Then, for each student r and each attribute a, compute the $l_a^T \times l_a^T$ contingency table where each entry is the average of two attribute transition posterior probabilities indexed by the row and column numbers in the contingency table, which correspond with the transition numbers $c' \in \{1, \ldots, l_a^T\}$. Recall that $\hat{\alpha}_{c'ra}$ denotes Transition Component 4—the posterior probability that student r is in transition c' for attribute a. The $l_a^T \times l_a^T$ contingency table for attribute a is

$$\begin{bmatrix} \frac{\sum_{r=1}^{N}\left(\hat{\alpha}_{1_{ra}}\right)^2}{N} & \frac{\sum_{r=1}^{N}\hat{\alpha}_{1_{ra}}\hat{\alpha}_{2_{ra}}}{N} & \frac{\sum_{r=1}^{N}\hat{\alpha}_{1_{ra}}\hat{\alpha}_{3_{ra}}}{N} & \cdots & \frac{\sum_{r=1}^{N}\hat{\alpha}_{1_{ra}}\hat{\alpha}_{(l_a^T)_{ra}}}{N} \\ \frac{\sum_{r=1}^{N}\hat{\alpha}_{2_{ra}}\hat{\alpha}_{1_{ra}}}{N} & \frac{\sum_{r=1}^{N}\left(\hat{\alpha}_{2_{ra}}\right)^2}{N} & \frac{\sum_{r=1}^{N}\hat{\alpha}_{2_{ra}}\hat{\alpha}_{3_{ra}}}{N} & \cdots & \frac{\sum_{r=1}^{N}\hat{\alpha}_{2_{ra}}\hat{\alpha}_{(l_a^T)_{ra}}}{N} \\ \frac{\sum_{r=1}^{N}\hat{\alpha}_{3_{ra}}\hat{\alpha}_{1_{ra}}}{N} & \frac{\sum_{r=1}^{N}\hat{\alpha}_{3_{ra}}\hat{\alpha}_{2_{ra}}}{N} & \frac{\sum_{r=1}^{N}\left(\hat{\alpha}_{3_{ra}}\right)^2}{N} & \cdots & \frac{\sum_{r=1}^{N}\hat{\alpha}_{3_{ra}}\hat{\alpha}_{(l_a^T)_{ra}}}{N} \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ \frac{\sum_{r=1}^{N}\hat{\alpha}_{(l_a^T)_{ra}}\hat{\alpha}_{1_{ra}}}{N} & \frac{\sum_{r=1}^{N}\hat{\alpha}_{(l_a^T)_{ra}}\hat{\alpha}_{2_{ra}}}{N} & \frac{\sum_{r=1}^{N}\hat{\alpha}_{(l_a^T)_{ra}}\hat{\alpha}_{3_{ra}}}{N} & \cdots & \frac{\sum_{r=1}^{N}\left(\hat{\alpha}_{(l_a^T)_{ra}}\right)^2}{N} \end{bmatrix} \quad (35)$$

Then, the polychoric reliability metric for attribute a, $\hat{\rho}_{\text{poly}_a}$, can be computed by using the polychoric correlation for the contingency table.

Average Maximum Transition Reliability Metric for the PTDCM

The average maximum transition metric for the PTDCM, $\max\{\hat{\alpha}_a\}$, is computed the same way as the average maximum transition metric for the TDCM (Schellman & Madison, in press). For each attribute, first compute Transition Component 4: the marginal posterior probabilities for each attribute's transitions for each student. Then, obtain each student's maximum attribute transition posterior probability, $\max\{\hat{\alpha}_{ra}\}$. Finally, compute the average of all students' maximum attribute transition posterior probabilities:

$$\max\{\hat{\alpha}_a\} = \sum_{r=1}^{N}\frac{\max\{\hat{\alpha}_{ra}\}}{N} \quad (36)$$

For example, if an assessment measures two dichotomous attributes at two testing occasions, then the total number of latent classes is 16: [0000], [0001], [0010], …, [1110], and [1111]. To compute the $\max(\hat{\alpha}_a)$ for the first attribute, one must compute four marginal posterior probabilities for each student (Transition Component 4, $\hat{\alpha}_{c'_{ra}}$): the sum of the class-level posterior probabilities that include the transition [00] for the

first attribute (the classes with the profile [0*0*]), the sum of the class-level posterior probabilities that include the transition [01] for the first attribute (the classes with the profile [0*1*]), the sum of the class-level posterior probabilities that include the transition [10] for the first attribute (the classes with the profile [1*0*]), and the sum of the class-level posterior probabilities that include the transition [11] for the first attribute (the classes with the profile [1*1*]). Then, compute each student's maximum attribute transition posterior probability for Attribute 1, $\max\{\hat{\alpha}_{r1}\}$. Finally, compute the average of all students' maximum attribute transition posterior probabilities, $\max\{\hat{\alpha}_{1}\}$.

The computation for the marginal posterior probability changes for different numbers of proficiency statuses and testing occasions, but all computations would follow the same logic. For example, if an assessment measures two attributes, one dichotomous and one trichotomous attribute, at two testing occasions, then the total number of latent transitions or classes is 36: [0000], [0001], [0002], …, [1211], and [1212]. To compute the average maximum transition reliability metric for the second attribute, one must compute nine marginal posterior probabilities: the sum of the class-level posterior probabilities with the profile [*0*0]), the sum of the class-level posterior probabilities with the profile [*0*1]), the sum of the class-level posterior probabilities with the profile [*0*2]), the sum of the class-level posterior probabilities with the profile [*1*0]), the sum of the class-level posterior probabilities with the profile [*1*1]), the sum of the class-level posterior probabilities with the profile [*1*2]), the sum of the class-level posterior probabilities with the profile [*2*0]), the sum of the class-level posterior probabilities with the profile [*2*1]), and the sum of the class-level posterior probabilities with the profile [*2*2]). Then, compute each student's maximum attribute

transition posterior probability for Attribute 2, $\max\{\hat{\alpha}_{r2}\}$. Finally, compute the average of all students' maximum attribute transition posterior probabilities, $\max\{\hat{\alpha}_{2}\}$.

It is helpful to compute the proportion of students who have maximum transition probabilities greater than .9 (i.e., the proportion of students for whom the model is very confident in their classifications), greater than .8, greater than .7, and so on to accompany the average maximum transition reliability metric. The greater the proportion of students who have maximum transition probabilities greater than larger checkpoints along the zero-to-one scale, the more confident the model is overall in student classifications.

Point Biserial Reliability Metric for the PTDCM

The longitudinal point biserial reliability metric for attribute a in the PTDCM, $\hat{\rho}_{\text{lpb}_a}$, utilizes Transition Components 2 and 4 and is given by the same equation in Schellman and Madison (in press)—the only difference is that the number of transitions is not always four because the attributes are not always dichotomous in the PTDCM. Therefore, the only change that needs to be made to Schellman and Madison's (in press) equation for the longitudinal point biserial reliability metric is to replace "4" with l_a^T:

$$\hat{\rho}_{\text{lpb}_a} = \frac{1}{l_a^T}\sum_{c'=1}^{l_a^T}\left[\frac{\frac{1}{N}\sum_{r=1}^{N}\left[(\hat{\alpha}_{c'ra})^2 - \left(p_{a,c'}\right)^2\right]}{p_{a,c'}\left(1-p_{a,c'}\right)}\right] \tag{37}$$

Equation 38 is essentially an average of the point biserial reliabilities of each transition for the attribute. Instead of computing an average, one could weight each transition's point biserial reliability by its corresponding base rate:

$$\hat{\rho}_{\text{wpb}_a} = \sum_{c'=1}^{l_a^T}\left(p_{a,c'}\right)\frac{\frac{1}{N}\sum_{r=1}^{N}\left[(\hat{\alpha}_{c'ra})^2 - \left(p_{a,c'}\right)^2\right]}{p_{a,c'}\left(1-p_{a,c'}\right)} \tag{38}$$

The weighted version is needed when a base rate for a transition is equal to zero because, in such a case, the unweighted version will not work unless you drop that transition out of the average.

Parallel Forms Reliability Metric for the PTDCM

The parallel forms reliability metric for attribute a in the PTDCM, $\hat{\rho}_{\text{lpf}_a}$, utilizes Transition Components 1, 2, 3, and 4 and is given by the same equation in Schellman and Madison (in press). As with the longitudinal point biserial metric, the only change that needs to be made to Schellman and Madison's (in press) equation for the longitudinal parallel forms reliability metric is to replace "4" with l_a^T:

$$\hat{\rho}_{\text{lpf}_a} = \frac{1}{l_a^T}\sum_{c'=1}^{l_a^T}\left[\frac{\sum_c^C \frac{1}{p_c}\left[\frac{1}{N}\sum_{r=1}^N(\hat{\alpha}_{c'ra})(\Pr\{\boldsymbol{\alpha_C} = \boldsymbol{\alpha_c}|\boldsymbol{X_r}\})\right]^2 - \left(p_{a,c'}\right)^2}{\frac{1}{N}\sum_{r=1}^N\left[\left(\hat{\alpha}_{c'ra}\right)^2 - \left(p_{a,c'}\right)^2\right]}\right] \tag{39}$$

Equation 40 is essentially an average of the parallel forms reliabilities of each transition for the attribute. Instead of computing an average, one could weight each transition's parallel forms reliability by its corresponding base rate:

$$\hat{\rho}_{\text{wlpf}_a} = \sum_{c'=1}^{l_a^T}(p_{a,c'})\left[\frac{\sum_c^C \frac{1}{p_c}\left[\frac{1}{N}\sum_{r=1}^N(\hat{\alpha}_{c'ra})(\Pr\{\boldsymbol{\alpha_C} = \boldsymbol{\alpha_c}|\boldsymbol{X_r}\})\right]^2 - \left(p_{a,c'}\right)^2}{\frac{1}{N}\sum_{r=1}^N\left[\left(\hat{\alpha}_{c'ra}\right)^2 - \left(p_{a,c'}\right)^2\right]}\right] \tag{40}$$

If any base rate is equal to zero, I recommend changing it to be .00001 prior to computing the metrics.

Information Gain Reliability Metric for the PTDCM

The unweighted and weighted information gain reliability metrics for attribute a in the PTDCM, $\hat{\rho}_{\text{lig}_a}$ and $\hat{\rho}_{\text{wlig}_a}$, respectively, are computed using Transition Components 2 and 4 and require this series of equations:

$$H1_{c'_a} = -p_{a,c'} \ln(p_{a,c'}) - (1 - p_{a,c'}) \ln(1 - p_{a,c'}) \tag{41}$$

$$H2_{c'_a} = -\frac{1}{N} \sum_{r=1}^{N} [\hat{\alpha}_{c'_{ra}} \times \ln(\hat{\alpha}_{c'_{ra}}) + (1 - \hat{\alpha}_{c'_{ra}}) \times \ln(1 - \hat{\alpha}_{c'_{ra}})] \tag{42}$$

$$\hat{\rho}_{\text{lig}_a} = \frac{1}{l_a^T} \sum_{c'=1}^{l_a^T} [1 - \exp(-2[H1_{c'_a} - H2_{c'_a}])] \tag{43}$$

$$\hat{\rho}_{\text{wlig}_a} = \sum_{c'=1}^{l_a^T} [(p_{a,c'}) \times (1 - \exp(-2[H1_{c'_a} - H2_{c'_a}]))] \tag{44}$$

In sum, in this chapter, I extended TDCM reliability metrics to the PTDCM setting, resulting in eight reliability metrics for the PTDCM. For the rest of this paper I refer to these metrics using the following abbreviations: polychoric reliability metric for the PTDCM (polychoric), average maximum transition reliability metric for the PTDCM, point biserial reliability metric for the PTDCM (PB), weighted PB (PBW), parallel forms reliability metric for the PTDCM (PF), weighted PW (PFW), information gain reliability metric for the PTDCM (IG), and weighted IG (IGW).

Factors that Might Impact the Reliability of DGPs

Schellman and Madison (in press) showed that the TDCM reliability metrics all increased as the number of items measuring each attribute increased, as the item quality increased, and as the number of attributes increased, but the reliabilities were relatively unimpacted by the magnitude of the correlations within the attributes (across testing occasions). These authors additionally showed that the metric with the lowest reliabilities was the information gain metric, and the metric with the greatest reliabilities was the test-retest metric with the polychoric correlation. The point biserial and parallel forms metrics were all similar (Schellman & Madison, 2021; Schellman & Madison, in press).

Madison and colleagues (2021) reported that as the number of attribute proficiency statuses for polytomous attributes increased and holding test length constant, the reliability of the attribute decreased. I expected similar results for the PTDCM reliability metrics, which I evaluated in a simulation study. Chapter 4 presents the results of my simulation study to evaluate the performance of the PTDCM reliability metrics under various DCM assessment conditions.

Disclaimers About Diagnostic Growth Percentiles and Their Reliabilities

As discussed previously, the primary student result from a DCM is the multidimensional attribute profile that indicates the student's estimated proficiency status with respect to each measured attribute. Underlying this primary student result are the posterior probabilities that the student is in each latent class. One of the key arguments in favor of using DCMs in education is that the attribute profiles provide criterion-referenced results about what students do and do not know. DCMs are designed to provide reliable and accurate student classifications—not reliable and accurate student posterior probabilities of attribute proficiency. Therefore, it is inappropriate to rank-order students by their marginal posterior probabilities of attribute proficiency because such rank-ordering does not adhere to the intended purpose of DCMs and is not meaningful statistically or practically.

However, the adjusted DGP metrics use students' posterior probabilities to quantify relative growth in the DCM framework. One may wonder whether the adjusted DGPs are violating the proper use of DCM results by incorporating posterior probabilities into a metric designed to rank-order growth. I argue that the adjusted DGP metrics are not inappropriately using posterior probabilities as a means of rank-ordering students because

the adjusted DGPs consider the entire distribution of a student's most likely class rather than only part of that distribution (e.g., a marginal or maximum posterior probability), and the posterior probabilities are used in the adjusted DGPs to weight the basic DGPs, so the adjusted DGPs do not use posterior probabilities directly to rank-order students. The context of ordering student growth using the adjusted DGP metrics is different from the content of using posterior probabilities to directly rank-order students.

Another disclaimer about my conceptualization of DGPs is that I am borrowing the PTDCM reliability for all four types of DGPs, though one may argue that this "borrowing" is inappropriate. It is possible that the PTDCM reliability metrics are only appropriate for the basic DGP, and the adjusted DGPs should have a different reliability metric because the PTDCM reliability metrics are extensions of reliability metrics that have been used to measure student classification reliability—not reliability for posterior probabilities, and the basic DGP is the only DGP that includes only student classifications while the adjusted DGPs include student classifications and posterior probabilities. On the other hand, some may argue that the average maximum transition metric is most aligned with the basic DGP because it focuses on classification consistency, and the basic DGP includes only student classifications, while the polychoric, PB, PF, and IG metrics are more aligned with the adjusted DGPs because these metrics and the adjusted DGPs all incorporate the posterior probabilities in addition to student classifications.

So, what is the most appropriate reliability metric for each DGP type? We do not yet know. This way of using posterior probabilities is new and requires further research to determine whether it is appropriate for specific contexts. Additionally, more research is

needed to explore other ways to conceptualize DGP reliability and evaluate whether the different types of DGPs require different reliability metrics. My approaches for conceptualizing DGPs and their reliabilities are a few of the potentially many ways to incorporate model uncertainty, penalize students for forgetting, and measure reliability. As mentioned previously, this study is meant to explore the efficacy and validity of my conceptualizations and to encourage further theoretical and practical research and discussion amongst researchers and practitioners, which must occur prior to implementing DGPs in practice. For the purposes of this dissertation, I treat the reliability of each attribute for each data set as the reliability for all four DGP metrics (i.e., all four DGP metrics have the same reliability for each attribute in each data set), regardless of which reliability metric is used.

CHAPTER 4

SIMULATION STUDY

In Chapter 3, I answered the research question (1d): How can DGP reliability be conceptualized and computed? I developed reliability metrics for the PTDCM, and I explained that PTDCM reliability is DGP reliability. In this chapter, I outline the simulation study I conducted to evaluate DGP reliability under various DCM conditions. The simulation study I conducted for this dissertation seeks to answer the second research question presented in Chapter 1:

2. How reliable are DGPs under different DCM assessment conditions?
    a. How do the reliability metrics for DGPs perform under different assessment conditions?

These research questions (2) and (2a) have a nuanced difference: I am interested in knowing how reliable DGPs are under various assessment conditions, but answering this research question simultaneously answers the research question about how the PTDCM reliability metrics perform under different assessment conditions because PTDCM reliability is DGP reliability.

I begin this chapter by explaining the simulation study design, including my approach to generating data and each analysis that I conducted on the data. Then, I present the results of the simulation study. This chapter concludes with a discussion of the simulation study, including its significance, limitations, and future research.

Simulation Study Design

The fixed factors in this study include the model, the sample size, the item type, the number of testing occasions, and the base rates for T1. The manipulated factors include the item quality, the number of items to measure each attribute, the attribute proficiency growth, the attribute correlations, the number of proficiency statuses per attribute, and the number of attributes. Table 8 summarizes the design of this simulation study.

Fixed Factors

The generating and estimating models for this study are both the PTDCM. Thus, the estimating model was the true model. No misspecified conditions were included in this study. I chose to fix the sample size at 1,000 students because this sample size was in line with TDCM research (e.g., Schellman & Madison, in press), and it closely resembles the sample sizes used in the empirical data analyses in Chapter 5. I also chose to fix the conditions to all use a pre-/post-test design (i.e., two testing occasions) because this practice was in line with other longitudinal DCM research (e.g., Madison, 2019; Schellman & Madison, in press) and the empirical data analyses in Chapter 5 each include only two testing occasions.

As in other TDCM-related research (Madison, 2019; Schellman & Madison, in press), this study uses only simple-structure items because it better isolates the impact of test length, which can be hidden when complex items are included. In the PTDCM framework with imposed measurement invariance, the number of item parameters for a simple-structure item is the same as the number of proficiency statuses for the corresponding attribute. Thus, simple-structure items that measure an attribute with three

proficiency statuses each have three item parameters, and simple-structure items that measure an attribute with five proficiency statuses each have five item parameters.

I set the base rates for T1 for each proficiency status. For attributes with three proficiency statuses and the labels *beginning*, *proficient*, and *advanced*, the base rates for T1 are .6, .25, and .15, respectively. This means that at T1, 60% of students were beginning, 25% were proficient, and 15% were advanced for the trichotomous attribute. For attributes with five proficiency statuses and the labels *beginning*, *developing*, *proficient*, *distinguished*, and *advanced*, the base rates for T1 are .45, .25, .15, .1, and .05, respectively. This means that at T1, 45% of students were beginning, 25% were developing, 15% were proficient, 10% were distinguished, and 5% were advanced for the attribute with five proficiency statuses. I set these base rates so that the lowest proficiency status had the greatest proportion of students, and the highest proficiency status had the smallest proportion of students at the pre-test. The decreasing proportions at T1 are reflective of a pre-testing occasion or initial measurement where students have not had the opportunity to learn the material or are about to receive instruction.

The labels of beginning, proficient, and advanced for the trichotomous attribute and beginning, developing, proficient, distinguished, and advanced for the attribute with five proficiency statuses are intended for illustrative purposes and are not meant to apply qualitative meaning to students' classifications. In practice, experts and stakeholders should determine the appropriate labels for the student classifications that are most meaningful in their specific setting.

Manipulated Factors

I varied the item quality to be lower or higher—I specifically did not label these levels of item quality as "low" or "high" in an absolute sense because the interpretation of whether specific item quality levels are low or high depends on specific assessment contexts. I generated the item parameters using item discrimination as a starting point. For this dissertation, I define item discrimination for the PTDCM as the difference in IRPs for students in different proficiency statuses, denoted $d_{l_a,(l_b l_b')}$, where l_a is the total number of proficiency statuses for attribute a, and l_b and $l_{b'}$ are different proficiency statuses such that $l_b \in \{1, 2, \dots, l_a\}$, $l_{b'} \in \{1, 2, \dots, l_a\}$, and $l_b \neq l_{b'}$. As the number of proficiency statuses increases, the number of discriminations increases. For this dissertation, I focused only on the discriminations for adjacent proficiency statuses and for the maximum and minimum proficiency statuses.

Thus, for a trichotomous attribute, there are two discrimination values with adjacent proficiency statuses: (1) the IRP for proficient students minus the IRP for beginning students, denoted $d_{3,10}$, and (2) the IRP for advanced students minus the IRP for proficient students, denoted $d_{3,21}$. These discrimination values are given by the simplified PDCM notation,

$$d_{3,10} = \frac{\exp(\lambda_0 + \lambda_1)}{1 + \exp(\lambda_0 + \lambda_1)} - \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)} \tag{456}$$

$$d_{3,21} = \frac{\exp(\lambda_0 + \lambda_1 + \lambda_2)}{1 + \exp(\lambda_0 + \lambda_1 + \lambda_2)} - \frac{\exp(\lambda_0 + \lambda_1)}{1 + \exp(\lambda_0 + \lambda_1)} \tag{467}$$

For the lower item quality conditions, I sampled $d_{3,10}$ and $d_{3,21}$ from $N(.2, .025)$. For the higher item quality conditions, I sampled $d_{3,10}$ and $d_{3,21}$ from $N(.3, .025)$. I chose these sampling distributions for the adjacent discriminations so the maximum

discrimination, $d_{3,20}$, would be about .4 and .6 for the lower and higher item quality conditions, respectively. I resampled if the generated discrimination was less than zero or greater than one.

For an attribute with five proficiency statuses, there are four discrimination values with adjacent proficiency statuses: (1) the IRP for developing students minus the IRP for beginning students, denoted $d_{5,10}$, (2) the IRP for proficient students minus the IRP for developing students, denoted $d_{5,21}$, (3) the IRP for distinguished students minus the IRP for proficient students, denoted $d_{5,32}$, and (4) the IRP for advanced students minus the IRP for distinguished students, denoted $d_{5,43}$. These discrimination values are given by

$$d_{5,10} = \frac{\exp(\lambda_0 + \lambda_1)}{1 + \exp(\lambda_0 + \lambda_1)} - \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)} \tag{4847}$$

$$d_{5,21} = \frac{\exp(\lambda_0 + \lambda_1 + \lambda_2)}{1 + \exp(\lambda_0 + \lambda_1 + \lambda_2)} - \frac{\exp(\lambda_0 + \lambda_1)}{1 + \exp(\lambda_0 + \lambda_1)} \tag{4948}$$

$$d_{5,32} = \frac{\exp(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3)}{1 + \exp(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3)} - \frac{\exp(\lambda_0 + \lambda_1 + \lambda_2)}{1 + \exp(\lambda_0 + \lambda_1 + \lambda_2)} \tag{490}$$

$$d_{5,43} = \frac{\exp(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)}{1 + \exp(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)} - \frac{\exp(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3)}{1 + \exp(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3)} \tag{501}$$

For the lower item quality conditions, I sampled $d_{5,10}$, $d_{5,21}$, $d_{5,32}$, and $d_{5,43}$ from $N(.125, .025)$. For the higher item quality conditions, I sampled $d_{5,10}$, $d_{5,21}$, $d_{5,32}$, and $d_{5,43}$ from $N(.175, .025)$. I chose these sampling distributions for the adjacent discriminations so the maximum discrimination, $d_{5,40}$, would be about .5 and .7 for the lower and higher item quality conditions, respectively. I resampled if the generated discrimination was less than zero or greater than one.

After generating the discrimination values, I sampled the intercept parameter, λ_0, for each item from $U[-2.19, -0.41]$ to give beginning students an IRP between about 10% and 40%. Then, given the generated discrimination $d_{l_a,10}$ and intercept parameter λ_0, I used Equations 46 and 48 (the equations for $d_{3,10}$ and $d_{5,10}$) to solve for the main effect, λ_1, which is the main effect for the proficient and developing students for the trichotomous and attributes with five proficiency statuses, respectively. Then, given the generated discrimination $d_{l_a,21}$, the generated intercept parameter λ_0, and the derived main effect parameter λ_1, I used Equations 47 and 49 (the equations for $d_{3,21}$ and $d_{5,21}$) to solve for the main effect, λ_2, which is the main effect for the advanced and proficient students for the trichotomous and attributes with five proficiency statuses, respectively. I used the same process to derive the remaining main effects (i.e., λ_3 and λ_4) for the attribute with five proficiency statuses. In item parameter generation, I also required that the IRP for the students in the greatest proficiency status was greater than .5 because students who are maximally proficient should have a greater than 50% chance of answering each item correctly, as is common practice in DCM literature.

After generating the item parameters using this approach, I computed the observed true distributions of the different IRP values for all replications and all conditions. Table 9 shows the minimum, mean, median, and maximum values for the resulting true IRP discriminations for the different types of conditions in my simulation study. Table 9 shows that, on average, the discriminations generated were the desired discriminations because, for the lower item quality conditions, $d_{3,10}$ and $d_{3,21}$ are about .2, on average, and $d_{5,10}$, $d_{5,21}$, $d_{5,32}$, and $d_{5,43}$ are about .125, on average, and for the higher item quality conditions, $d_{3,10}$ and $d_{3,21}$ are about .3, on average, and $d_{5,10}$, $d_{5,21}$,

$d_{5,32}$, and $d_{5,43}$ are about .175, on average. These adjacent discriminations together yielded the desired maximum discriminations of about .4 for the conditions with three proficiency statuses and lower-quality items, about .6 for the conditions with three proficiency statuses and higher-quality items, about .5 for the conditions with five proficiency statuses and lower-quality items, and about .7 for the conditions with five proficiency statuses and higher-quality items.

I varied the number of items per attribute to be 8 or 12 at each testing occasion to be comparable with test lengths used in other PTDCM studies that showed adequate model performance (e.g., Madison & Bao, 2018). Madison and Bao (2018) also investigated conditions with as few as four items per attribute and as many as 16 items per attribute, but if I use 16 items per attribute in my simulation study, which has only simple-structure items, then conditions with three attributes would have 48 items to measure three attributes, which is a long assessment for so few attributes, in relation to traditional DCM literature. It goes against the primary desirable feature of DCMs in that they allow for shorter test lengths than IRT-based assessments. Of course, polytomous attributes are more complex and require more items than dichotomous attributes, but we still must consider the number of items we can realistically expect students to answer within a given testing occasion. Therefore, I used eight or 12 items per attribute to consider how realistically short and long assessments perform under different diagnostic assessment conditions.

I varied the number of proficiency statuses per attribute to be three or five. Bao (2019) used three proficiency statuses. I added attributes with five proficiency statuses to

this study because having more proficiency statuses better illustrates the utility of the DGP metrics by adding more variability in student classifications and transitions.

I varied the number of attributes to be one or three attributes to illustrate how the DGP reliability performs with unidimensional and multidimensional diagnostic assessments. However, for the conditions with three attributes, I only used three proficiency statuses per attribute. I did not use five proficiency statuses for conditions with more than one attribute due to the computational demand of estimating a multidimensional polytomous attribute model with a large number of proficiency statuses per attribute.

Using Proficiency Growth and Attribute Correlations to Generate Attribute Profiles

To align with the study by Schellman and Madison (in press), I varied the target attribute correlations to be 0, .25, .5, and .75, where I fixed the correlation within each simulation condition. For the conditions with multidimensional assessments ($A = 3$), I applied the attribute correlation for attributes within the same testing occasion (within correlations) and for attributes at different testing occasions (between correlations), but for conditions with unidimensional assessments ($A = 1$), the correlation only referred to the correlations between the single attribute and itself at different testing occasions. Although attribute correlations below about .4 are not expected in applications with educational data, it is worthwhile for this study to evaluate the research questions under some extreme conditions as long as some realistic conditions (i.e., attribute correlations of .5 and .75) are also evaluated.

Low correlations have also been used in simulation studies that investigated the reliability of SGPs. For example, Monroe and Cai's (2015) study used the MIRT model

to measure the same latent variable at multiple testing occasions (i.e., a unidimensional assessment given at multiple testing occasions). These authors generated the latent variable to have between-testing occasion correlations ranging from .05 to .95 in increments of .05, and they used an autoregressive correlation matrix, which was designed to show that the between-testing occasion correlations decrease as more time passes between testing occasions (e.g., the between-testing occasion correlation should be greater for Testing Occasions 1 and 2 than for Testing Occasions 1 and 4). In the autoregressive correlation matrix, some between-testing correlations were as low as .000125 for some generation conditions. Therefore, it is reasonable that the current study includes lower correlations, such as 0 and .25.

To explain how I generated students' proficiency statuses, I first present my approach generally for any number of attributes and testing occasions. I sampled students' proficiency statuses by first generating TA multivariate random normal variables of length N, denoted $\boldsymbol{T}$, where $\boldsymbol{T} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is a TA-length vector of 0s, and $\boldsymbol{\Sigma}$ is a $TA \times TA$ matrix with diagonal elements equal to one and off-diagonal elements equal to the desired correlation for the simulation condition, ρ. Specifically,

$$\boldsymbol{T} = \begin{Bmatrix} T_{11} \\ \dots \\ T_{ta} \\ \dots \\ T_{TA} \end{Bmatrix} \sim N\left(\begin{Bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{Bmatrix}, \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \rho & \rho & \ddots & \dots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \right) \tag{5251}$$

This approach allows the underlying distributions for the attribute profiles to have specified correlations within and between testing occasions. Note that for unidimensional assessments, only between correlations are relevant. For unidimensional assessments, $\boldsymbol{T}$ simplifies to

$$\boldsymbol{T} = \begin{Bmatrix} T_{11} \\ T_{21} \end{Bmatrix} \sim N\left(\begin{Bmatrix} 0 \\ 0 \end{Bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \quad (5352)$$

In Equation 53, $T = 2$ and $A = 1$, so the first variable in $\boldsymbol{T}$, T_{11}, corresponds with the attribute proficiency at T1, and the second variable in $\boldsymbol{T}$, T_{21}, corresponds with the attribute proficiency at T2. For the rest of this section, I use the one attribute situation to explain my generation approach.

Next, I used the cumulative base rates as the percentiles to yield the desired proportions of students in each proficiency status for T1. For example, for the trichotomous attribute, I used percentile ranks of .6, .85, and 1 to divide the students' T_1 values into groups of the desired size for each proficiency status. I varied the attribute proficiency growth to be no growth, moderate growth, or large growth. For the trichotomous attribute, no growth corresponds with base rates of .6, .25, and .15 for the beginning, proficient, and advanced groups, respectively. These are the same base rates that I used for T1. For the trichotomous attribute, moderate growth corresponds with base rates of .15, .6, and .25 for the beginning, proficient, and advanced groups, respectively. For the trichotomous attribute, large growth corresponds with base rates of .15, .25, and .6 for the beginning, proficient, and advanced groups, respectively. Table 10 summarizes the base rates for the different levels of growth for the attribute with five proficiency statuses. In the same way that I used the cumulative base rates as the percentiles T_1 to yield the desired proportions of students in each proficiency status for T1, I used the cumulative base rates for the level of growth specified in the simulation condition as the percentiles T_2 to yield the desired proportions of students in each proficiency status for T2. In sum, to generate attribute profiles, I discretized multivariate random normal

variables that had the desired correlation (Pearson correlation), which yielded categorical variables that had the same desired correlation (polychoric correlation).

After applying this approach to attribute profile generation, I evaluated the true attribute correlations for the generated attribute profiles to see the resulting distributions of the true correlations between all generated attribute proficiency statuses. Figure 2 shows the true distributions of all attribute correlations across all replications of all conditions. Table 11 shows the summary statistics for the distributions in Figure 2. Figure 2 and Table 11 show that I correctly generated the attribute profiles to yield the desired attribute correlations because the distributions are centered around the target attribute correlations.

Evaluation Metrics for Simulation Study

For each condition, I report the convergence rates across all replications within each condition. I also report the average item parameter estimation accuracy for each condition by computing the mean absolute difference (MAD) between the generated and estimated IRPs. I compute MAD in terms of the IRPs rather than the item parameters themselves because IRPs are on the probability scale, so they are easier to interpret than DCM item parameters, which are on the logit scale. Each item with three proficiency statuses has three MAD values because it has three IRPs—one for each proficiency status: beginning, proficient, and advanced. Each item with five proficiency statuses likewise has five MAD values because it has five IRPs—one for each proficiency status: beginning, developing, proficient, distinguished, and advanced. I computed each MAD value using Equation 54:

$$MAD_{l_a} = |\pi_{l_a} - \widehat{\pi_{l_a}}| \tag{54}$$

where π_{l_a} is the true IRP for the proficiency group l_a, and $\widehat{\pi_{l_a}}$ is the estimated IRP for the proficiency group l_a. I report the average MAD value for each condition for all IRPs together and for each IRP separately.

I report the average student classification accuracy at the profile and attribute levels for each condition. The average profile or class correct classification rate (CCCR) for a single replication is computed as

$$CCCR = \frac{\sum_{r=1}^{N} x_r}{N} \tag{55}$$

$$x_r = \begin{cases} 1 \text{ if } \boldsymbol{\alpha_r} = \widehat{\boldsymbol{\alpha_r}} \\ 0 \text{ if } \boldsymbol{\alpha_r} \neq \widehat{\boldsymbol{\alpha_r}} \end{cases}$$

where $\boldsymbol{\alpha_r}$ is student r's true attribute profile, and $\widehat{\boldsymbol{\alpha_r}}$ is student r's estimated attribute profile. For the conditions with three attributes, I considered the CCCR values for students' most likely latent classes at the pre-test and post-test (*CCR Pre* and *CCR Post*, respectively) separately and for their longitudinal most likely latent classes together (*CCCR*).

I used Equation 56 to compute the average attribute correct classification rate (ACCR) for a single replication:

$$ACCR_t = \frac{\sum_{a=1}^{A} \frac{\sum_{r=1}^{N} x_{rat}}{N}}{A} \tag{56}$$

$$x_{rat} = \begin{cases} 1 \text{ if } \alpha_{rat} = \widehat{\alpha_{rat}} \\ 0 \text{ if } \alpha_{rat} \neq \widehat{\alpha_{rat}} \end{cases}$$

where α_{rat} is student r's true proficiency status for attribute a at testing occasion t, and $\widehat{\alpha_{rat}}$ is student r's estimated proficiency status for attribute a at testing occasion t. Notice that I compute the ACCR values by averaging the attribute-level classification

accuracy across attributes within the same testing occasion. However, I report the ACCR for each testing occasion separately.

I also report the average reliabilities for each condition using three of the reliability metrics introduced in Chapter 3: the polytomous polychoric reliability metric, the polytomous average maximum transition metric, and the polytomous longitudinal PB metric. I selected the polytomous PB metric for use with my simulation study because, in previous studies with the PB, PF, and IG metrics (Johnson & Sinharay, 2020; Schellman & Madison, in press), the PB and PF metrics yield comparable reliabilities, but the IG metric yields lower reliabilities.

In addition to reporting the reliability metrics, to further investigate the model performance and the average maximum transition reliability metric, I computed nine other evaluation metrics to examine the robustness of the PTDCM reliability: I computed the proportion of students' maximum transitions that are greater than .1, greater than .2, and so on, up to greater than .9. The greater the maximum transition probabilities, the more confident the model is in the student classifications. I hope to see large proportions of students with large maximum transition probabilities.

Finally, I computed the basic and adjusted DGPs for all students in all replications of all conditions for all attributes in the condition. Specifically, I computed the minimum, mean, median, and maximum of each type of DGP for each attribute in each condition to report how students' DGPs are impacted across different DCM assessment conditions.

The Full Model Approach Versus the Calibrate-and-Score Approach

Previous research with polytomous attributes measured over multiple testing occasions has introduced two key approaches for estimating the item parameters and

student transitions: the full model approach and the calibrate-and-score approach (Madison et al., 2023). Note that, in this section, I am only considering situations with two testing occasions, but both of these approaches can be used with more than two testing occasions. The full model approach uses the PTDCM to estimate transition probabilities, longitudinal MLCs, and longitudinal posterior probabilities directly. The calibrate-and-score approach first calibrates the PDCM for the pre-test item responses and then uses the resulting estimated item parameters to score the post-test item responses, but it allows the base rates at each testing occasion to be freely estimated. In other words, the post-test base rates are not fixed to use the base rates estimated from the pre-test, so even though the calibrate-and-score approach does not directly estimate transition probabilities as the full model approach does, it estimates base rates at each testing occasion. Because the calibrate-and-score approach does not yield longitudinal MLC or longitudinal posterior probabilities (longitudinal posterior probabilities are just called "posterior probabilities" in the full model approach), one must combine the cross-sectional results obtained from calibrating the pre-test and scoring the post-test.

For the calibrate-and-score approach, students' cross-sectional MLCs from each testing occasion are produced separately from calibrating the pre-test and then scoring the post-test. The cross-sectional MLCs can be combined to obtain students' longitudinal MLCs. For example, consider a three-attribute assessment with three proficiency statuses per attribute that was administered at two testing occasions. Suppose that the result of calibrating the pre-test was that Student X's estimated pre-test attribute profile was [102] (i.e., proficient for the first attribute, beginning for the second attribute, and advanced for the third attribute), and the result of scoring the post-test was that Student X's estimated

post-test attribute profile was [212] (i.e., advanced for the first and third attributes, and proficient for the second attribute). Therefore, combining the pre- and post-test attribute profiles gives a longitudinal attribute profile of [102212], which corresponds with Student X's longitudinal MLC.

The PTDCM reliability metrics require students' longitudinal posterior probabilities, but the calibrate-and-score approach yields only cross-sectional posterior probabilities for the pre- and post-test separately. In the calibrate-and-score approach, students' longitudinal posterior probabilities can be obtained by computing the *cross-product* of the cross-sectional posterior probabilities. Specifically, to obtain longitudinal posterior probabilities from cross-sectional posterior probabilities for a three-attribute assessment with three proficiency statuses per attribute, one must first create an $N \times l_a^A = N \times 3^3 = N \times 27$ matrix of the N students' posterior probabilities for the 27 latent classes at the pre-test. Then, one must create an analogous matrix of the N students' posterior probabilities for the 27 latent classes at the post-test. All rows in each of these matrices sum to 1 because of the properties of DCM posterior probabilities. Next, one must create a $N \times l_a^{TA}$ matrix of posterior probabilities for T testing occasions by computing the cross-product of the pre- and post-test matrices. To illustrate the cross-product, consider the probability of students being in Class 1 at the pre-test and being in Class 1 at the post-test. To compute the posterior probabilities of students belonging to Class 1 at the pre-test and Class 1 at the post-test (i.e., the first column of the $N \times l_a^{TA}$ matrix), one must multiply students' cross-sectional posterior probabilities of belonging to Class 1 at the pre-test by their cross-sectional posterior probabilities of belonging to Class 1 at the post-test. In other words, one must multiply the first column of the pre-test

matrix with the first column of the post-test matrix. The resulting product becomes the first column of the $N \times l_a^{TA}$ matrix. One must similarly multiply all combinations of latent class posterior probabilities across the pre- and post-test to obtain the rest of the longitudinal posterior probabilities. All rows in the $N \times l_a^{TA}$ matrix will sum to 1 and, thus, maintain the properties of DCM posterior probabilities.

At this point, one would have obtained students' longitudinal MLCs by combining the cross-sectional MLCs and would have obtained students' longitudinal posterior probabilities by computing the cross-product of the cross-sectional posterior probability matrices. However, it is possible that the longitudinal MLC does not agree with the class that corresponds with the maximum longitudinal posterior probability. For example, consider again Student X, who had an estimated pre-test MLC of [102], an estimated post-test MLC of [212], and a longitudinal MLC of [102212], which was estimated via the calibrate-and-score approach. Suppose that Student X's maximum longitudinal posterior probability obtained from the cross-products of the cross-sectional posterior probabilities indicated that Student X's most likely longitudinal MLC should be [101112]. This profile indicates that the pre-test profile should be [101] and the post-test profile should be [112], which does not agree with the cross-sectional MLCs.

When this disagreement happens, one can choose whether to let the class corresponding with the maximum longitudinal posterior probability override the longitudinal MLC or vice versa. For my simulation study, I chose to let the class corresponding with the maximum longitudinal posterior probability override the longitudinal MLC, and I updated the affected students' cross-sectional MLCs accordingly. Unfortunately, no matter which class one chooses to be the final class, one

set of posterior probabilities will not be accurate. With my choice, the cross-sectional posterior probabilities no longer correspond with the class that I determined to be the most likely longitudinal class based on the longitudinal posterior probabilities. If I had made the opposite choice, then the cross-sectional posterior probabilities would match the final class, but the longitudinal posterior probabilities would not. Because my study requires the longitudinal posterior probabilities, my choice to let the longitudinal posterior probabilities dictate the final class and have mismatched cross-sectional posterior probabilities does not have any detrimental impact on the results of my study.

Fortunately, this disagreement should not happen for the majority of students. In my simulation study, I had 48 conditions with three attributes. I replicated each of these conditions 50 times, and each replication included 1,000 students. Thus, I generated three-attribute data for a total of 2,400,000 students. Of these 2,400,000 students, only 17 (less than 0.005%) had disagreements between their longitudinal MLCs and the class that corresponded with their maximum longitudinal posterior probability. So, it is important to realize that this disagreement is possible, but it is not a common occurrence.

Despite the benefits of the calibrate-and-score approach in its simplicity and faster estimation, this approach has some key limitations. In addition to the limitation related to the disagreement discussed above, the greatest limitation of the calibrate-and-score approach to estimation is related to invariance. In the full model approach with the same items used at each testing occasion, invariance is assumed, and the model estimates one set of item parameters across all testing occasions. In this situation, the model has information from student responses at all testing occasions to make very precise item parameter estimates. In other words, it has more data to use for estimation. In the

calibrate-and-score approach, invariance is also assumed, but item parameter estimation only uses student response data from T1. Therefore, in theory, the calibrate-and-score approach has less response data to use for item parameter estimation than the full model approach.

Determining Which Approach is Most Appropriate for My Simulation Study

I conducted preliminary analyses to evaluate the convergence rates for my simulation conditions using the full model and the calibrate-and-score approaches. I found that it was reasonable to use the full model approach for the one-attribute conditions because the estimation time was reasonable, and the convergence rates were high. However, for the three-attribute conditions, the convergence rates with the full model approach were significantly lower than the convergence rates with the calibrate-and-score approach. Therefore, I decided that I probably needed to use the calibrate-and-score approach to estimate and analyze the three-attribute conditions. Before proceeding with using the calibrate-and-score approach, I investigated whether the calibrate-and-score approach was appropriate for my simulation study.

In this investigation, I conducted a mini-analysis using one replication of one condition—the condition with three attributes, three proficiency statuses per attribute, large growth, attribute correlations of .75, and 12 higher-quality items per attribute. For this replication, I used the full model approach and the calibrate-and-score approach to estimate the model. The full model converged successfully after 10 hours and 23 minutes. The calibration for the pre-test converged after three hours. The scoring for the post-test finished running after only two seconds. I used the methods described in the previous section to obtain the longitudinal MLCs and posterior probabilities for the calibrate-and-

score approach. For this analysis, all students' longitudinal MLCs agreed with their classes that corresponded with their maximum longitudinal posterior probabilities.

For both approaches, I evaluated the item parameter estimation accuracy, student classification accuracy, and classification reliability. Unlike in my simulation study, for this analysis, I computed the item parameter estimation accuracy using the item parameters themselves, which are on the logit scale, rather than using the IRPs, which are on the probability scale. I chose this approach because I needed only to compare the accuracy across these two approaches and did not need to interpret the level of estimation accuracy—interpretation is easier for the probability scale than for the logit scale, but relative comparisons are equally easy with the logit and probabilities scales. I found that the full model approach had a mean average difference of 0.129 and a median average difference of 0.106 between the true and estimated item parameters. The calibrate-and-score approach had a mean average difference of 0.486 and a median average difference of 0.164 between the true and estimated item parameters. Therefore, the full model approach yielded greater item parameter estimation accuracy than the calibrate-and-score approach. The only other study that investigated the relative performance of these approaches (Madison et al., 2023) did not report on the item parameter recovery of the two approaches.

When I computed the CCCR across all students for both approaches, I found that the full model approach had an overall CCCR of .481, and the calibrate-and-score approach had an overall CCCR of .461. Additionally, the full model approach had a pre-test CCCR of .689 and a post-test CCCR of .724, while the calibrate-and-score approach had a pre-test CCCR of .675 and a post-test CCCR of .728. Therefore, the full model

approach yielded slightly greater student classification accuracy than the calibrate-and-score approach, except for the CCCR at the post-test, in which case the calibrate-and-score approach yielded better classification accuracy. My findings related to student classification accuracy are consistent with the findings in the study conducted by Madison and colleagues (2023).

I computed the polytomous polychoric reliability metric for each attribute for the full model approach and for the calibrate-and-score approach. The polychoric reliabilities for the full model approach for Attributes 1, 2, and 3 were .937, .937, and .947, respectively. The polychoric reliabilities for the calibrate-and-score approach for Attributes 1, 2, and 3 were .917, .901, and .920, respectively. These classification reliability values were all high, but the full model approach yielded reliabilities that were greater than the reliabilities from the calibrate-and-score approach. However, the decreases in reliability were slight, with an average decrease of .028 across all attributes, which was comparable to the findings from Madison and colleagues' (2023) study of the full model approach and calibrate-and-score approach.

I concluded that because (1) this is a simulation study in which I generated the data to be invariant, and (2) my results were consistent with the results from Madison and colleagues' (2023) study, in which they concluded that these two approaches were similar in terms of performance, it would be appropriate to use the calibrate-and-score approach for my three-attribute conditions. The calibrate-and-score approach yields results that would be generally comparable to results from the full model approach. In other words, the calibrate-and-score approach loses a small amount of precision with parameter estimation, but it does not change the values themselves or the interpretations of the

values. Therefore, for my simulation study, I used the full model approach for the one-attribute conditions, and I used the calibrate-and-score approach for the three-attribute conditions.

Summary of Simulation

With two levels for the item quality, two levels for the number of items to measure each attribute, three levels for the attribute proficiency growth, four levels for the attribute correlations, two levels for the number of proficiency statuses per attribute, and two levels for the number of attributes, a fully-crossed simulation study design would yield 192 conditions. However, for this dissertation, due to the time needed to estimate the full PTDCM for more complex designs (based on pilot studies I conducted in preparation for this dissertation), I excluded the conditions with three attributes and five proficiency statuses. Therefore, this simulation study included 144 generation conditions. I generated the data and conducted all analyses using R Statistical Software (v4.3.2, R Core Team, 2023). I estimated the PTDCM using the full model approach for the one-attribute conditions and the PDCM using the calibrate-and-score approach for the three-attribute conditions using Mplus Version 8 (Muthén & Muthén, 1998-2017).

Simulation Study Results

In this section, I present the results of the simulation study. First, I discuss the convergence rates across all conditions. Second, I discuss the results for the one-attribute conditions. Then, I discuss the results for the three-attribute conditions. For each set of conditions, I first present item parameter estimation accuracy results and student classification accuracy results. Finally, I present PTDCM reliability and DGP results. In the final section of this chapter, I discuss this simulation study. The tables and figures in

this section use a new variable that I call “Icat,” which is a combination of the number of items and the number of proficiency statuses per attribute. This variable allows for easier discussion of the simulation study conditions. For example, “I8_cat3” refers to the conditions with eight items and three proficiency statuses per attribute.

Convergence Rates

Table 12 shows the convergence rate for every condition. Figure 3 shows a histogram of the convergence rates for all conditions in the simulation study with an emphasis on the number of attributes for each condition. Overall, convergence rates ranged from .640 to 1, with a mean rate of .891 and a median rate of .900. Figure 3 shows that the one-attribute conditions generally had greater convergence rates than the three-attribute conditions, which aligned with my expectations because the three-attribute models were more complex and more difficult to estimate even with the calibrate-and-score approach. These data showed that 29 of the 96 one-attribute conditions and one of the 48 three-attribute conditions had convergence rates of 1. For the one-attribute conditions, the condition that had the lowest convergence rate of .720 was the condition with one attribute with five proficiency statuses measured by 12 items under the assumption of no growth and with a between-testing occasion attribute correlation of 0. For the three-attribute conditions, two conditions tied for the lowest convergence rate of .640. These were the conditions with three attributes with three proficiency statuses measured by eight items under the assumption of large growth and attribute correlations of .25 and .5.

If we break the rates down in terms of the number of attributes and the item quality, we see that convergence rates for the one-attribute conditions with lower-quality

items ranged from .740 to 1, with a mean rate of .920 and a median rate of .940. Convergence rates for the one-attribute conditions with higher-quality items ranged from .720 to 1, with a mean rate of .929 and a median rate of .970. Convergence rates for the three-attribute conditions with lower-quality items ranged from .640 to .880, with a mean rate of .770 and a median rate of .780. Convergence rates for the three-attribute conditions with higher-quality items ranged from .680 to 1, with a mean rate of .876 and a median rate of .900. Therefore, convergence rates improved with higher-quality items.

In general, the conditions with five proficiency statuses had lower convergence rates than the conditions with three proficiency statuses, which was expected because increasing the number of proficiency statuses increases the model and estimation complexity. Additionally, increasing the number of items that measured each attribute or increasing the item quality resulted in increased convergence rates, in general. The greatest convergence rates occurred for conditions with one attribute with three proficiency statuses measured by 12 items under the assumption of moderate growth. In sum, although the convergence rates varied, they were all acceptably high.

Results for the One-Attribute Conditions

Item Parameter Estimation Accuracy Results for the One-Attribute Conditions

Table 13 shows the average MAD values for each type of IRP for each condition. Figure 4 displays the data in Table 13. For the MAD data, smaller values indicate greater estimation accuracy. Recall that the conditions with three proficiency statuses have three IRPs, and the conditions with five proficiency statuses have five IRPs. I computed the MAD values for the different IRPs separately and together (the “All” IRP type). In Table 13 and Figure 4, I used “IRP0”, for example, to denote the IRP for the student group with

the proficiency status [0]. Figure 4 shows several key trends for the manipulated factors in this simulation study and for the different types of IRPs.

First, in general, the estimation accuracy across all conditions was acceptable, with many MAD values less than .10 for all but some values for IRP3 and IRP4 and values within the moderate- and large-growth conditions. Even the greatest average MAD value was .170 (for IRP2 in the condition with 12 high-quality items measuring an attribute with five proficiency statuses under large growth and with an attribute correlation of 0), which was not so extreme as to cause serious concern about item parameter estimation accuracy, but this maximum value was an outlier. The average and median MAD values across all conditions and all IRP types were .057 and .063, respectively. This mean of .057 can be interpreted to mean that, on average, IRPs were estimated to be within plus or minus 5.7% of the original value. Therefore, in general, the IRP and the item parameters were generally well recovered. However, the manipulated factors in the simulation study did impact the item parameter recovery: The conditions with the greatest item parameter estimation accuracy (i.e., the lowest MAD values) for all IRP types averaged together were the conditions with 12 higher-quality items measuring an attribute with three proficiency statuses. The conditions with the lowest item parameter estimation accuracy (i.e., the greatest MAD values) for all IRP types averaged together were the conditions with eight lower-quality items measuring an attribute with five proficiency statuses.

On average, across all simulation conditions, IRP0 was the IRP type that had the greatest estimation accuracy with a mean value of .021 compared with mean values of .077, .068, .092, and .070 for IRP1, IRP2, IRP3, and IRP4, respectively. This result was

expected because the IRP0s depend on only one item parameter, while the other IRP types depend on multiple item parameters. It was interesting to find that the IRP type with the worst estimation accuracy on average was IRP3, which depended on four item parameters, and not IRP4, which depended on five item parameters. Further, in some situations, IRP1 had worse estimation accuracy than IRP2, IRP3, and IRP4. To support these results, I computed the MAD values for the item parameters themselves and found the general trend that the intercept parameter was estimated with the greatest accuracy, followed by Main Effect 1 (the main effect for the students with the proficiency status [1]), then Main Effect 2, then Main Effect 3, and Main Effect 4 generally had the worst estimation accuracy, as expected. In future research, I plan to investigate more into how the accuracy for the IRP types was not ordered as expected while the accuracy for the item parameters was ordered as expected. The MAD values for the item parameters were similarly influenced by the manipulated factors, as were the MAD values for the IRPs described in this section.

In terms of the manipulated factors in the simulation study, the average MAD values for conditions with three and five proficiency statuses were .037 and .077 (average difference of .040), respectively, so as the number of proficiency statuses increased, the MAD values increased, and the item parameter estimation accuracy decreased. The average MAD values for conditions with lower- and higher-quality items were .066 and .055 (average difference of .011), respectively, so as the item quality increased, the MAD values decreased, and the item parameter estimation accuracy increased. The average MAD values for conditions with eight and 12 items per attribute were .063 and .058 (average difference of .005), respectively, so as the number of items per attribute

increased, the MAD values decreased, and the item parameter estimation accuracy increased. Further, the average MAD values for conditions with I12_cat3, I8_cat3, I12_cat5, and I8_cat5 were .032, .041, .075, and .078, respectively. Therefore, as the number of proficiency statuses decreased and the number of items per attribute increased, the average MAD values decreased, and the item parameter estimation accuracy increased. Again, these findings were expected because having more items to measure less complex attributes results in more efficient estimation. To summarize, the average differences in MAD values were greater when only the number of proficiency statuses varied (.040) than when only item quality varied (.011), which were greater than when only the number of items varied (.005). Therefore, MAD values decreased more when the number of proficiency statuses decreased than when the item quality increased or the number of items increased, but the MAD values decreased more when the item quality increased than when the number of items increased.

Across all conditions, growth did not show a consistent pattern, with average MAD values of .060, .055, and .067 for the no-growth, moderate-growth, and large-growth conditions. However, within each Icat type, the no-growth conditions showed more consistent and lower MAD values than the moderate-growth conditions, which showed more consistent and lower MAD values than the large-growth conditions, as evidenced by the increasing spread of the lines as you move across the growth panels in Figure 4. The consistency of the MAD values was further impacted by the number of proficiency statuses. As the growth and number of proficiency statuses both increased, the consistency of the MAD values decreased, resulting in a wider distribution of MAD values in the large-growth panel of Figure 4 than in the no-growth and moderate-growth

panels. In terms of attribute correlations, the average MAD values for conditions with attribute correlations of 0, .25, .5, and .75 were .063, .062, .061, .056, with average differences of .001, .001, and .005, respectively. Therefore, the conditions with greater attribute correlations yielded slightly lower MAD values and, therefore, greater item parameter estimation accuracy on average.

Student Classification Accuracy Results for the One-Attribute Conditions

Table 14 shows the CCR data for the one-attribute conditions for the simulation study, and Figure 5 displays these data. These data include three types of CCRs: attribute level at the pre-test (ACCR Pre), attribute level at the post-test (ACCR Post), and latent class or profile level (CCCR). In general, the CCRs across all conditions were lower than in typical DCM simulation studies, which often report classification accuracies greater than .9 (e.g., de la Torre, 2009; Madison & Bradshaw, 2018). This finding of lower classification accuracy with the PTDCM than with the TDCM with dichotomous attributes was consistent with previous PTDCM simulation study results with similar numbers of items and proficiency statuses per attribute (Madison & Bao, 2018).

In general, the conditions that showed high attribute-level classification accuracy (greater than about .8) were the conditions with 12 higher-quality items measuring an attribute with three proficiency statuses, and of these conditions, the ones with greater attribute correlations generally showed greater classification accuracy. Under moderate growth and/or with greater attribute correlations, some conditions with only eight higher-quality items measuring an attribute with three proficiency statuses also achieved attribute-level classification accuracy values greater than .8. In Table 14, these highly accurate conditions are represented in bold font. The conditions that showed the lowest

attribute-level classification accuracy were the conditions with eight lower-quality items measuring an attribute with five proficiency statuses. These trends were expected because the highly accurate conditions had simpler models and more information than the models in the more complex conditions.

For this simulation study, the CCRs ranged from .164 to .877, with mean and median values of .559 and .564, respectively. Splitting the results by CCR type, the ACCRs (averaged across the pre- and post-test values) ranged from .319 to .877, with mean and median values of .627 and .613, respectively, and the CCCRs ranged from .164 to .771, with mean and median values of .422 and .398, respectively. This finding that the ACCR values were greater than the CCCR values was also consistent with previous DCM literature (e.g., Schellman, 2021). It was expected that attribute-level classification accuracy would be greater than class-level classification accuracy because attribute-level classification accuracy requires matching the true and estimated proficiency status for only one attribute and one testing occasion at a time, while class-level accuracy requires matching the true and estimated proficiency statuses for multiple attributes and testing occasions simultaneously because the whole profile must be accurately recovered. Additionally, Figure 5 shows that the attribute-level proficiency statuses at the pre-test, which had ACCR values ranging from .481 to .873, with mean and median values of .660 and .642, were recovered more accurately than the attribute-level proficiency statuses at the post-test, which had ACCR values ranging from .319 to .877, with mean and median values of .593 and .607. This finding was similar to findings from previous studies (Madison & Bradshaw, 2018; Yu et al., 2023), in which the post-test classifications were

more accurate than the pre-test classifications. This phenomena of classification accuracy being different for different testing occasions requires more research.

On average, as growth increased, CCRs decreased slightly, with average CCRs of .595, .541, and .540 (average differences of .054 and .001, respectively) and median CCRs of .593, .564, and .546 for the no-growth, moderate-growth, and large-growth conditions. The effect of growth on the CCRs showed a few key patterns related to consistency, as seen in Figure 5. First, when growth was low, the CCRs were closer together with a smaller range on the zero-to-one scale (a range of .634 for the no-growth conditions versus a range of .713 for the large-growth conditions). The range of the CCRs increased as the growth increased because the CCCR and ACCR Post values decreased, while the ACCR Pre values remained about the same across growth conditions. When growth was low, the ACCR values for the pre-test and post-test were nearly identical, while the CCCR values were more distinct. When growth was low, the CCCR values were more influenced by the attribute correlations than the ACCR values. Here, the greater attribute correlations corresponded with greater CCCR values. When the growth was larger, the ACCR values for the pre-test and post-test were much more distinct, and the CCCR values were less influenced by the attribute correlation values because the lines within each CCR type in the large-growth panel in Figure 5 were more compact, especially for the conditions with higher-quality items. However, even with large growth when attribute correlation had less of an impact on CCRs, the trend that larger attribute correlations correspond with larger CCRs held: Across all conditions, the attribute correlations of 0, .25, .5, and .75 yielded average CCR values of .546, .551, .559, and .579, with average differences of .005, .008, and .020, respectively.

Figure 5 also shows that, as expected, when the number of proficiency statuses decreased, the item quality increased, and/or the number of items per attribute increased, the CCR values increased because the model was less complex and easier to estimate. The impact of the number of proficiency statuses was greater than the impact of the item quality, which was greater than the impact of the number of items. Specifically, the average CCR values (averaged across the pre- and post-test ACCRs and the class-level CCRs) for the conditions with three and five proficiency statuses were .700 and .417 (average difference of .283), respectively, the average CCR values for the conditions with lower- and higher-quality items were .498 and .619 (average difference of .121), respectively, and the average CCR values for the conditions with eight and 12 items per attribute were .536 and .581 (average difference of .045), respectively. When considering the number of items and number of proficiency statuses together, the average CCR values for the conditions with I12_cat3, I8_cat3, I12_cat5, and I8_cat5 were .733, .666, .430, and .405, respectively. To summarize, the average differences in CCR values were greater when only the number of proficiency statuses varied (.283) than when only the item quality varied (.121), which were greater than when only the number of items varied (.046), which were greater than when only the attribute correlation varied (.004, .009, and .020). Therefore, CCR values increased more when the number of proficiency statuses decreased than when the item quality increased or the number of items increased.

PTDCM Reliability Results for the One-Attribute Conditions

Table 15 shows the reliabilities for the polytomous polychoric, polytomous average maximum transition, and polytomous PB reliability metrics for the one-attribute conditions in this simulation study. Figure 6 displays this data and shows several trends

across the reliability metrics and across the manipulated simulation study factors that were analogous to those trends described above with item parameter estimation and student classification accuracy. Figures 7-9 show the same results but with separate plots for each reliability metric and with the item quality values on the same plots so the influence of the manipulated factors on each reliability metric can be better investigated. I chose to separate the plots for the reliability results but not for the item parameter estimation accuracy and student classification accuracy results because reliability was the focus of this simulation study.

First, as with previous simulation studies using the DCM and TDCM with dichotomous attributes (Johnson & Sinharay, 2020; Schellman & Madison, in press), the polychoric metric showed the greatest levels of reliability (.729 on average), followed by the average maximum transition metric (.612 on average), followed by the PB metric (.315 on average). Figure 6 shows that the range of polychoric reliabilities (.398) was greater than the range of average maximum transition reliabilities (.325), which was less than the range of PB reliabilities (.376), but the PB metric had the most compact lines in the figure, so it was less influenced by the attribute correlation than the polychoric and average maximum transition metrics. The reliabilities in this simulation study were generally lower than the reliabilities for other DCM and TDCM simulation studies that use dichotomous attributes. However, the reliabilities in this simulation study were consistent with the reliabilities in the previous PTDCM simulation study (Madison & Bao, 2018), which reported a minimum polytomous polychoric reliability greater than .7 for attributes with three proficiency statuses and four to 16 items per attribute. In my simulation study, the minimum average polytomous polychoric reliability for the

conditions with three proficiency statuses and higher-quality items was .731. In Table 15, I flagged the reliability values that can be considered acceptable, good, very good, or excellent based on the recommendations from Schellman and Madison (in press), and all of the reliability values for the average maximum transition or PB metrics were considered less than acceptable based on these criteria. The polychoric metric had some values that were acceptable, good, or very good (generally, for conditions with higher-quality items with less complex Q-matrices), but no values were considered to be excellent. It is possible that, because these were polytomous attributes that had greater data requirements, increasing the number of items per attribute or the sample size would help get these reliability values to more desirable thresholds.

As with the item parameter estimation accuracy and student classification accuracy, within each reliability metric, the conditions with the greatest reliabilities were the conditions with 12 higher-quality items measuring an attribute with three proficiency statuses, and of these conditions, the conditions with greater attribute correlations generally showed greater reliabilities. The conditions that showed the lowest reliabilities were the conditions with eight lower-quality items measuring an attribute with five proficiency statuses. Again, these trends were expected because the conditions with high reliability had simpler models and more information than the conditions with low reliability.

Figures 6-9 show that increasing the number of items, increasing the item quality, and/or decreasing the number of proficiency statuses per attribute resulted in greater reliability. However, for reliability, the influence of the item quality was greater than the influence of the number of proficiency statuses, which was greater than the influence of

the number of items. For item parameter estimation accuracy and classification accuracy, the number of proficiency statuses had the greatest influence, so it was different to find that the item quality had the greatest influence on reliability. Specifically, the average reliabilities for conditions with three and five proficiency statuses were .601 and .504 (average difference of .097), respectively, the average reliabilities for conditions with lower- and higher-quality items were .495 and .610 (average difference of .115), respectively, the average reliabilities for conditions with eight and 12 items per attribute were .517 and .587 (average difference of .070), respectively, and the average reliabilities for conditions with I12_cat3, I8_cat3, I12_cat5, and I8_cat5 were .639, .535, 562, and .472, respectively. To summarize, the average differences in reliability values were greater when only the item quality varied (.115) than when only the number of proficiency statuses varied (.097), which were greater than when only the number of items varied (.070). Therefore, reliability values increased much more when the item quality increased than when the number of proficiency statuses decreased or the number of items increased.

Figures 7-9 do not all use the same scale points on the vertical axes, but they all use the same range (.45), so the scale holds true across these plots and the relative slopes and visual differences can be compared across the plots.

In terms of growth, Figure 7 shows that growth did not have a consistent impact on the polychoric metric because the no-growth conditions showed greater reliabilities than the large-growth conditions, which showed greater reliabilities than the moderate-growth conditions. For the conditions with lower-quality items, the reliability values dropped drastically when moving from no growth to moderate growth, but the values

increased again when moving to large growth. However, these differences were not as significant for the conditions with higher-quality items. Figure 7 also shows that the lines for the attribute correlation did not cross, so greater attribute correlations coincided with greater polychoric reliabilities—this finding was in line with expectations. Specifically, on average, across all conditions and reliability metrics, the reliability values for conditions with attribute correlations of 0, .25, .5, and .75 were .536, .545, .557, and .571, with average differences of .009, .012, and .014, respectively, so the influence of attribute correlation held across all conditions. In terms of item quality, the moderate-growth conditions were the conditions that were most influenced by decreased item quality, as evidenced by the separation between reliabilities for the conditions with higher and lower-quality items.

Figure 8 shows different results for the average maximum transition metric in comparison with the polychoric metric in Figure 7. Most notably, the slopes in Figure 8 were much sharper than the slopes in Figure 7 because the average maximum transition reliability metric was more heavily influenced by the number of proficiency statuses than the average polychoric reliability metric—when the number of proficiency statuses increased, the average maximum transition reliability decreased much more than the average polychoric reliability. As with the polychoric reliability, the average maximum transition reliabilities decreased as the growth increased, specifically when moving from no growth to moderate growth, but the moderate- and large-growth conditions had comparable average maximum transition reliabilities, especially for the conditions with lower-quality items. Additionally, as growth increased, the influence of the attribute correlation decreased, as evidenced by the compact lines in the large-growth panel of

Figure 8. Specifically, when the growth was large and the model was complex (eight lower-quality items measuring an attribute with three proficiency statuses), the condition with an attribute correlation of 0 had greater average maximum transition reliability than the conditions with attribute correlations of .25, .5, and .75. Therefore, the average maximum transition reliability metric was generally not as influenced by the attribute correlations as was the polychoric reliability metric, especially when growth was large. Finally, the item quality was least influential for the no-growth conditions, as evidenced by the overlapping lines in the no-growth panel in Figure 8. As growth increased, the conditions with lower-quality items had consistently lower average maximum transition reliability values.

Figure 9 shows the results for the average PB metric, which were different from the results for the polychoric and average maximum transition metrics. Like the average maximum transition metric, the slopes in Figure 9 were much sharper than the slopes in Figure 7 because the PB reliability metric was more heavily influenced by the number of proficiency statuses than the average polychoric reliability metric. The PB metric showed similar trends as the other two metrics in terms of growth. However, the PB metric appears to not have been as influenced by the attribute correlation as were the other two metrics because, within all levels of growth and item quality, the lines in Figure 9 were very compact and almost identical across the levels of attribute correlation. Finally, unlike the polychoric and average maximum transition metric, the PB metric was consistently influenced by item quality across all levels of growth, as evidenced by the separation of the values for conditions with lower- and higher-quality items in Figure 9.

Finally, Table 16 and Figure 10 show the average proportions of maximum posterior probabilities for the one-attribute conditions in the simulation study that were greater than .1, .2, .3, …, and .9. As mentioned previously, the greater a student's maximum posterior probability, the more confident the PTDCM is in their classification, so the greater the proportion of students' maximum posterior probabilities above specific values on the zero-to-one scale, the more confident the PTDCM is across all student classifications. This analysis is directly related to the average maximum transition metric because it depends on students' maximum posterior probabilities. Therefore, the better the results from this analysis, the more trustworthy the results from the reliability metrics, especially the average maximum transition metric, because the proportion of maximum posterior probabilities greater than a specific value is most highly correlated with the average maximum transition metric, but it is positively correlated with all of the reliability metrics.

Figure 10 shows that the proportions of maximum transitions greater than checkpoints on the zero-to-one scale decreased at a slower rate for the conditions with more items and fewer proficiency statuses, especially for the conditions with higher-quality items. For the conditions with lower-quality items, as growth increased, the number of items per attribute appears to have had less of an impact, but the number of proficiency statuses was still influential. Additionally, the conditions with no growth had less steep slopes than the conditions with moderate and large growth. Thus, increasing growth resulted in decreased confidence in the student classifications from the PTDCM. In terms of attribute correlations, the conditions with attribute correlations of .75 showed the greatest confidence in student classifications for each set of items and proficiency

statuses, but the other levels of attribute correlations showed comparable confidence. The influence of the attribute correlation decreased as the growth increased.

DGP Results for the One-Attribute Conditions

To summarize the DGP results from the simulation study, I computed the average minimum, mean, median, and maximum of each type of DGP for each condition across all replications. Tables 17-20 and Figures 11-14 show these aggregated statistics for each condition. I chose to aggregate the minimum, mean, median, and maximum DGPs across all replications of each condition so I could better discuss how the manipulated simulation study factors and the types of DGPs impacted each statistic.

Table 17 and Figure 11 show the average minimum DGPs for each condition. The average minimum DGP results across the DGP types behaved as expected. First, the CWP DGP had an average minimum value of 0 across all conditions. Second, the average minimum values for the basic DGPs were all close to 0, with an average value of .021. Finally, the DGP with the greatest average minimum values was the CW DGP, with an average mean minimum value of .110. This finding makes sense because, in general, the CW DGP metric increases the DGPs for students who had a low basic DGP.

The number of proficiency statuses had a slightly stronger impact on the average minimum DGP values than the item quality and the number of items. Conditions with three and five proficiency statuses had mean minimum DGP values of .040 and .026 (average difference of .014), respectively. On average, as the item quality increased, the average minimum DGPs decreased slightly, with mean values of .035 and .031 (average difference of .004) for conditions with lower- and higher-quality items, respectively. On average, as the number of items per attribute increased, the average minimum DGPs

decreased slightly, with mean values of .035 and .031 (average difference of .004) for conditions with eight and 12 items, respectively. On the other hand, the attribute correlations had a slightly stronger impact on the average minimum DGP values than the number of proficiency statuses, with mean values of .052, .035, .027, and .020 (average differences of .017, .008, and .007, respectively) for conditions with attribute correlations of 0, .25, .5, and .75, respectively. As the growth increased, the average minimum DGPs decreased—this decrease was more significant when moving from no growth to moderate growth than when moving from moderate growth to large growth: the average minimum DGPs had mean values of .065, .020, and .015 (average differences of .045 and .005, respectively) for conditions with no growth, moderate growth, and large growth, respectively Additionally, as growth increased, the average minimum DGPs got more consistent across conditions, as evidenced by the more compact lines in the large-growth panel of Figure 11. The ranges of average minimum DGP values were .472, .154, and .114 for the no-growth, moderate-growth, and large-growth conditions. In sum, the minimum DGPs behaved as expected across the simulation conditions.

Table 18 and Figure 12 show the average mean DGPs for each condition. Figure 12 shows greater distinction between the DGP types than Figure 11. In general, the basic DGP had the greatest average mean DGPs, with a mean value of .769, followed by the CW DGPs, with a mean value of .743, then the PW DGPs, with a mean value of .674, and then the CWP DGPs, with a mean value of .634, as expected. In terms of growth, it is not helpful to consider the average mean DGP values for each level of growth because as growth increased, the basic and CW DGPs decreased, but the CWP and PW DGPs decreased, so the average mean DGP values for each level of growth were not very

different (.693, .714, and .709 for the no-growth, moderate-growth, and large-growth conditions, respectively). It is more helpful to consider the ranges of the average mean DGP values within each level of growth. The average mean plots showed similar patterns as the average minimum plots: as growth increased, the average mean DGPs got more consistent with ranges of .474, .310, and .224 for the conditions with no growth, moderate growth, and large growth, respectively. This trend means that as growth increased, the influence of the other manipulated simulation study factors decreased.

Even when growth was large, the other manipulated factors had some influence over the average mean DGP values: the average mean DGPs were greater for the conditions with fewer proficiency statuses, lower for the conditions with higher-quality items, slightly greater for the conditions with fewer items per attribute, and greater for the conditions with greater attribute correlations. On average, the mean DGPs for conditions with three and five proficiency statuses were .731 and .679 (average difference of .052), respectively. On average, the mean DGPs for conditions with lower and higher item quality were .719 and 691 (average difference of .028), respectively. The number of items did not appear to have had a strong influence on the average mean DGPs, with average mean DGPs of .709 and .701 (average difference of .008) for conditions with eight and 12 items per attribute, respectively. The average mean DGPs were less influenced by the attribute correlations than the average minimum DGPs, with average mean DGP values of .664, .682, .715, and .760 (average differences of .018, .033, and .045, respectively) for conditions with attribute correlations of 0, .25, .5, and .75, respectively. Therefore, the average mean DGPs were most impacted by the number of

proficiency statuses, followed by the attribute correlations and the item quality, and they were least impacted by the number of items per attribute.

Table 19 and Figure 13 show the average median DGPs for each condition. Figure 13 also shows patterns similar to those in Figure 12. However, the average median basic DGPs were more distinct from the other DGP metrics. This was due to the fact that many students' basic DGPs were 1 because they achieved the maximum proficiency status by the post-test regardless of their proficiency statuses at the pre-test, especially for the large-growth conditions. The influence of the DGP type on the average median DGPs was the same as with the average mean DGPs: In general, the basic DGP had the greatest average median DGPs, with a mean value of .864, followed by the CW DGPs, with a mean value of .791, then the PW DGPs, with a mean value of .751, and then the CWP DGPs, with a mean value of .728, as expected. In terms of growth, on average, the conditions with no growth, moderate growth, and large growth had average median DGP values of .744, .786, and .821, with average differences of .042 and .035, respectively. As with the average minimum and mean DGPs, as growth increased, the consistency of the average median DGPs decreased with ranges of .480, .385, and .354 for the no-growth, moderate-growth, and large-growth conditions, respectively.

The influence of the other manipulated factors on the average median DGP values had the same trends as with the average mean DGP values: the average median DGPs were greater for the conditions with fewer proficiency statuses, lower for the conditions with higher-quality items, slightly higher for the conditions with fewer items per attribute, and greater for the conditions with greater attribute correlations. On average, the median DGPs for conditions with three and five proficiency statuses were .828 and

.739 (average difference of .089), respectively. On average, the median DGPs for conditions with lower and higher item quality were .794 and .773 (average difference of .021), respectively. The number of items did not appear to have had a strong influence on the average median DGPs, with average mean DGPs of .787 and .780 (average difference of .007) for conditions with eight and 12 items per attribute, respectively. On average, the median DGP values were .732, .757, .799, and .846 (average differences of .025, .042, and .047, respectively) for conditions with attribute correlations of 0, .25, .5, and .75, respectively. Therefore, as with the average mean DGPs, the average median DGPs were most impacted by the number of proficiency statuses, followed by the attribute correlations and the item quality, and they were least impacted by the number of items per attribute.

Table 20 and Figure 14 show the average maximum DGPs for each condition. The average maximum DGPs were largely equal to 1 across all conditions. In fact, only 113 out of the 384 (29.4%) conditions in Table 20 had average maximum values less than 1, and none of these 113 values were for the basic DGP: 27 were for the CW DGP, 44 were for the CWP DGP, and the other 42 were for the PW DGP. Figure 14 shows that the average maximum DGPs were specifically impacted by the number of proficiency statuses for the attribute, as the majority of the average maximum DGPs that were not equal to 1 occurred for conditions with a greater number of proficiency statuses. However, the conditions with eight higher-quality items measuring an attribute with three proficiency statuses under moderate growth showed the lowest average maximum DGP values, specifically for the CW and PW DGPs. Table 20 and Figure 14 did not show meaningful patterns of influence for growth, the item quality, the number of items, or the

attribute correlations, as all average mean maximum DGP values were greater than .993 across all manipulated factors, and the average maximum DGP values within each level of each manipulated factor were all greater than .999, so I exclude interpretation for the influence of the other manipulated factors on the average maximum DGPs.

Results for Three-Attribute Conditions

Recall that this simulation study did not include any conditions with three attributes and five proficiency statuses because such a complex model is difficult to estimate. Therefore, I cannot comment on the influence of the number of proficiency statuses for the three-attribute conditions because they all used three proficiency statuses per attribute. Additionally, although the calibrate-and-score approach that I used for the three-attribute conditions yields results that were comparable to the full model approach that I used for the one-attribute conditions, I did not directly compare the one-attribute and three-attribute conditions, so the modeling approach itself could not be considered a confounding variable.

Item Parameter Estimation Accuracy Results for the Three-Attribute Conditions

Table 21 and Figure 15 show the average MAD values between the true and estimated IRPs for the three-attribute conditions in the simulation study. As with the one-attribute conditions, IRP0 had the lowest MAD values (.021, on average, compared with .074 and .060 for IRP1 and IRP2, respectively) and, therefore, the greatest estimation accuracy, which was expected because the accuracy of the estimation of IRP0 depended on only one item parameter. In general, the IRP2s were estimated more accurately than the IRP1s, which was interesting because the IRP2s depended on more item parameters than the IRP1s. However, the mean and median MAD values across all conditions and

IRP types were .052 and .056, respectively. Therefore, the item parameters were generally well-recovered for the three-attribute conditions. As with the one-attribute conditions, to support these results, I computed the MAD values for the item parameters themselves and found the general trend that the intercept parameter was estimated with the greatest accuracy, and Main Effect 2 had the worst estimation accuracy, as expected. In future research, I plan to investigate more into how the accuracy for the IRP types was not ordered as expected while the accuracy for the item parameters was ordered as expected. The MAD values for the item parameters were similarly influenced by the manipulated factors, as were the MAD values for the IRPs described in this section.

In terms of the manipulated factors in the simulation study, Figure 15 shows no apparent impact of growth on item parameter estimation accuracy—the average MAD values for each level of growth were .051, .051, and .052 (average differences of 0 and .001, respectively) for the no-growth, moderate-growth, and large-growth conditions, respectively. However, the item quality, the number of items per attribute, and the attribute correlations did impact item parameter estimation accuracy. As the item quality increased, the average MAD values decreased because conditions with lower- and higher-quality items had average MAD values of .061 and .042 (average difference of .019), respectively. As the number of items per attribute increased, the average MAD values decreased, so estimation accuracy increased. Specifically, conditions with eight and 12 items per attribute had average MAD values of .058 and .045 (average difference of .013), respectively. Additionally, as the attribute correlations increased, the average MAD values decreased because conditions with attribute correlations of 0, .25, .5, and .75 had average MAD values of .055, .054, .051, and .046, with average differences of

.001, .003, and .005, respectively, across all conditions. Some of the attribute correlation lines within IRP type, growth level, and item quality in Figure 15 did cross, so this trend was not consistent within all subsets of conditions. Therefore, increasing item quality had a greater impact on the MAD values than increasing the number of items, which had a greater impact on the MAD values than increasing the attribute correlation.

Student Classification Accuracy Results for the Three-Attribute Conditions

Table 22 and Figure 16 show the average CCR values for the three-attribute conditions. Unlike the one-attribute conditions, the three-attribute conditions had CCR values for the *cross-sectional* (i.e., one testing occasion; not longitudinal) pre- and post-test latent classes, which were different from the ACCR values. For the one-attribute conditions, the ACCR Pre and ACCR Post values were the same as the CCR Pre and CCR Post values, respectively. I compared these estimated classes with students' true classes and individual attribute proficiency statuses. For the three-attribute conditions, I averaged the ACCR values across all attributes. In general, the results were as expected because the average CCR values were the lowest for the CCCRs, which had an average CCR value of .218, and greatest for the ACCRs, which had average CCR values of .759 and .716 for the ACCR Pre and ACCR Post values, respectively. The average CCR Pre value was .475, and the average CCR Post value was .407.

For the no-growth conditions, the average ACCR Pre and ACCR Post values were nearly identical, and the CCR Pre and CCR Post values were nearly identical, as evidenced by the overlapping lines in the no-growth panel of Figure 16. However, as growth increased, the ACCR Pre and CCR Pre values did not change much, but the ACCR Post and CCR Post values decreased. Growth appears to have not had any other

consistent impact on the average CCR values—the average CCR values across all conditions with no growth, moderate growth, and large growth had mean average CCR values of .547, .490, and .508, respectively.

As the item quality increased, the number of items increased, and/or the attribute correlations increased, the CCR values increased, except for some of the ACCR Post and CCR Post conditions with large growth and lower-quality items, which had negative slopes in Figure 16. On average, the conditions with lower- and higher-quality items had CCR values of .425 and .605 (average difference of .180), respectively. On average, conditions with eight and 12 items per attribute had CCR values of .473 and .557 (average difference of .084), respectively. On average, conditions with attribute correlations of 0, .25, .5, and .75 had CCR values of .478, .492, .519, and .571, with average differences of .014, .027, and .052, respectively. Therefore, like the MAD values discussed above, increasing item quality had a greater impact on the CCR values than increasing the number of items, which had a greater impact on the CCR values than increasing the attribute correlation.

PTDCM Reliability Results for the Three-Attribute Conditions

Table 23 and Figure 17 show the average reliability metrics for the three-attribute conditions. Figures 18-20 show the same data but with a separate plot for each reliability metric so the patterns of the impact of the manipulated factors can be more closely inspected. As in the one-attribute conditions, the polytomous polychoric metric had the greatest values, with an average value of .767, followed by the polytomous average maximum transition metric, with an average value of .690, followed by the polytomous PB metric, with an average value of .459. In Table 23, I flagged the reliability values that

can be considered acceptable, good, very good, or excellent based on the recommendations from Schellman and Madison (in press), and like the one-attribute conditions, all of the reliability values for the average maximum transition or PB metrics were considered less than acceptable based on these criteria. The polychoric metric had some values that were acceptable, good, or very good (generally, for conditions with higher-quality items with less complex Q-matrices), but no values were considered to be excellent.

As with item parameter estimation accuracy and student classification accuracy, growth did not have a consistent impact on classification reliability with average reliabilities of .652, .620, and .645 for the no-growth, moderate-growth, and large-growth conditions, respectively. Also, like the other evaluation metrics, on average, the reliabilities increased when the item quality increased, the number of items per attribute increased, and/or the attribute correlation increased. On average, the reliabilities for the conditions with lower- and higher-quality items were .573 and .704 (average difference of .131), respectively. On average, the reliabilities for the conditions with eight and 12 items per attribute were .602 and .676 (average difference of .074), respectively. On average, the reliabilities for the conditions with attribute correlations of 0, .25, .5, and .75 were .614, .626, .644, and .672, with average differences of .012, .018, and .028, respectively. Therefore, like the MAD and CCR values discussed above, increasing item quality had a greater impact on the reliabilities than increasing the number of items, which had a greater impact on the reliabilities than increasing the attribute correlation.

As with the one-attribute conditions, although Figures 18-20 do not use the same vertical scale points, the ranges of the scales are identical, so the relative scales of Figures

18-20 are identical, and visual comparisons can be relied upon. Figures 18-20 highlight the strictly increasing lines across all three reliability metrics, indicating that as the number of items per attribute increased, the reliabilities increased. Although Figures 18-20 show some crossing lines, these figures generally illustrate the direct relationship between the attribute correlation and reliability. In comparing Figures 18-20, it appears that the polychoric metric was less influenced by item quality than the other two metrics because there was less separation between the lines for lower- and higher-quality items in Figure 18 than in Figures 19 and 20. Additionally, the range of the polychoric metric values (.357) was greater than the ranges for the average maximum transition and PB metrics (.224 and .303, respectively). Finally, the polychoric metric appears to have been more influenced by the attribute correlation than the other two metrics, as evidenced by the clear separation and spacing between the lines corresponding to the different attribute correlations in Figure 18. In Figures 19 and 20, the attribute correlation lines were less distinct than they were in Figure 18.

Table 24 and Figure 21 show the average proportions of maximum posterior probabilities for the three-attribute conditions in the simulation study that were greater than .1, .2, .3, …, and .9. Figure 21 shows that, in general, the conditions with higher-quality items, 12 items per attribute, and/or greater attribute correlations had slopes that decreased at slower rates than the conditions with lower-quality items, eight items per attribute, and/or lower attribute correlations, which means that the model was more confident in student classifications when the item quality was greater, the number of items was greater, and/or the attribute correlations were greater. Across the conditions with lower-quality items, about 50% of students had maximum transitions that were

greater than .55 to .75, and across the conditions with higher-quality items, about 50% of students had maximum transitions that were greater than .7 to .85, which means that the PTDCM was, in general, more confident in student classifications for the conditions with higher-quality items than for the conditions with lower-quality items.

DGP Results for the Three-Attribute Conditions

In this section, I obtained DGP summary statistics (average minimum, mean, median, and maximum) results for each attribute separately, but I discuss the results for only Attribute 1 and the entire attribute profile (the averages of the attribute-level DGPs across all three attributes) because I generated all attributes the same way, so there should not be meaningful differences between the results for the different attributes. However, I included tables and figures to show the results for all attributes. Tables 25-28 and Figures 22-25 show the average minimum DGP values for Attribute 1, Attribute 2, Attribute 3, and the entire attribute profile, respectively. Tables 29-32 and Figures 26-29 show the average mean DGP values for Attribute 1, Attribute 2, Attribute 3, and the entire attribute profile, respectively. Tables 33-36 and Figures 30-33 show the average median DGP values for Attribute 1, Attribute 2, Attribute 3, and the entire attribute profile, respectively. Tables 37-40 and Figures 34-37 show the average maximum DGP values for Attribute 1, Attribute 2, Attribute 3, and the entire attribute profile, respectively.

Let us discuss the average minimum DGPs for Attribute 1 and the attribute profile. Figure 22 shows similar trends as in Figure 11 (the average minimum DGPs for the one-attribute conditions). First, the CWP DGP had an average minimum value of 0 across all conditions. Second, the average minimum values for the basic DGPs were all close to 0, with an average value of .088. Finally, the DGP with the greatest average

minimum values was the CW DGP, with an average mean minimum value of .135. Growth did not have a consistent impact on the average minimum DGPs for Attribute 1, with values of .130, .023, and .015 for the no-growth, moderate-growth, and large-growth conditions, respectively. However, like the one-attribute conditions, as the growth increased, the average minimum DGPs got more consistent, as evidenced by the overall ranges of the average minimum DGPs within each level of growth (.461, .138, and .078 for the no-growth, moderate-growth, and large-growth conditions, respectively). In terms of the other manipulated factors, the average minimum DGP values were most influenced by the attribute correlations—specifically for the greater correlations, followed by the item quality and the number of items per attribute. Specifically, on average, conditions with lower- and higher-quality items had average minimum DGPs of .061 and .051 (average difference of .010), respectively. On average, conditions with eight and 12 items per attribute had average minimum DGPs of .054 and .058 (average difference of .004), respectively. Figure 22 shows that the number of items per attribute had an inconsistent influence on the average minimum DGPs because some slopes were negative while some were positive. On average, conditions with attribute correlations of 0, .25, .5, and .75 had average minimum DGPs of .077, .076, .049, and .023, with average differences of .001, .027, and .026, respectively. Figure 25 shows similar patterns as in Figure 22, but the attribute profile average minimum DGPs were generally larger than the Attribute 1 DGPs with average minimum DGP values of .152 and .209 for the basic and CW DGPs, respectively. The attribute profile average minimum CWP DGP was still 0 across all conditions—as with Attribute 1.

Next, let us discuss the average mean DGPs for Attribute 1 and the attribute profile. Figure 26 shows greater distinction between the DGP types than Figure 22. As expected, in general, the basic DGP had the greatest average mean DGPs, with a mean value of .748, followed by the CW DGPs, with a mean value of .735, then the PW DGPs, with a mean value of .667, and then the CWP DGPs, with a mean value of .627. Unlike the one-attribute conditions, growth showed a consistent impact on the average mean DGP values with values of .675, .690, and .717 (average differences of .015 and .027, respective) for the no-growth, moderate-growth, and large-growth conditions, respectively. However, just like with the one-attribute conditions, the relative ranges of the average mean DGPs were of more interest than the averages because as growth increased, the average mean DGPs got more consistent, which means that as growth increased, the influence of the other manipulated factors decreased. The ranges for the average mean DGP values were .324, .172, and .104 for the conditions with no growth, moderate growth, and large growth, respectively.

Across all conditions, the other manipulated factors had some influence over the average mean DGP values: the average mean DGPs were greater for the conditions with higher-quality items, greater for the conditions with more items per attribute, and greater for the conditions with greater attribute correlations. On average, the mean DGPs for conditions with lower and higher item quality were .691 and .698 (average difference of .007), respectively. On average, the mean DGPs for conditions with eight and 12 items per attribute were .691 and .697 (average difference of .006), respectively. On average, the mean DGPs for conditions with attribute correlations of 0, .25, .5, and .75 were .675, .684, .698, and .721, with average differences of .009, .014, and .023, respectively.

Therefore, the average mean DGPs were most impacted by the growth and the attribute correlations, and they were least impacted by the item quality and the number of items per attribute. The attribute profile average mean DGP results were nearly identical to the results for Attribute 1.

Next, let us discuss the average median DGPs for Attribute 1 and the attribute profile. Figure 30 also shows patterns similar to those in Figure 26. However, for the average median DGPs, the average median basic DGPs were more distinct from the other DGP metrics. This was due to the fact that many students' basic DGPs were 1 because they achieved the maximum proficiency status by the post-test regardless of their proficiency statuses at the pre-test, especially for the large-growth conditions. The influence of the DGP type on the average median DGPs was the same as with the average mean DGPs: As expected, in general, the basic DGP had the greatest average median DGPs, with a mean value of .851, followed by the CW DGPs, with a mean value of .789, then the PW DGPs, with a mean value of .757, and then the CWP DGPs, with a mean value of .736. On average, the conditions with no growth, moderate growth, and large growth had average median DGP values of .720, .764, and .866, with average differences of .044 and .102, respectively. Like the average mean DGPs, as growth increased, the average median DGPs got more consistent, but only for the CW, CWP, and PW DGPs. The basic DGPs diverged from the other DGPs more as growth increased.

Across all conditions, the other manipulated factors had some influence over the average median DGP values: the average median DGPs were greater for the conditions with higher-quality items, greater for the conditions with more items per attribute, and greater for the conditions with greater attribute correlations. On average, the median

DGPs for conditions with lower and higher item quality were .765 and .802 (average difference of .037), respectively. On average, the median DGPs for conditions with eight and 12 items per attribute were .775 and .791 (average difference of .016), respectively. On average, the median DGPs for conditions with attribute correlations of 0, .25, .5, and .75 were .756, .767, .788, and .822, with average differences of .011, .021, and .034, respectively. Therefore, the average median DGPs were most impacted by the growth, followed by the item quality, then the attribute correlation, and they were least impacted by the number of items per attribute. The attribute profile average median DGP showed similar patterns as did the Attribute 1 average median DGP, but the attribute profile average median DGP values were generally lower than the Attribute 1 values. Most notably, the influence of increasing growth did not have as strong an impact on the attribute profile average median basic DGPs as it did for the Attribute 1 average median basic DGPs because Figure 33 shows that the basic DGP lines were not separated from the other DGP lines like they were in Figure 30.

Finally, let us discuss the average maximum DGPs for Attribute 1 and the attribute profile. Tables 37-39 show that only nine (9.38%), 14 (14.6%), and six (6.25%) of the 96 three-attribute conditions had Attribute 1, Attribute 2, and Attribute 3 average maximum DGPs less than 1, respectively. Recall that 29.4% of the one-attribute conditions had average maximum DGP values less than 1. Tables 37-40 show that all average maximum DGP values were greater than .990. Tables 37-39 and Figures 34-36 show that the lowest average maximum DGPs occurred for the conditions with eight lower-quality items, no growth or moderate growth, and attribute correlations of 0, .25, or .5. Figure 37 shows that the average maximum values for the profile-level DGPs were

much more varied than the average maximum values for the attribute-level DGPs except for the basic DGPs, which all had average maximum profile-level values of 1. Figure 37 also shows that, in general, as the growth increased, the item quality increased, the number of items per attribute increased, and/or the attribute correlations increased, the average maximum DGPs increased. Because all of the average maximum DGP values were so large, I exclude the average values for each condition and rely on the visual interpretations of Figure 37.

To summarize this section, the DGP metrics generally performed as expected across all attributes, at the profile level, and across simulation conditions.

Simulation Study Discussion

In this chapter, I conducted a simulation study to investigate the impact of a few DCM assessment conditions, including the item quality, the number of items to measure each attribute, the level of attribute proficiency growth, the attribute correlations, the number of proficiency statuses per attribute, and the number of attributes on PTDCM reliability and DGP values. My goal was to answer the second research question presented in Chapter 1:

2. How reliable are DGPs under different DCM assessment conditions?
    a. How do the reliability metrics for DGPs perform under different DCM assessment conditions?

Findings from the simulation study showed that, under some conditions, the PTDCM and, therefore, PTDCM classifications can be highly reliable with values greater than .8 and .9. Because DGPs come directly from the PTDCM person estimates, DGPs too can be highly reliable. Despite these high reliability values for some conditions and

some reliability metrics, very few conditions were considered to have had acceptable, good, very good, or excellent reliability according to the criteria from Schellman and Madison (in press). It is possible that, because these were polytomous attributes that had greater data requirements, increasing the number of items per attribute or the sample size would help get these reliability values to more desirable thresholds.

For the one-attribute and three-attribute conditions and within each reliability metric, the conditions with the greatest reliabilities (.820 to .926 for the polychoric metric) were the conditions with 12 higher-quality items measuring an attribute with three proficiency statuses and any level of growth, and of these conditions, the conditions with greater attribute correlations generally showed greater reliabilities. The conditions that showed the lowest reliabilities (.521 to .686 for the polychoric metric) were the conditions with eight lower-quality items measuring an attribute with five proficiency statuses. Table 41 summarizes the relative impacts of the manipulated factors on the reliabilities and other key evaluation metrics discussed below. The values in Table 41 match the values presented in the previous section for each comparison of relative impact. In Table 41, one can see how the change in each manipulated factor (e.g., changing the number of proficiency statuses from three to five: 3→5) changed each evaluation metric overall. These trends for reliability were expected because the conditions with high reliability had simpler models and more information than the conditions with low reliability. Additionally, in general, the polychoric metric had the greatest reliabilities, and the PB metric had the lowest reliabilities, but the polychoric metric was more influenced by the attribute correlations than the other two metrics.

Growth did not have as strong or as consistent an impact on reliability as did the other manipulated factors: decreasing the number of proficiency statuses, increasing the item quality, increasing the number of items per attribute, and/or increasing the attribute correlation yielded greater reliabilities. However, the impact of item quality was greater than the impact of the number of proficiency statuses, which was greater than the impact of the number of items, which was greater than the impact of the attribute correlation. The finding that DGP reliability increased when attribute correlations increased was counter to research that has shown SGP reliability to decrease with increased between-testing occasion correlations (Monroe & Cai, 2015), which indicates favorability for DGPs over SGPs.

Consistent with previous PTDCM research (Madison et al., 2023), this simulation study yielded reliabilities that were generally lower than those of cross-sectional LCDM, TDCM, or PDCM reliabilities, but this result was expected because the PTDCM is a much more complex model than are the LCDM, TDCM, and PDCM. We will see in the next chapter that the empirical data analyses that I conducted to illustrate DGPs with real data yielded slightly greater reliabilities than the simulation study. For example, the polychoric reliability was .941 (very good reliability, according to Schellman and Madison, in press) for the second empirical data analysis with one attribute with three proficiency statuses and an attribute correlation of .808.

The DGP results generally matched my expectations. Specifically, the CWP DGP had average minimum values of 0 across all conditions, the average minimum values for the basic DGP were all close to 0, the DGP with the greatest average minimum values was the CW DGP, the basic DGP had the greatest average mean and median values, the

CWP DGP had the lowest average mean and median values, and the majority of average maximum DGP values were equal to 1. The average mean DGPs were most impacted by the growth and the number of proficiency statuses, followed by the attribute correlations and the item quality, and they were least impacted by the number of items per attribute. Additionally, as the growth increased, the overall ranges of the average mean DGPs decreased, which means that as growth increased, the average mean DGPs were not as influenced by the other manipulated factors. For the one-attribute conditions, the average median DGPs were most impacted by the number of proficiency statuses and the attribute correlations, followed by the growth and the item quality, and they were least impacted by the number of items per attribute. For the three-attribute conditions, the average median DGPs were most impacted by the growth and the item quality, followed by the attribute correlations, and they were least impacted by the number of items per attribute. For the one- and three-attribute conditions, growth did not have a consistent impact across all DGP metrics—as growth increased, the average median DGPs got more consistent for the CW, CWP, and PW DGPs, but the basic DGPs diverged from the other DGPs, which dampened the overall impact of growth on the average median DGPs.

In terms of the other evaluation metrics, the item parameters were recovered well, on average, across all conditions. Growth did not have as strong or as consistent an impact on reliability as did the other manipulated factors: decreasing the number of proficiency statuses, increasing the item quality, increasing the number of items per attribute, and/or increasing the attribute correlation resulted in increased item parameter estimation accuracy. However, the number of proficiency statuses had a greater impact than the item quality, which had a greater impact than the number of items per attribute,

which had a greater impact than the attribute correlations. It is known that sample size has a large impact on item parameter recovery (e.g., Madison & Bradshaw, 2018), but sample size was not manipulated in this study. Future research with the PTDCM reliability metrics should include sample size as a manipulated factor.

Like reliability, student classification accuracy was lower in my study than in DCM studies that use dichotomous attributes measured at one testing occasion, but they were consistent with Madison and colleagues (2023). Consistent with expectations, the attribute-level CCRs were greater than the attribute profile-level or class-level CCRs. When growth was low, the ACCR values for the pre-test and post-test were nearly identical, while the CCCR values were more distinct, and the CCCR values were more influenced by the attribute correlations than the ACCR values. When the growth was larger, the ACCR values for the pre-test and post-test were much more distinct, and the CCCR values were less influenced by the attribute correlation values. In terms of the other manipulated factors, decreasing the number of proficiency statuses, increasing the item quality, increasing the number of items per attribute, and increasing the attribute correlation yielded greater CCR values, but like the item parameter estimation accuracy, the number of proficiency statuses had a greater impact than the item quality, which had a greater impact than the number of items per attribute, which had a greater impact than the attribute correlations.

Significance, Limitations, and Future Research

This simulation study presents a significant contribution to the longitudinal DCM literature because it systematically investigated the performance of new PTDCM reliability metrics under different DCM assessment conditions and showed that DGPs can

be acceptably reliable under specific conditions. Additionally, the results of this study can help to inform data reviews in practical situations because I found that item quality was the most influential factor for PTDCM reliability. The empirical item quality in practical situations needs to be investigated so longitudinal assessments with items that measure polytomous attributes can have the necessary discriminations to yield high reliabilities. My results indicate that practitioners should focus their efforts more on developing high-quality items rather than developing more items because increasing the test length will not help their reliabilities as much as increasing their item quality. However, more research is needed in this area because, as with all simulation studies, this study was limited in that I could not explore all assessment conditions that I wanted to explore.

A major limitation for PTDCM studies, in general, is that the software available for estimating these complex models is limited. I used Mplus to estimate the PTDCM, but Mplus took an average of one hour and 10 minutes for the three-attribute conditions to finish running. This lengthy estimation time limited the number of conditions and replications I could include in my study. Additionally, more complex models (i.e., three attributes with five proficiency statuses, more than two testing occasions, and complex item types rather than simple-structure items) were infeasible to include with this simulation study because current estimation algorithms cannot support such complex models. As of right now, R has packages such as mirt (Chalmers, 2012) that can estimate the PDCM for one attribute, but the mirt package cannot provide the posterior probabilities for all latent classes for multiple attributes or multiple testing occasions (M. Madison, personal communication, June 17, 2024), so it cannot be used to compute

PTDCM reliability or DGP metrics. Hopefully, estimation software will soon allow for more complex PTDCM assessment conditions to be investigated via simulation studies.

CHAPTER 5

EMPIRICAL DATA ANALYSIS

In Chapter 3, I answered the research questions: (1a) What are adjustments to DGPs to account for uncertainty in student classification from the DCM? (1b) How can DGPs be adjusted to consider penalties for forgetting? (1c) What are the implications of DGP interpretability for different adjusted DGPs? And (1d) How can DGP reliability be conceptualized and computed? In this chapter, I explore the developed DGP and reliability metrics using empirical data. This chapter includes analyses of two empirical data sets. The purpose of these empirical data analyses was to illustrate the DGP metrics and compare the eight reliability metrics for empirical data with differing numbers of proficiency statuses and attributes. First, I describe the data sets I used for this study and how I analyzed them. Then, I present the results of the empirical data analysis.

Empirical Data Sets

The first data set, which I refer to as *A4_cat2*, included response data from 879 middle school math students who participated in a large-scale mathematics study that used a 21-item assessment as a pre-test and as a post-test with a six- to eight-week instructional intervention between the two testing occasions (Bottge et al., 2014; Bottge et al., 2015; Madison & Bradshaw, 2018). The 21-item assessment measured four attributes with four, six, five, and six simple-structure, open-ended items, respectively. Table 42 shows the Q-matrix for the A4_cat2 data set. The four attributes in this data were all dichotomous and measured problem-solving skills in mathematics domains,

including ratios and proportional relationships, measurement and data, number system, and geometry. The intervention was an instructional program that included mathematics problem-solving situations in video-based scenarios, which is a technique that has been shown to help students with disabilities work through the math in contextual problems (Bottge et al., 2001). The students in this sample were 54% male and 78% white. The grade level breakdown was 15% sixth grade, 64% seventh grade, and 21% eighth grade. The purpose of analyzing A4_cat2 for this dissertation was to illustrate the utility of the DGP metrics with data from a multidimensional assessment that was administered at multiple testing occasions. However, A4_cat2 was limited in illustrating the DGPs and their reliabilities because it included only dichotomous attributes. DGPs can be applied in dichotomous situations, but it is important to illustrate how DGPs manifest with polytomous attributes because they allow for greater separation of student groups via a greater number of transitions than dichotomous attributes have.

The second data set, which I refer to as *A1_cat3*, was unidimensional, so I could not use it to illustrate some of the aggregation techniques for the DGP metrics. However, the attribute in this data set was trichotomous, so I could use A1_cat3 to illustrate the DGPs and their reliabilities for attributes with more than two proficiency statuses. The A1_cat3 data set included response data from 910 fourth-grade English language arts students who took a 25-item assessment at the beginning of the school year and a 25-item assessment at the end of the school year. I do not have access to demographic data for this sample of students.

Evaluation of the Empirical Data Analyses

To evaluate these empirical data analyses for A4_cat2 and A1_cat3, in the next two sections I report observed attribute correlations (between and within testing occasion), IRPs for each student group to evaluate item quality, latent class proportions at the pre-test and post-test, attribute base rates, and attribute transitions, but I focus the results mostly on the DGPs and their reliabilities.

For each attribute, I computed all eight PTDCM reliability metrics (PB, PBW, PF, PFW, IG, IGW, polychoric, and average maximum transition) that were introduced in Chapter 3, and I compared them to see how they perform with different types of empirical data. Finally, I computed all students' basic DGPs, CW DGPs, CWP DGPs, and PW DGPs. I include summaries of the DGPs across attributes and students and demonstrate how to aggregate DGPs to the student and class levels. The results illustrate how the DGPs compare with each other and how they all perform in practice. The results related to reliability and the DGPs illustrate how DGPs could be used in practice and highlight the benefits and limitations related to their use. In Chapter 6, I compare the DGP benefits and limitations to those of SGPs to see which SGP limitations are lessened or remain. Finally, I conclude with recommendations for using the DGP in practice.

Results for the Empirical Data Analysis with Four Attributes with Two Proficiency Statuses

I begin this section with analyses that are commonly conducted for empirical analyses with DCMs. Then, I dive into details related to the DGPs and reliabilities resulting from the analyses with A4_cat2. Table 43 shows the attribute correlations for between and within testing sessions. The off-diagonal within- and between-testing

occasion correlations ranged from .220 to .680, with an average value of .408 and median value of .380. In general, the within-testing occasion correlations were greater for the post-test than for the pre-test, which had mean off-diagonal correlations of .489 and .614, respectively. The within-testing occasion correlations were greater than the between-testing occasion correlations, which had overall average correlations of .552 and .301, respectively. Finally, for the between-testing occasion correlations, the average between correlations of each attribute with itself over time (.329) was greater than the average between correlations of each attribute and all other attributes over time (.291).

Figure 38 shows the IRPs for the non-proficient and proficient student groups for each item. The items are grouped based on the Q-matrix, so each subplot in Figure 38 corresponds with a particular attribute. In the context of A4_cat2, item discrimination is the difference between the IRP for the students in the proficient group and the IRP for the students in the non-proficient group. Most items were high quality, with discriminations ranging from .210 to .810, an average discrimination of .539 and a median discrimination of .550. None of these items would have been flagged as being low quality in a data review or item analysis, which typically uses cut-off values around .15 to flag dichotomous items with too low discrimination (Crocker & Algina, 1986). The easiest item for non-proficient students was Item 2. All items across all attributes were generally difficult for non-proficient students, but the most difficult item for non-proficient students was Item 7, followed closely by Item 17. The easiest item for proficient students was Item 11, followed closely by Item 18. The most difficult item for proficient students was Item 17. Attribute 1 had one very easy item (Item 13) and one more difficult item (Item 17) for proficient students, with IRPs of .896 and .226, respectively. The other two items

were of moderate to easy difficulty for proficient students, who had IRPs of .555 and .69. Attribute 2 had five easy items for proficient students, who had IRPs ranging from .813 to .973 for these five items. The sixth item for Attribute 2 (Item 3) was difficult for proficient students, who had an IRP of .335, but it was still highly discriminating between proficient and non-proficient students. Attribute 3 had two very easy items (Items 5 and 6) for proficient students, with IRPs of .932 and .908. The three other items were of moderate difficulty for proficient students, with IRPs ranging from .584 to .619. Attribute 4 had one very easy item for proficient students (Item 18), who had an IRP of .972. The other five items were of moderate difficulty for proficient students, who had IRPs ranging from .464 to .782 for these five items.

Figure 39 shows the latent class proportions for the pre-test only. At the pre-test, the A4_cat2 assessment had 16 latent classes—the fully-crossed proficiency statuses across four dichotomous attributes. Figure 39 shows that the largest latent classes at the pretest were the first and last latent classes, which was expected because the first class included the students who were non-proficient for all attributes, and the last class included the students who were proficient for all attributes. Because the attributes were moderately correlated, it was likely that proficiency in one of the attributes was related to proficiency in another attribute, so I expected the first and last classes together to carry a large proportion of students. Figure 40 shows the latent class proportions for the post-test only. Again, the largest classes were the first and last classes, but the last class (proficiency for all attributes) at the post-test was much larger than it was at the pre-test, which was expected because students experienced an intervention between the pre-test and post-test, so I expected that they grew and acquired more attribute proficiencies.

However, comparing the latent class proportions at each testing occasion does not help us interpret growth—it only gives us information about student proficiencies at each testing occasion separately. Instead, we need to consider transition probabilities, but before discussing transition probabilities, let us discuss the base rates for the four attributes at each testing occasion.

Figure 41 shows the base rates for each attribute for the pre-test and post-test. The base rate shows the marginal latent class proportions, summed across the latent classes that include proficiency for the attribute. In other words, the base rate is the proportion of students who are proficient for each attribute at each testing occasion. At the pre-test, Attributes 1 and 3 were the most difficult because they had the smallest proportions of students who were proficient for these attributes at the pre-test. The easiest attribute at the pre-test was Attribute 2, with a base rate of .625. At the post-test, the base rates for Attribute 1 increased at the greatest rate, with an increase of .235. Attributes 3 and 4 increased by .160 and .191, respectively. Attribute 2 had the smallest increase in base rate over time, with an increase of .064. At the post-test, Attribute 3 was still the attribute with the lowest base rate, while Attributes 2 and 4 were the attributes with the greatest base rates.

As mentioned previously, the latent classes and base rates in Figures 39, 40, and 41 did not show the full picture for the longitudinal setting. It is important to also consider for each attribute the proportion of students who had each attribute transition. For dichotomous attributes measured at two testing occasions, there are four transitions for each attribute: non-proficient at the pre-test and non-proficient at the post-test (denoted [00]), non-proficient at the pre-test and proficient at the post-test (denoted [01]),

proficient at the pre-test and non-proficient at the post-test (denoted [10]), and proficient at the pre-test and proficient at the post-test (denoted [11]). Figure 42 shows the proportion of students who had each transition for each attribute. All four plots in Figure 42 showed results that I would expect and hope to see. Specifically, for each attribute, the transition with the smallest proportion of students was the transition [10]. It is good if transition [10] has few students because this transition shows forgetting over time—moving from proficient to non-proficient. In education, we expect that students learn and do better over time, so we do not expect forgetting, although we know that, in reality, some students will forget. The proportions of students with the transition [10] ranged from .066 to .11 across the four attributes, which was appropriately small so as not to cause concern. All attributes showed that most students did not change proficiency statuses over time because the sum of the transition proportions for the [00] and [11] transitions was greater than .5 for all attributes. The attribute that showed the greatest proportion of learning was Attribute 1 because it had a transition probability of .301 for transition [01], and the attribute that showed the smallest proportion of learning was Attribute 2 because it had a transition probability of .16 for the transition [01]. Table 44 shows the conditional transition matrices for each attribute in A4_cat. These matrices are transformations of the values from Figure 42 that showed the proportion of students from each pre-test proficiency status who moved to each proficiency status at the post-test. For example, 52.2% of the students who were non-proficient for Attribute 1 at the pre-test were still non-proficient for Attribute 1 at the post-test, 47.8% of the students who were non-proficient for Attribute 1 at the pre-test transitioned to proficient for Attribute 1 at the post-test, 17.8% of the students who were proficient for Attribute 1 at the pre-test

transitioned to non-proficient for Attribute 1 at the post-test, and 82.2% of the students who were proficient for Attribute 1 at the pre-test were still proficient for Attribute 1 at the post-test. These conditional transition matrices allow for easier interpretation of student growth over time. Table 44 shows that Attribute 3 had the greatest proportion of students who were proficient at the pre-test transitioning to non-proficient at the post-test. So even though the overall proportion of students who had the transition [10] for Attribute 3 was .11 (see Figure 42), these .11 students accounted for about 30% of the students who were proficient at the pre-test. This example shows how the conditional transition probabilities can tell a different story from the raw transition probabilities. All four attributes had about 50% of the students who were non-proficient at the pre-test transition to proficient at the post-test.

Based on the results presented so far, A4_cat2 demonstrated high-quality characteristics for a DCM analysis (i.e., highly-discriminating items, base rates increased over time, and few students were classified into transitions with forgetting). Now, I present the reliability and DGP-related results. Figure 43 and Table 45 show the values for the reliabilities for each attribute. In Table 45, I used the suggested reliability levels for what can be considered acceptable, good, very good, or excellent reliability from Schellman and Madison (in press) to provide qualitative descriptions for values of the PB, PF, IG, polychoric, and average maximum transition metrics. Schellman and Madison (in press) did not include the PBW, PFW, or IGW metrics in their suggestions for describing the levels of reliability. As shown in Table 45, across the PB, PF, IG, polychoric, and average maximum transition metrics, all four attributes had good, very good, or excellent reliability, which was better than the simulation study results in which

most conditions did not have even acceptable reliability and no conditions had excellent reliability, according to the suggestions from Schellman and Madison (in press). The values for the reliabilities were greater than or equal to .759 for all metrics except for IG and IGW, which had reliabilities ranging from .534 to .627. As in the simulation study, the polychoric reliability metric showed the greatest reliabilities across all four attributes, followed closely by the average maximum transition metric, which had average values of .931 and .909, respectively. The IG and IGW metrics showed the lowest reliabilities for all four metrics, with average values of .549 and .598, respectively. The weighted versions of the PB, PF, and IG metrics all showed slightly greater reliabilities than their non-weighted versions, with average values of .799 and .815 for PB and PBW and .827 and .843 for PF and PFW, respectively. The PB, PBW, PF, and PFW metrics all yielded comparable levels of reliability. These results agree with findings from previous TDCM reliability studies (Madison, 2019; Schellman & Madison, in press). Figure 43 shows that Attribute 2 had the greatest reliabilities across all eight metrics. The other three attributes all had similar reliabilities, but Attribute 3 had slightly greater reliability than Attributes 1 and 4, and Attribute 4 had greater reliabilities than Attribute 1 except for the PFW and polychoric metrics.

Recall that the average maximum transition reliability metric is the average of students' maximum posterior probabilities, regardless of the latent class with which the posterior probabilities correspond. To further evaluate the average maximum transition reliability metric, it is helpful to determine the proportion of students' maximum posterior probabilities that were greater than a few checkpoints on the zero-to-one scale. The greater the maximum posterior probability, the greater the PTDCM's confidence in

students' classifications. Figure 44 and Table 46 show, for each attribute, the proportion of maximum posterior probabilities that were greater than .1, .2, .3, …, .8, and .9. As expected, the lines in this plot were decreasing because fewer maximum posterior probabilities were greater than the checkpoints at the greater end of the zero-to-one scale. Attribute 2 had the least steep slopes, which means that the PTDCM was more confident in students' classifications for Attribute 2 than for the other attributes, which all had nearly identical lines in Figure 44. For all attributes, large proportions of students (i.e., greater than .777) had maximum posterior probabilities greater than .8. The proportions of students with maximum posterior probabilities greater than .9 ranged from .661 (for Attribute 1) to .821 (for Attribute 2). In summary, the model was generally highly confident in student classifications for all attributes in the A4_cat2 data.

Because the attributes showed good to excellent reliability, and the model showed high confidence in student classifications, it follows that students' DGPs were also highly reliable because DGPs come directly from students' transitions and posterior probabilities. Figure 45 shows the basic DGPs for each transition for each attribute. For example, students who had the transition [00] for Attribute 1 had a basic DGP of .522; students who had the transition [01] for Attribute 1 had a basic DGP of 1; students who had the transition [10] for Attribute 1 had a basic DGP of .178; and students who had the transition [11] for Attribute 1 had a basic DGP of 1. Recall from Chapter 3 that it is expected that the basic DGPs for students who have the greatest proficiency status at the last testing occasion are all be equal to 1 regardless of their proficiency statuses at the previous testing occasion(s). Figure 45 also shows that the smallest basic DGP for each attribute corresponded with transition [10], which was also expected because transition

[10] was the transition with forgetting. Note that the basic DGPs can be directly derived from the conditional transition probabilities in Table 44—the basic DPGs are the cumulative sums of the conditional transition probabilities as you move across the columns of Table 44.

I chose to display the plots in Figure 45 as stacked bar plots because they illustrate how the basic DGP increases as the post-test proficiency status increases within each pre-test proficiency group. The basic DGPs for Attribute 1, for example, can be interpreted as follows: Students with the transition [00] for Attribute 1 showed growth greater than or equal to (in this case, it was "equal to") 52.2% of the students who were in the non-proficient group at the pre-test. Students with the transition [10] for Attribute 1 showed growth greater than or equal to (in this case, it was "equal to") 17.8% of the students who were in the proficient group at the pre-test. Students with the transition [01] or [11] for Attribute 1 showed growth greater than or equal to 100% of the students in the sample because they reached the maximum proficiency status. These interpretations are somewhat trivial for dichotomous attributes. The analysis of A1_cat3 below shows that the interpretation of basic DGPs for polytomous attributes is more interesting than the interpretations of basic DGPs for dichotomous attributes.

I computed each student's transition for each attribute and assigned them to have the basic DGP that corresponded with their most likely transition for each attribute. Thus, all students in the A4_cat2 data set had a basic DGP for each attribute. After determining each student's basic DGP for each attribute, I computed their adjusted DGPs for each attribute. To evaluate how the adjusted DGP metrics performed with the A4_cat2 data set, I wanted to examine their distributions conditional on students' basic DGPs. For

example, the first panel in Figure 46 shows the distributions of students' CW DGP values with a box plot for each basic DGP for Attribute 1. In this first panel of Figure 46, we can see that the distribution of the CW DGP metric for students who had a basic DGP of .178 was lower than the distribution of the CW DGP metric for students who had a basic DGP of .522, which was lower than the distribution of the CW DGP metric for students who had a basic DGP of 1. This panel shows that adjusting DGPs to account for uncertainty in student classifications generally did not change students' DGPs drastically—students with very low basic DGPs did not have very high CW DGP values, and students with very high basic DGPs did not have very low CW DGP values. However, all the distribution of the CW DGP values for students with a basic DGP of .178 did overlap with the distribution of the CW DGP values for students with a basic DGP of .522, which overlapped with the distribution of the CW DGP values for students with a basic DGP of 1, but these overlaps only occurred in outlying values—not in the center of the distributions. The other panels in Figure 46 and Figures 47, 48, and 49 can be similarly interpreted with the conclusion that adjusting the DGP to account for uncertainty in student classification and to penalize for forgetting did not drastically change students' locations on the zero-to-one scale, in general. In other words, students with low basic DGPs generally had low adjusted DGPs, and students with high basic DGPs generally had high adjusted DGPs.

To investigate students' DGPs more closely, I created Figures 50, 51, 52, and 53. In this section, I focus on Figure 50 (Attribute 1) as an example, but any of these plots can be similarly interpreted. Figure 50 shows all students' basic and adjusted DGPs, and each line in Figure 50 corresponds with a student in the A4_cat2 data. The first panel of

Figure 50 shows all students who had a basic DGP of .178 for Attribute 1. The second panel of Figure 50 shows all students who had a basic DGP of .522 for Attribute 1. The third panel of Figure 50 shows all students who had a basic DGP of 1 for Attribute 1. The line plots in Figures 50, 51, 52, and 53 give a slightly different picture of the distributions of the DGPs than the box plots in Figures 46, 47, 48, and 49 because the line plots showed how individual students' DGPs changed with each adjustment. Figures 50, 51, 52, and 53 showed that although the basic DGP is theoretically a continuous variable, only a few values on the zero-to-one scale were values for the basic DGP because of the discrete nature of the transitions, and the fewer proficiency statuses an attribute has, the fewer basic DGP values the attribute has. The adjusted DGPs, on the other hand, have more values, so the theoretically continuous scale of the DGP is realized with the adjusted metrics. Thus, the adjusted DGPs have more variability than the basic DGP, and they can be used to better differentiate students than the basic DGP, which has the same value for all students who have the same transition.

For each group of students (for each panel in Figure 50), we can extract the trends in the adjusted DGPs. For example, for students with a basic DGP of .178, there were two general patterns; for students with a basic DGP of .522, there were three general patterns; and for students with a basic DGP of 1, there were three general patterns. I further inspected these patterns to better illustrate how and why they occur by extracting the transition probabilities and DGPs for one example student for each trend.

Figure 54 shows the line plots for these eight example students, but rather than separating Figure 54 into panels based on students' basic DGPs, I put them all into one plot. The points on the lines in Figure 54 correspond with the students' DGP values. The

legend below the plot in Figure 54 first shows the student ID number. Then, it shows each student's posterior probability for each transition—the first posterior probability is for transition [00], the second posterior probability is for transition [01], the third posterior probability is for transition [10], and the fourth posterior probability is for transition [11]. I can use the transition probabilities and basic DGPs to explain and interpret each student's DGP values. For example, Student 861 (the red line) had a 100% chance of belonging to the transition [11], so their adjusted DGPs were all 1 because the model was certain in Student 861's classification, and there was a 0% chance that Student 861 belonged to the forgetting transition—transition [10].

Alternatively, Student 185 (the orange line) had a basic DGP of 1, a CW DGP of .944, a CWP DGP of .932, and a PW DGP of .938. Student 185 had a .06 probability of having the transition [00], a .801 probability of having the transition [01], a .034 probability of having the transition [10], and a .106 probability of having the transition [11]. Student 185 had a basic DGP of 1 because their most likely transition was [01], in which the final testing occasion had the greatest proficiency status. Student 185's CW DGP was less than their basic DGP because they had a non-zero probability of belonging to each other transition. Student 185's CWP and PW DGPs were less than their CW DGP because they had a non-zero probability of belonging to the transition [10], so they got penalized for having a chance (even though it was a small chance) of forgetting instead of learning.

Figure 54 shows a third and final pattern from the group of students who had a basic DGP of 1. This student, Student 256, had a basic DGP of 1, a CW DGP of .600, a CWP DGP of .430, and a PW DGP of .515. Student 256 had a .02 probability of having

the transition [00], a .002 probability of having the transition [01], a .475 probability of having the transition [10], and a .503 probability of having the transition [11]. Student 256 had a basic DGP of 1 because their most likely transition was [11], in which the final testing occasion had the greatest proficiency status. However, their next most likely transition was [10] with a probability of nearly .5. Therefore, when applying the adjusted DGP metrics, Student 256's DGP dropped because there was a large probability that their true transition for Attribute 1 was [10] instead of [11].

Figure 54 shows three patterns for students who had a basic DGP of .522 for Attribute 1, which corresponds with students who had the transition [00]. The first student in this set of patterns is Student 835 (the green line). Student 835 had a basic DGP of .522, a CW DGP of .720, a CWP DGP of .670, and a PW DGP of .695. Student 835 had a .347 probability of having the transition [00], a .198 probability of having the transition [01], a .139 probability of having the transition [10], and a .316 probability of having the transition [11]. Student 835's greatest transition probability was for transition [00], but their second greatest transition probability was for transition [11], and their probabilities for these two transitions were nearly identical (.347 and .316, respectively). Therefore, when adjusting the DGP to account for uncertainty in student classifications, Student 835's DGP increased because they had a sizeable probability of belonging to a transition that corresponds with a basic DGP of 1. However, Student 835's CWP DGP was less than their CW DGP because they had a non-zero probability of belonging to the transition [10].

The second pattern for the group of students who had a basic DGP of .522 is illustrated by Student 382 (the blue-green line). Student 382 was in a similar situation as

Student 861 because the model was extremely confident in their classifications, but in the case of Student 382, the model was confident that the student belonged to the transition [00], so Student 382's basic and adjusted DGPs were equivalent because they had a 0% chance of belonging to the transition [10].

The third and final pattern for the group of students who had a basic DGP of .522 is illustrated by Student 103 (the blue line). Student 103 was in a situation similar to that of Student 256, where Student 103's most likely transition was [00], but their next most likely transition was [10]—the forgetting transition, and it had a large probability (i.e., .484). Therefore, when adjusting for uncertainty in student classifications and penalizing for having a chance of forgetting, Student 103's DGP decreased.

The final two lines in Figure 54 were for students who had basic DGPs of .178, which corresponds with the students who were in the transition [10]. The first student, Student 484 (the purple line), was in a situation similar to Student 103 but with different transitions. In Student 484's case, their most likely transition was [10], but their next most likely transition was [11], and it had a high probability of being Student 484's true transition (i.e., .474). Therefore, when adjusting the DGP, Student 484's DGP increased because it was possible that the student did not forget and instead maintained the same proficiency status.

The last pattern (the pink line) is for Student 828, whose most likely transition was [10], but in this case, the model was very confident in Student 828's classification because they had a high probability (i.e., .938) of belonging to the transition [10]. Student 828's CW DGP was greater than their basic DGP because they had a non-zero probability

(i.e., .062) of belonging to the transition [11], but their CWP DGP was 0 because of their high probability of belonging to the transition [10].

The eight example students illustrated in Figure 54 characterize the types of patterns that can be extracted from Figures 50, 51, 52, and 53 and illustrate how the DGP metrics perform with different combinations of different magnitudes of transition probabilities.

The final two analyses of the DGPs for the A4_cat2 data set were related to the different ways to aggregate DGPs across attributes and across students. The first analysis was for aggregation across attributes but not across students. Every student had a basic DGP and three adjusted DGPs for each attribute. Recall that, to aggregate across attributes, one can compute the average of each type of DGP across all attributes to obtain attribute profile-level DGPs. Figure 55 shows all profile-level DGPs. The most important point to make about this plot is that students' profile-level basic DGPs could have taken any one of 81 different values because each attribute had three basic DGPs, as shown in Figure 45, and each student could have had any one of these values for each attribute, so the total number of profile-level basic DGPs was $3 \times 3 \times 3 \times 3 = 81$. Therefore, if separation between students is desirable, using a multidimensional assessment and focusing only on students' profile-level basic DGPs will separate students, and the use of adjusted DGPs for the purpose of separating students might not be necessary. However, the profile-level basic DGP still does not factor in model uncertainty or penalize students for forgetting; the choice of which DGP(s) to use in practice depends on the dimensionality (in unidimensional assessments, there is no

profile-level DGP; so if separation is desired, then one must use the adjusted DGPs) of the assessment and stakeholders' desired uses and interpretations of results.

Figure 56 shows the DGPs averaged across students for each combination of attribute and DGP type. The standard errors for the means in Figure 56 ranged from .007 to .011, with mean and median values of .009 and .009, respectively. The first thing to notice about Figure 56 is that there was no noticeable difference between the mean values for the different DGP types. They ranged from .715 to .827, with mean and median values of .786 and .783, respectively. In general, the average DGP across all students was greatest for Attribute 2 and smallest for Attribute 3. Figure 56 is helpful for knowing which attribute had the greatest average DGP across all students, but it does not help us compare different groups of students, which may be desirable in practice.

I computed the means in Figure 56 using all students in the sample. However, the A4_cat2 data set included two groups of students: the 456 students who received the control condition and the 422 students who received the treatment condition between the first and second testing occasions. Figure 57 shows the average DGPs across students within each experimental group. The standard errors for the mean ranged from .010 to .016, with mean and median values of .013 and .013, respectively. Figure 57 shows that across all attributes and DGP types, the treatment group had greater mean DGPs than the control group, and these mean differences were significant at the 95% confidence level for all but four comparisons: Attribute 2 basic DGP, Attribute 2 CWP DGP, Attribute 2 PW DGP, and Attribute 3 CWP DGP. The finding that the treatment group generally showed greater growth than the control group was consistent with previous studies that used the A4_cat2 data set (e.g., Madison 2019).

Results for the Empirical Data Analysis with One Attribute with Three Proficiency Statuses

This section shows the results of the same analyses as in the previous section but for the A1_cat3 data set. The results in this section show how DGPs look different for different types of attributes. As in the previous section, this section starts with typical DCM analyses and then shows the reliability and DGP results.

After fitting the PTDCM to the A1_cat3 data, I found that the between-testing occasion correlation of this attribute with itself was .808, which was reasonably high and greater than the average between-testing occasion correlations of each attribute with itself over time from the A4_cat2 data set. Because A1_cat3 was unidimensional, there were no other attribute correlations to consider.

Table 47 and Figure 58 show the IRPs for each proficiency group for each item. In general, the IRPs followed the pattern I would expect: beginning students had the lowest IRPs, and advanced students had the greatest IRPs. However, a few items (Items 6, 8, 9, and 10) had the same IRPs for the beginning and proficient groups. For Item 10, advanced students had an IRP that was only slightly greater than the IRPs for the beginning and proficient students. For a trichotomous attribute, there were three discriminations: the difference between the IRPs for the beginning and proficient students, the difference between the IRPs for the proficient and advanced students, and the difference between the IRPs for the beginning and advanced students. For the A1_cat3 data set, the proficient/beginning discriminations ranged from 0 to .461, with a mean discrimination of .212 and a median discrimination of .218. The advanced/proficient discriminations ranged from .035 to .421, with a mean discrimination

of .243 and a median discrimination of .220. The advanced/beginning discriminations ranged from .035 to .739, with a mean discrimination of .455 and a median discrimination of .493. These mean discrimination values were comparable to the lower-quality items that I used in my simulation study.

Because trichotomous attributes had more than one discrimination (as with dichotomous attributes), they could not be interpreted with the same criteria that dichotomous attributes use (i.e., a cut-off value of .15). In general, the discriminations for the A1_cat3 data set were moderate to high with only a few non-discriminating or poorly discriminating items. The easiest item for advanced students was Item 12, with an IRP of .971, closely followed by Item 24 (the IRP was .970), Item 4 (the IRP was .961), and Item 17 (the IRP was .944). The most difficult item for advanced students was Item 10, with an IRP of .286, followed closely by Item 6, with an IRP of .293. The easiest item for proficient students was Item 12, with an IRP of .865. The most difficult item for proficient students was Item 9, with an IRP of .183, followed closely by Item 8 (the IRP was .198), Item 6 (the IRP was .220), and Item 10 (the IRP was .250). Finally, the easiest item for beginning students was Item 12, with an IRP of .455. The most difficult item for beginning students was Item 9, with an IRP of .183, followed closely by Item 18 (the IRP was .184) and Item 8 (the IRP was .198).

Figure 59 shows the latent class proportions for the pre-test only. At the pre-test, 51.5% of students were classified as beginning, 33.3% of students were classified as proficient, and 15.2% of students were classified as advanced. Figure 60 shows the latent class proportions for the post-test only. At the post-test, 36.9% of students were classified as beginning, 28.5% of students were classified as proficient, and 34.6% of students were

classified as advanced. Because the A1_cat3 data set was unidimensional, the base rates for the attribute were the same as the latent class proportions at each testing occasion. The patterns in Figures 59 and 60 generally matched my expectations because the pre-test showed many students who had lower proficiency statuses, and the post-test showed that students learned because more students had greater proficiency statuses.

Based on a comparison of Figures 59 and 60 to evaluate student growth, and with an attribute correlation of .808, 25 items for one attribute with three proficiency statuses and lower-quality items, the A1_cat3 analysis was most similar to the one-attribute, I12_cat3 simulation condition with an attribute correlation of .75, lower-quality items, and moderate growth.

Figure 61 shows the proportion of students who had each transition for the attribute. The transition with the greatest proportion of students was the [00] transition. The transitions with the smallest proportions of students were the transitions with forgetting (i.e., [10], [20], and [21]) and the transition that moved from the lowest proficiency status to the greatest proficiency status (i.e., [02]), which was expected. Figure 61 shows that 14.9% of students moved from beginning to proficient, and 19.6% of students moved from proficient to advanced. A total of 64.6% of students maintained their same proficiency statuses over time.

Table 48 shows the conditional transition probabilities for A1_cat3. As described previously with A4_cat2, the conditional transition probabilities give a different picture from the raw transition probabilities in Figure 61. Table 48 shows that the conditional probabilities of forgetting (i.e., .007, .009, and. 013) were very small, which was desirable. The conditional probability that students who were beginning at the pre-test

transitioned to advanced at the post-test (i.e., .004) was also very small. The conditional proportions of students whose proficiency statuses increased by one level were larger for students who were proficient (i.e., .589) at the pre-test than for students who were beginning at the pre-test (i.e., .289).

The rest of this section focuses on the PTDCM reliability and DGP results for A1_cat3. Figure 62 shows the reliabilities for each of the eight metrics for the attribute in the A1_cat3 data set. Figure 62 shows results similar to those in Figure 43: the polychoric and average maximum transition metrics had the greatest reliabilities, the weighted versions of the PB, PF, and IG metrics had greater reliabilities than the unweighted versions, and the IG and IGW metrics had the smallest values. However, for A1_cat3, the IGW value was closer to the PB and PF values than it was for A4_cat2 because the PB and PF reliabilities were smaller for A1_cat3 than they were for A4_cat2. Based on the recommended ranges from Schellman and Madison's (in press) analysis, the PB, PF, and IG reliability values for A1_cat3 were all less than acceptable, but the polychoric reliability was very good, and the average maximum transition reliability was acceptable.

These results were consistent with the simulation study results presented in Chapter 4, but the A1_cat3 reliabilities were greater than the reliabilities for the one-attribute, I12_cat3 simulation condition with an attribute correlation of .75, lower-quality items, and moderate growth shown in Table 15. Specifically, for this condition, the simulation had average reliabilities of .704, .625, and .267 for the polychoric, average maximum transition, and PB metrics, respectively. Comparing those values with the reliabilities of .941, .809, and .518 for the polychoric, average maximum transition, and PB metrics for the A1_cat3 analysis shows that the empirical data had much greater

reliabilities than the simulation study. This difference was likely because the empirical data analysis had 25 items to measure the attribute, while the simulation study had at most 12 items per attribute.

Figure 63 shows the proportion of students' maximum posterior probabilities that were greater than various checkpoints along the zero-to-one scale. Analogous to the finding that A1_cat3 had lower reliability than A4_cat2, Figure 63 shows that the PTDCM was less confident in student classifications for A1_cat3 than it was for A4_cat2 because the proportion of maximum posterior probabilities greater than specific values was smaller for A1_cat3 than it was for A4_cat2. It is likely that the decreased reliability and generally smaller maximum posterior probabilities were due to the unidimensional nature of A1_cat3. When diagnostic assessments are multidimensional with attributes that are moderately to highly correlated, the model can borrow diagnostic information between the attributes. In a unidimensional assessment, the model has no other source of diagnostic information other than the single attribute being measured, so it was not surprising that A1_cat3 showed lower reliabilities and confidence than A4_cat2. In sum, A1_cat3 had very good polychoric and acceptable average maximum transition reliability; therefore, the DGPs for A1_cat3 also had reliabilities that were acceptable to very good.

Figure 64 shows the seven basic DGPs for A1_cat3. Students who were beginning at the pre-test and beginning at the post-test had a basic DGP of .706. Students who were beginning at the pre-test and proficient at the post-test had a basic DGP of .996. Students who were proficient at the pre-test and beginning at the post-test had a basic DGP of .010. Students who were proficient at the pre-test and proficient at the post-test had a

basic DGP of .413. Students who were advanced at the pre-test and beginning at the post-test had a basic DGP of .014. Students who were advanced at the pre-test and proficient at the post-test had a basic DGP of .022. All students who were advanced at the post-test had a basic DGP of 1. Comparing Figures 64 and 45 highlights how the nature of the basic DGP changes when the number of proficiency statuses increases. Specifically, the basic DGPs for A1_cat3 had more variability than the basic DGPs for A4_cat2 because A1_cat3 had more proficiency statuses and, therefore, more transitions for each attribute.

Additionally, as mentioned in the previous section, the interpretations for the basic DGPs for A1_cat3 were more interesting than the interpretations for the basic DGPs for A4_cat2. Students with the transition [00] showed growth greater than or equal to (in this case, "equal to") 70.6% of students who were in the beginning group at the pre-test. Students with the transition [01] showed growth greater than or equal to 99.6% of students who were in the beginning group at the pre-test. Students with the transition [02] showed growth greater than or equal to 100% of students who were in the beginning group at the pre-test. Students with the transition [10] showed growth greater than or equal to (in this case, "equal to") 1% of students who were in the proficient group at the pre-test. Students with the transition [11] showed growth greater than or equal to 41.3% of students who were in the proficient group at the pre-test. Students with the transition [12] showed growth greater than or equal to 100% of students who were in the proficient group at the pre-test. Students with the transition [20] showed growth greater than or equal to (in this case, "equal to") 1.4% of students who were in the advanced group at the pre-test. Students with the transition [21] showed growth greater than or equal to 2.2% of students who were in the advanced group at the pre-test. Students with the transition [22]

showed growth greater than or equal to 100% of students who were in the advanced group at the pre-test.

Figure 65 shows the distributions for the adjusted DGPs for A1_cat3 conditional on students' basic DGPs. As with the box plots for A4_cat2, the primary purpose of this analysis was to ensure that adjusting the DGPs did not drastically change students' general locations on the zero-to-one scale. Other than the transitions that had very few students (the transitions with basic DGPs of .010, .014, and .022), the distributions of the box plots in Figure 65 all increased as the basic DGP increased, with some overlap between the distributions. Figure 65 shows that the distributions of the adjusted DGPs for each basic DGP were nearly identical, which means that in the A1_cat3 data, students' adjusted DGPs were nearly all identical. This can also be seen in Figure 66. Figure 66 shows the distributions of the adjusted DGPs at the student level. The student lines in Figure 66 were nearly flat for the CW, CWP, and PW DGPs. The absolute difference between students' CW DGPs and CWP DGPs ranges from 0 to .025, with mean and median values of .000 (not equal to zero but very near zero). I found comparable results when comparing students' CW and PW DGPs and CWP and PW DGPs. This *flat-line trend* for the adjusted DGPs occurred because the DGPs for the transitions with forgetting were all very small (i.e., .010, .014, and .022), and students' posterior probabilities of belonging to any of the forgetting transitions were very small, with mean and median values of .009 and .0001, respectively, across all students. So, when the adjusted DGPs weighted the basic DGPs by the posterior probabilities and added penalties for forgetting, the portions of the sums in the adjusted DGP metrics that corresponded with the forgetting transitions were so small that adding a penalty for

forgetting resulted in CWP DGPs and PW DGPs that were nearly identical to the CW DGPs. I do not expect that this flat-line trend will appear in all other DGP applications with polytomous attributes. However, it will appear in applications where, as desired, very few students are in forgetting transitions.

The final analysis I conducted on the A1_cat3 data set was for aggregating the DGPs. Because A1_cat3 was unidimensional, students' profile-level DGPs were equivalent to their attribute-level DGPs for the single attribute in the data. Therefore, I could further aggregate DGPs within each student. However, I aggregated the DGPs across all students to find the average basic and adjusted DGPs. For A1_cat3, the average basic DGP was .807, the average CW DGP was .789, the average CWP DGP was .798, and the average PW DGP was .789. These aggregated DGPs for A1_cat3 had values comparable to those obtained from aggregating the DGPs for A4_cat2, but notice that the averages of the adjusted DGPs were identical because of the flat line trend.

Empirical Data Analysis Discussion

These two empirical data analyses illustrated DGPs and their reliabilities with different types of attributes. In this chapter, I showed how students' DGPs changed with the adjustments for classification uncertainty and penalties for forgetting. I also showed students' real DGPs had reliabilities that were acceptable to excellent for these two data sets. The results from these analyses illustrate how to quantify student growth in a DCM framework, which focuses on their statistically determined classifications and transitions. Based on these analyses but without conducting any sort of focus groups with students or teachers, I would recommend that, when reporting students' attribute-level DGPs to stakeholders, the DGPs be accompanied by students' transition probabilities for each

attribute so detailed interpretations of students' relative growth can be achieved by being able to clearly explain their DGPs.

Although these analyses allowed for demonstrating the methods introduced in this dissertation, they had some key limitations. First, the empirical data sets I used in this study were not originally designed to be diagnostic assessments, so these analyses are considered *retrofitting* studies where DCMs are fit to empirical data from assessments that were designed to be used in a CTT or IRT framework, which is an ill-advised practice in DCM literature. However, as Bradshaw (2011) states, "development of psychometric theory of a model necessarily precedes its application" (p. 138). Therefore, until DCMs are more widely applied and more diagnostic assessments are created from the ground up in the DCM framework and administered in longitudinal settings, we have to settle for retrofitting studies to at least begin to illustrate the utility of developed methods and models, but the full utility remains to be unearthed.

A second key limitation of these empirical data analyses is that the only longitudinal data that I have access to are the A4_cat2 and A1_cat3 data sets, but ideally, I would have analyzed empirical data for a multidimensional assessment with polytomous attributes measured over at least three testing occasions. In the future, if I ever have data like this, I will apply the same analyses to show how the DGPs and their reliabilities perform in more complex assessment conditions.

CHAPTER 6

CONCLUSION

In this dissertation, I introduced a diagnostic version of the student growth percentile (SGP) metric called the diagnostic growth percentile (DGP). The DGP is intended to be used to quantify students' growth relative to their peers who experienced similar categorical gain scores, which in the longitudinal DCM framework manifest as students' most likely transitions of attribute proficiency for all measured attributes over time. I also introduced one adjustment to the basic DGP that factors in uncertainty in model classifications (i.e., consider the probability that students' most likely transitions are not their true transitions) and two adjustments to the basic DGP that additionally penalize students for having non-zero probabilities of belonging to a transition that includes forgetting (i.e., when the proficiency status decreases over time). To evaluate the reliability of these developed DGP metrics, I extended the TDCM reliability metrics (Madison, 2019; Schellman & Madison, in press) to the PTDCM framework to accommodate polytomous attributes. I conducted a simulation study and two empirical data analyses to investigate and illustrate the DGP metrics and their reliabilities.

In this last chapter, I reflect on the simulation study and empirical data analyses. I discuss what I learned from these studies and how they addressed the research questions presented in Chapter 1. I also discuss the significance of this dissertation as it fills gaps in the longitudinal DCM literature. Finally, I discuss the limitations of these studies and

plans for future research. This chapter also addresses the three key discussion questions presented in Chapter 1:

1. Which SGP limitations remain with DGPs?
2. Which SGP limitations are lessened when using DGPs?
3. Do DGPs have any new limitations that SGPs do not have?

Reflection on Simulation Study and Empirical Data Analyses

The simulation study aimed to evaluate how reliable DGPs can be and how the PTDCM reliability metrics perform under the manipulation of six assessment conditions: the item quality, the number of items to measure each attribute, the level of attribute proficiency growth, the attribute correlations, the number of proficiency statuses per attribute, and the number of attributes. Findings from the simulation study showed that, under some conditions, the PTDCM and, therefore, PTDCM classifications can be highly reliable with values greater than .8 and .9. Because DGPs come directly from the PTDCM person estimates, DGPs too can be highly reliable. However, consistent with previous PTDCM research (Madison et al., 2023), this simulation study yielded reliabilities that were generally lower than those of cross-sectional LCDM, TDCM, or PDCM reliabilities, but this result was expected because the PTDCM is a much more complex model than are the LCDM, TDCM, and PDCM.

In terms of how the PTDCM reliability metrics performed under the simulation conditions, the simulation study results were expected: For the one-attribute and three-attribute conditions and within each reliability metric, the conditions with the greatest reliabilities were the conditions with 12 higher-quality items measuring an attribute with three proficiency statuses and any level of growth, and of these conditions, the conditions

with greater attribute correlations generally showed greater reliabilities. The conditions that showed the lowest reliabilities were the conditions with eight lower-quality items measuring an attribute with five proficiency statuses.

Additionally, the polytomous polychoric reliability metric yielded the greatest reliabilities, and the polytomous PB metric yielded the lowest reliabilities. Growth did not have as strong of an influence on reliability as did the other manipulated factors: As the item quality increased, the number of items per attribute increased, the number of proficiency statuses decreased, and/or the attribute correlation increased, the reliabilities increased. However, the impact of item quality was greater than the impact of the number of proficiency statuses, which was greater than the impact of the number of items, which was greater than the impact of the attribute correlation.

The empirical data analyses aimed to illustrate how to compute DGPs for real data and to visualize and interpret DGPs. The empirical analyses demonstrate that DGPs can be used with dichotomous or polytomous attributes in unidimensional or multidimensional assessments. Specifically, these analyses showed that adjusting basic DGPs to account for uncertainty in classifications and to penalize for forgetting did not drastically change students' DGP in the sense that students who had low basic DGPs had low adjusted DGPs, and students who had high basic DGPs had high adjusted DGPs. In sum, the adjusted DGP metrics work because that they are logical, given students' most likely transitions and their entire profile of transition probabilities, as illustrated with the example students in Figure 54.

Together, the simulation study and empirical data analyses showed that the developed DGP metrics are sound metrics for quantifying students' relative growth for

specific applications in which validity arguments are prepared and thorough analyses of the appropriateness of the application of DGPs are conducted. No psychometric methods should ever be used without proper validity arguments for the specific intended use(s) desired by researchers, practitioners, and stakeholders.

Comparison of Student Growth Percentiles and Diagnostic Growth Percentiles

Sireci and colleagues (2016) wrote a report that highlighted six key limitations of SGPs and recommended that SGPs "should be abandoned and not used in education" (Sireci et al., 2916, p. 2). In this section, I discuss each of their six limitations in the context of DGPs to determine whether these limitations still exist, and if they do still exist, are they lessened or worsened with DGPs? Of course, DGPs have not yet been used in practice with teachers and students, so I cannot speak to how teachers and students might misinterpret DGPs, but I can predict potential sources of confusion based on the sources of confusion with SGPs.

Before discussing the six limitations from Sireci and colleagues' (2016) report, I first want to discuss the one new limitation that DGPs have and SGPs do not have: for a single attribute, the basic DGP does not differentiate between students who have the same transition. However, this limitation does not exist with the adjusted DGPs because the adjusted DGPs are true continuous variables that better account for students' individual differences via their posterior probabilities. Therefore, as mentioned previously in this dissertation, if the goal of using DGPs is to differentiate students using a continuous variable, then basic DGPs should not be used. However, if the diagnostic assessment is multidimensional, then the profile-level basic DGP can be used to differentiate students.

Another nuance with this limitation is counter to one of the benefits of SGPs: SGPs give students credit for growth not reflected in their categorical growth (e.g., changes in their achievement levels over time). On the other hand, DGPs (basic and adjusted) focus only on categorical growth via students' most likely transitions from the DCM, so DGPs do not factor in total or scaled scores that could give any more specific "credit" for student growth. This is not a limitation of DGPs, however, because the purpose of DCMs is to estimate students' most likely transitions, and DGPs are relative measures of students' transitions in comparison with other students' transitions. Stakeholders who use DCMs and who might use DGPs someday are not concerned with students' total scores or estimating scaled scores. However, the adjusted DGPs do give students credit for having non-zero probabilities of belonging to other transitions, which would increase their DGPs if the other transitions have greater basic DGPs, but this is not the same type of "credit" that people mean when they are talking about SGPs.

The first limitation of SGPs is that "SGPs are not what people think they are" (Sireci et al., 2016, p. 2), and these authors discuss two primary issues here: (1) stakeholders do not understand how to compute SGPs (i.e., quantile regression versus cohort approaches) and (2) the words "growth" and "percentile" are misleading. To address issue (1), one of the key benefits of DGPs over SGPs is that quantile regression is not required. Basic DGPs come directly from students' transitions, and the computation for the adjusted DGPs is a simple manipulation of the basic DGPs. However, DGP computation requires that the TDCM or PTDCM is used to obtain students' transitions and posterior probabilities, and proper application of the TDCM or TPDCM requires a psychometrician's expertise. This should not be a problem because DGPs are intended to

be used for assessments that are already using DCMs as the psychometric modeling framework, and such assessments should already employ psychometricians who specialize in DCMs, so computing DGPs would be a simple addition to their post-calibration analyses.

To address issue (2), "growth" and "percentile" are part of the name for DGPs because I wanted to use a name that directly linked the SGP and DGP literature, as the DGP is an adaptation and modification of the SGP for the DCM framework. However, the limitations related to the nomenclature described by Sireci and colleagues (2016) remain with DGPs. First, using DGPs to interpret growth is different from interpreting growth from categorical gain scores, which in the DCM framework are transitions. For example, using transitions to interpret growth allows us to say that Student X with the transition [01] for Attribute 1 showed more growth than Student Y with the transition [00] for Attribute 1 because Student X moved from non-proficiency to proficiency for Attribute 1 by T2, but Student Y did not yet move from non-proficiency to a greater proficiency status. DGPs do not allow for such comparisons because large DGP values do not always indicate large categorical gain scores, and small DGP values do not always indicate small categorical gain scores. Consider the example students in Figure 54. Student 256 (the light green line) most likely had the transition [11], which, from a gain score perspective, would indicate no growth because Student 256's proficiency status for the attribute did not change. Student 835 (the green line) most likely had the transition [00], which would also indicate no growth from a gain score perspective. Student 256 had a greater basic DGP than Student 835, but Student 835 had greater adjusted DGPs than Student 256 even though a gain score perspective would have shown zero growth for

both students. Additionally, Student 185 most likely had the transition [01], which indicates positive growth from a gain score perspective, but Student 185 and Student 256 had the same basic DGP even though Student 256 showed no growth from a gain score perspective. Thus, "growth" in the name of DGPs does not give any information about students' growth from a gain score perspective. Finally, "percentile" in the name for DGPs does not indicate anything about the computation for DGPs, which will likely be a source of confusion if/when DGPs are applied in practice.

The second limitation of SGPs is that "SGPs are unreliable" (Sireci et al., 2016, p. 2). Because DGPs are computed using metrics provided by the PTDCM, DGP reliability is PTDCM reliability, and my simulation study showed the assessment conditions that impact PTDCM reliability. DGP reliability can be high under proper assessment conditions (i.e., high-quality items, more items per attribute, fewer proficiency statuses per attribute, and/or greater attribute correlations). Within this second limitation, Sireci and colleagues (2016) also discuss issues with aggregating DGPs across students to obtain class-level SGPs, and they cited studies that found that the margin of error for SGPs spans about half of the 99-point SGP scale. My empirical data analysis showed different results when aggregating DGPs across students: As discussed previously, the standard errors for the mean DGPs in Figures 56 and 57 were at most .016, which was very small. So, the results of this dissertation did not indicate that DGPs should not be aggregated across students for group-level comparisons.

The third limitation of SGPs is that "Educators do not understand how to use SGPs" (Sireci et al., 2016, p. 2), and the primary issue these authors discuss with this limitation is that stakeholders use SGPs to evaluate schools and teachers and to draw

conclusions that would be appropriate from gain scores—not from SGPs (e.g., identify students who need remediation). Luckily, I conducted this dissertation prior to DGPs being used or needed in practice, which was not the case with much of the research for SGPs. Therefore, I hope to be able to head off some of these types of misuse and misinterpretations through comprehensive training and educational materials about what exactly DGPs are, how they are intended to be used, and the kinds of validity arguments that should be provided to support the use of DGPs for a specific situation.

The fourth limitation of SGPs is that "There is no validity evidence to support the use of SGPs" (Sireci et al., 2016, p. 2), and "we did not find any empirical studies…that provided positive results to defend the use of SGPs" (Sireci et al., 2016, p. 7). This dissertation provides two empirical data analyses that support the use of DGPs in practice. However, as I have stated several times throughout this dissertation, each proposed application of DGPs should be preceded by a thorough validity argument that provides evidence in support of the intended use. Therefore, although my dissertation demonstrates the DGPs and their reliabilities for the particular data sets I used, my results should not be used as arguments in support of any other use without being accompanied by other validity evidence for the particular use.

The fifth limitation of SGPs is that "Current use of SGPs violates the *Standards*" (Sireci et al., 2016, p. 2), and these authors specifically note that SGPs have been inappropriately used for teacher evaluation without proper validity evidence. To be clear, I do not recommend that DGPs be used for teacher evaluation. Again, this dissertation precedes the use of DGPs in practice, which is the ideal pattern of events, so hopefully, proper training and education about DGPs and my intended uses can minimize the

improper use of DGPs. However, no amount of training or documentation can completely eradicate all improper uses and misinterpretations of any psychometric method.

The sixth limitation of SGPs is that "SGPs encourage comparing students to each other, rather than to knowledge and skill areas they are being taught" (Sireci et al., 2016, p. 2). This limitation is critical and is still a prominent issue in the DGP framework. One of the primary features of DCMs is that they are strictly criterion-referenced and provide results that show individual students' strengths and areas for improvement with respect to the measured attributes. Creating a diagnostic version of the SGP (i.e., the purpose of this entire dissertation) is counter to the primary argument in favor of DCMs. As stated at the beginning of this dissertation, I am not yet recommending that DGPs be used in practice or that DGPs are necessarily good metrics to consider. The intention of this dissertation is to explore the possibility of such a metric to see if it would work or make sense. The studies I conducted in this dissertation showed that DGPs do work and can be highly reliable. However, whether DGPs should move from the realm of research to practice is a question that I cannot answer. I believe that more research needs to be conducted prior to the use of DGPs in practice, and validity arguments should indicate whether or not DGPs are feasible or appropriate for each particular use.

Limitations and Future Research

I would like to briefly recap the key limitations discussed in my simulation study and empirical data analyses. First, as discussed at the end of Chapter 3, I have a couple of disclaimers to make for the adjusted DGP metrics: (1) the use of posterior probabilities in the computation of a metric designed to rank-order student growth may be considered a questionable practice, and (2) it may be inappropriate to borrow the PTDCM reliability

for all four types of DGPs. To address Disclaimer #1, I argued that the adjusted DGP metrics are not inappropriately using posterior probabilities as a means of rank-ordering students because the adjusted DGPs consider the entire distribution of a student's most likely class rather than only part of that distribution (e.g., a marginal or maximum posterior probability), and the posterior probabilities are used in the adjusted DGPs to weight the basic DGPs, so the adjusted DGPs do not use posterior probabilities directly to rank-order students. To address Disclaimer #2, I explained that more research is needed to determine whether this new use of posterior probabilities is appropriate for specific contexts and whether there are other approaches for conceptualizing DGP reliability.

In addition to re-emphasizing that this dissertation is meant to be an exploration of the efficacy and validity of these ideas, and I do not intend for DGPs to be put into practice immediately, I would like to make one final point related to these disclaimers: as psychometricians, we are often challenged to balance model design and theory with stakeholders' needs, which sometimes violate the intended uses of psychometric methods. For example, stakeholders desired classifications from IRT-based assessments, so psychometricians figured out methods for extracting classifications from scaled-scores even though classification was not an intended use of IRT-based results. In this dissertation, I used learnings from IRT and SGP literature to anticipate the types of requests stakeholders might make for DCM-based results (i.e., a metric for comparing student growth; an adjustment to that metric to account for model uncertainty; an additional adjustment to that metric to penalize students for forgetting), and I figured out ways to accommodate those requests even though DCMs are intended to be criterion-referenced and avoid relative student comparisons. Thus, some of the work in this

dissertation does not naturally follow from the DCM literature, so it may be called into question. However, with additional research in the area of DGPs, we will be able to learn whether DGPs are truly appropriate for practical use or if they should remain theoretical.

Future research could explore an adaptive testing approach to mitigate the potential issues related to using posterior probabilities in the computation of metrics designed for comparing student growth (M. Madison, personal communication, July 2, 2024). In this adaptive design, students would be required to continue responding to items until their maximum posterior probability is greater than a desired level. This adaptive design would guarantee that the model is acceptably confident in all students' classifications, so the adjusted DGPs would not be necessary.

In terms of the simulation study, I was limited in the number of conditions I could include, as is the nature of all simulation studies. I was also limited because the software for estimating the PTDCM is limited due to the complexity of some PTDCM conditions. In future research, I would like to conduct a simulation study with more proficiency statuses for more attributes measured at more testing occasions and with complex item types, but such research depends on advancements in DCM estimation algorithms and software. Additionally, I did not investigate model misfit or misspecification in my simulation study. Future research should evaluate the extent to which the DGPs and their reliabilities are impacted by different types of model misfit and misspecification.

In terms of the empirical data analysis, the key limitations were (1) the data came from assessments that were not designed to be used with DCMs and (2) I could not simultaneously illustrate DGPs with multidimensional assessments with polytomous attributes or illustrate the DGPs for data with more than two testing occasions. When data

from a longitudinal diagnostic assessment is available and when I can find longitudinal data with multiple polytomous attributes (whether it is data from a diagnostic assessment or not), I will run the same empirical data analyses I ran for this dissertation to see how the DGPs and DGP reliability metrics perform with diagnostic data and with multiple polytomous attributes.

Conclusion

The basic DGP and its adjustments are viable approaches for comparing student growth in the DCM framework, but more research is needed in this area. My results highlighted a broad finding and recommendation with respect to applied work with the PTDCM: assessment developers do not need to make assessments longer to achieve high levels of reliability with the PTDCM—they need to make the items more discriminating. Increasing the number of items will not help as much as increasing the item quality, especially as the number of proficiency statuses increases, which may be desirable in practice because having a greater number of proficiency statuses provides more detailed feedback about student proficiencies. Increasing the number of proficiency statuses comes at a cost: Teaching and learning time is valuable and limited. Increasing test lengths will increase the amount of time students spend taking assessments, and the gain is not enough to warrant the time needed for longer assessments. Assessment developers who intend to use the PTDCM must be extremely diligent about creating thorough item development processes and train item writers to develop high-quality items that efficiently support student classifications for polytomous attributes over time.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Azen, R., & Walker, C. M. (2011). *Categorical Data Analysis for the Behavivorial and Social Sciences.* Taylor & Francis Group.

Bao, Y. (2019). A diagnostic classification model for polytomous attributes. (Unpublished doctoral dissertation). University of Georgia, Athens, GA.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51.

Betebenner, D. W., Van Iwaarden, A. R., Domingue, B., & Shang, Y. (2022). *SGP: Student Growth Percentiles & Percentile Growth Trajectories* (Version 2.0-0.0) [R package]. https://sgp.io

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-74.

Bottge, B. A., Heinrichs, M., Chan, S.-Y., & Serlin, R. C. (2001). Anchoring adolescents' understanding of math concepts in rich problem-solving environments. *Remedial and Special Education, 22*(5), 299–314.

Bottge, B. A., Ma, X., Gassaway, L., Toland, M. D., Butler, M., & Cho, S.-J. (2014). Effects of blended instructional models on math performance. *Exceptional Children, 80*(4), 423-437.

Bottge, B. A., Toland, M. D., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., & Ma, X. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Exceptional Children, 81*(2), 158–175.

Bradshaw, L. P. (2011). Combining scaling and classification: A psychometric model for scaling ability and diagnosing misconceptions. [Unpublished doctoral dissertation]. University of Georgia.

Bradshaw, L. (2016). Diagnostic classification models: A multivariate classification approach for cognitively complex assessment. In A. Rupp, & J. Leighton (Eds.), *Handbook of Cognition and Assessment* (pp. 297-326). Wiley-Blackwell.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2-14.

Bradshaw, L., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing, 16*(2), 99–118. https://doi.org/10.1080/15305058.2015.1107076.

Bradshaw, L., Masters, J., Famularo, L., Lee, H., & Azevedo, R. (2017-2021). *Diagnostic Inventories of Cognition in Education (DICE)* (Project No. R305A170441)

Bramlett, S. A. (2018). A method for detecting measurement invariance in the log-linear cognitive diagnosis model. [Unpublished doctoral dissertation]. University of Georgia.

Castellano, K. E., & Ho, A. D. (2013a). A practitioner's guide to growth models. In *Council of Chief State School Officers*.

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics, 38*(2), 190-215.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi:10.18637/jss.v048.i06.

Clauser, A.L., Keller, L.A., McDermott, K.A. (2016). Principals' uses and interpretations of student growth percentile data. *Journal of School Leadership, 26*(1), 6-33.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Group.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163-183.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*(1), 115-127.

Dong, Y., Ma, X., Wang, C., & Gao, X. (2021). An optimal choice of cognitive diagnostic model for second language listening comprehension test. *Frontiers in Psychology, 12*.

Dynamic Learning Maps Consortium (2022, December). *2021-22 Technical manual – year-end model.* University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://2022-ye-techmanual.dynamiclearningmaps.org/

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Every Student Succeeds Act, Pub. L. No. 114-95, 129 Stat. 1802 (2015).

Ford, C. (2015, September 20). *Getting started with quantile regression.* University of Viriginia Library Research Data Services + Sciences. https://data.library.virginia.edu/getting-started-with-quantile-regression/

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1-24. doi:10.18637/jss.v074.i02.

Georgia Department of Education (n.d.). *Understanding the Georgia Milestones achievement levels.* https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/achievement_levels.aspx

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

Harrison, A. J., Bradshaw, L. P., Naqvi, N. C., Paff, M. L., Campbell, J. M. (2017). Development and pscyhometric evaluation of the autism stigma and knowledge questionnaire. *Journal of Autism and Developmental Disorders, 47*, 3281-3295.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 72*(4), 191-210.

Jang, E. E. (2005). A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL. [Unpublished doctoral dissertation]. University of Illinois at Urbana–Champaign.

Johnson, M. S., & Sinharay, S. (2020). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics, 45*(1), 5-31.

Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing, 14*, 49-72.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Koenker, R. (2022). *quantreg: Quantile Regression* (Version 5.94) [R package].

Madaus, G., Russell, M, & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society.* Information Age Publishing Inc.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1-26. doi:10.18637/jss.v093.i14.

Madison, M. J. (2019). Reliably assessing growth with longitudinal diagnostic classification models. *Educational Measurement: Issues and Practice, 38*(2), 68-78.

Madison, M. J., & Bao, Y. (July, 2018). *A longitudinal and polytomous diagnostic classification model.* [Paper presentation]. International Meeting of Psychometric Society.

Madison, M. J., Bao, Y., Chung, S., Kim, J., & Bradshaw, L. (June, 2021). *A longitudinal DCM with polytomous attributes.* [Paper presentation]. National Council on Measurement in Education.

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika, 83*(4), 963-990.

Madison, M. J., Chung, S., Kim, J., & Bradshaw, L. P. (2023). Approaches to estimating longitudinal diagnostic classification models. *Behaviormetrika.* https://doi.org/10.1007/s41237-023-00202-5.

Michigan Department of Education. (n.d.). *Student growth percentiles (SGPS).* https://www.michigan.gov/mde/Services/ed-serv/Educator-Retention-Supports/educator-eval/student-growth/student-growth-percentiles-sgps

Monroe, S., & Cai, L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. *Educational Measurement: Issues and Practice, 34*(4), 21-30.

Monroe, S., Cai, L., & Choi, K. (2014). Student growth percentiles based on MIRT: Implications of calibrated projection. CRESST Report 842. *In National Center for Research on Evaluation, Standards, and Student Testing (CRESST).* National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus user's guide (8th ed.) Los Angeles, CA: Muthén & Muthén.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262.

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

Schellman, M. A. (2021). Diagnostic concept inventories for misconception classification accuracy and reliability. [Unpublished master's thesis]. University of Georgia.

Schellman, M., & Madison, M. (July, 2021). *Estimating the reliability of skill transition in longitudinal DCMs*. [Paper presentation]. International Meeting of the Psychometric Society.

Schellman, M., & Madison, M. (in press). Estimating the reliability of skill transitions in longitudinal diagnostic classification models. *Journal of Educational and Behavioral Statistics*.

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for student growth percentiles using the SIMEX method. *Educational Measurement: Issues and Practice, 34*.

Sireci, S. G., Wells, C. G., & Keller, L. A. (2016). Why we should abandon student growth percentiles. Research Brief 16-1. *Center for Educational Assessment.* Center for Educational Assessment University of Massachusetts Amherst.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model student estimates. *Journal of Classification, 30*, 251-275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37-50.

Wells, C. G., & Sireci, S. G. (2020). Evaluating random and systematic error in student growth percentiles. *Applied Measurement in Education, 33*(4), 349-361.

Yu, X., Zhan, P., & Chen, Q. (2023). Don't worry about the anchor-item setting in longitudinal learning diagnostic assessments. *Frontiers in Psychology, 14*.

Table 1

Summary of Notation and Indices

Component of the DCM Framework	Notation for Total Number	Notation for Index
Attributes	A	$a \in \{1, \ldots, A\}$
Students	N	$r \in \{1, \ldots, N\}$
Items	I	$i \in \{1, \ldots, I\}$
Testing Occasions	T	$t \in \{1, \ldots, T\}$
Latent Classes	C	$c \in \{1, \ldots, C\}$
Attribute-Level Transitions	C'	$c' \in \{1, \ldots, C'\}$
Proficiency Statuses	l_a	$\alpha_{ca} \in \{1, \ldots, l_a\}$

Table 2

Dummy Coding for a PDCM Attribute with Five Proficiency Statuses

Proficiency Status Label	Latent Class (c)	Attribute Profile (α_{ca})	Dummy Variables			
			α_{ca}^{1}	α_{ca}^{2}	α_{ca}^{3}	α_{ca}^{4}
Beginning	1	0	0	0	0	0
Developing	2	1	1	0	0	0
Proficient	3	2	1	1	0	0
Distinguished	4	3	1	1	1	0
Advanced	5	4	1	1	1	1

Table 3

Log Odds of a Correct Response for Each Proficiency Status for an Attribute with Five Proficiency Statuses in the PDCM Framework

Latent Class	Dummy Variables				Log Odds of a Correct Response
(c)	α_{ca}^{1}	α_{ca}^{2}	α_{ca}^{3}	α_{ca}^{4}	
1	0	0	0	0	$\lambda_{i,0}$
2	1	0	0	0	$\lambda_{i,0} + \lambda_{i}^{1}$
3	1	1	0	0	$\lambda_{i,0} + \lambda_{i}^{1} + \lambda_{i}^{2}$
4	1	1	1	0	$\lambda_{i,0} + \lambda_{i}^{1} + \lambda_{i}^{2} + \lambda_{i}^{3}$
5	1	1	1	1	$\lambda_{i,0} + \lambda_{i}^{1} + \lambda_{i}^{2} + \lambda_{i}^{3} + \lambda_{i}^{4}$

Table 4

Summary of Computation for the Number of Latent Classes and Attribute-Level Transitions

Attribute Type	Number of Testing Occasions	Model	Restrictions on l_a Over Time	Number of Latent Classes (C)	Number of Attribute-Level Transitions
Dichotomous	$T = 1$	LCDM	-	$C = 2^A$	-
Dichotomous	$T > 1$	TDCM	-	$C = 2^{AT}$	$C' = 2^T$
Polytomous	$T = 1$	PDCM	-	$C = \prod_{a=1}^{A} l_a$	-
Polytomous	$T > 1$	PTDCM	Can vary for each attribute*	$C = \prod_{t=1}^{T} \prod_{a=1}^{A} l_{at}$	$C'_a = \prod_{t=1}^{T} l_{at}$
Polytomous	$T > 1$	PTDCM	Stays the same for each attribute*	$C = \prod_{a=1}^{A} l_a^T$	$C'_a = l_a^T$

Note. *l_a need not be the same for all attributes in the assessment. Refer to Table 1 for the meaning of the notation used here.

Table 5

Transitions for an Assessment that Measures One Attribute with Five Proficiency Statuses at Two Testing Occasions

Most Likely Class	Attribute Profile at T1	Attribute Profile at T2	Attribute Profile/ Transition
1	0	0	00
2	0	1	01
3	0	2	02
4	0	3	03
5	0	4	04
6	1	0	10
7	1	1	11
8	1	2	12
9	1	3	13
10	1	4	14
11	2	0	20
12	2	1	21
13	2	2	22
14	2	3	23
15	2	4	24
16	3	0	30
17	3	1	31
18	3	2	32
19	3	3	33
20	3	4	34
21	4	0	40
22	4	1	41
23	4	2	42
24	4	3	43
25	4	4	44

Note. “T1” = first testing occasion, and “T2” = second testing occasion.

Table 6

Basic Diagnostic Growth Percentiles for a Trichotomous Attribute Measured at Two Testing Occasions

Cohort/ Transition	Computation for Basic DGP	Interpretation
[00]	$D_{B_{a,00}} = \frac{p_{a,00}}{p_{a,00} + p_{a,01} + p_{a,02}}$	Proportion of students who were non-proficient at T1 who are at least non-proficient at T2
[01]	$D_{B_{a,01}} = \frac{p_{a,00} + p_{a,01}}{p_{a,00} + p_{a,01} + p_{a,02}}$	Proportion of students who were non-proficient at T1 who are at least partially proficient at T2
[02]	$D_{B_{a,02}} = \frac{p_{a,00} + p_{a,01} + p_{a,02}}{p_{a,00} + p_{a,01} + p_{a,02}}$	Proportion of students who were non-proficient at T1 who are at least proficient at T2
[10]	$D_{B_{a,10}} = \frac{p_{a,10}}{p_{a,10} + p_{a,11} + p_{a,12}}$	Proportion of students who were partially proficient at T1 who are at least non-proficient at T2
[11]	$D_{B_{a,11}} = \frac{p_{a,10} + p_{a,11}}{p_{a,10} + p_{a,11} + p_{a,12}}$	Proportion of students who were partially proficient at T1 who are at least partially proficient at T2
[12]	$D_{B_{a,12}} = \frac{p_{a,10} + p_{a,11} + p_{a,12}}{p_{a,10} + p_{a,11} + p_{a,12}}$	Proportion of students who were partially proficient at T1 who are at least proficient at T2
[20]	$D_{B_{a,20}} = \frac{p_{a,20}}{p_{a,20} + p_{a,21} + p_{a,22}}$	Proportion of students who were proficient at T1 who are at least non-proficient at T2
[21]	$D_{B_{a,21}} = \frac{p_{a,20} + p_{a,21}}{p_{a,20} + p_{a,21} + p_{a,22}}$	Proportion of students who were proficient at T1 who are at least partially proficient at T2
[22]	$D_{B_{a,22}} = \frac{p_{a,20} + p_{a,21} + p_{a,22}}{p_{a,20} + p_{a,21} + p_{a,22}}$	Proportion of students who were proficient at T1 who are at least proficient at T2

Note Attribute a has the proficiency statuses "non-proficient", "partially proficient", and "proficient". "T1" = first testing occasion, and "T2" = second testing occasion. $p_{a,\alpha_1\alpha_2}$ is the proportion of all students who have the transition $[\alpha_1\alpha_2]$ for Attribute a, where α_1 is the proficiency status at T1 and α_2 is the proficiency status at T2.

Table 7

Basic Diagnostic Growth Percentiles for a Dichotomous Attribute Measured at Three Testing Occasions

Cohort/ Transition	Computation for Basic DGP	Interpretation
[000]	$D_{B_{a,000}} = \frac{p_{a,000}}{p_{a,000} + p_{a,001}}$	Non-proficient at T1, non-proficient at T2, and at least non-proficient at T3
[001]	$D_{B_{a,001}} = \frac{p_{a,000} + p_{a,001}}{p_{a,000} + p_{a,001}}$	Non-proficient at T1, non-proficient at T2, and at least proficient at T3
[010]	$D_{B_{a,010}} = \frac{p_{a,010}}{p_{a,010} + p_{a,011}}$	Non-proficient at T1, proficient at T2, and at least non-proficient at T3
[011]	$D_{B_{a,011}} = \frac{p_{a,010} + p_{a,011}}{p_{a,010} + p_{a,011}}$	Non-proficient at T1, proficient at T2, and at least proficient at T3
[101]	$D_{B_{a,110}} = \frac{p_{a,100}}{p_{a,100} + p_{a,101}}$	Proficient at T1, non-proficient at T2, and at least non-proficient at T3
[101]	$D_{B_{a,101}} = \frac{p_{a,100} + p_{a,101}}{p_{a,100} + p_{a,101}}$	Proficient at T1, non-proficient at T2, and at least proficient at T3
[110]	$D_{B_{a,110}} = \frac{p_{a,110}}{p_{a,110} + p_{a,111}}$	Proficient at T1, proficient at T2, and at least non-proficient at T3
[111]	$D_{B_{a,111}} = \frac{p_{a,110} + p_{a,111}}{p_{a,110} + p_{a,111}}$	Proficient at T1, proficient at T2, and at least proficient at T3

Note Attribute a has the proficiency statuses "non-proficient" and "proficient". . "T1" = first testing occasion, "T2" = second testing occasion, and "T3" = third testing occasion. $p_{a,\alpha_1\alpha_2\alpha_3}$ is the proportion of all students who have the transition $[\alpha_1\alpha_2\alpha_3]$ for Attribute a, where α_1 is the proficiency status at T1, α_2 is the proficiency status at T2, and α_3 is the proficiency status at T3. Due to space limitations, this table excludes "Proportion of students who were" from the beginning of each interpretation.

Table 8

Simulation Study Design

Factor	Number of Levels	Levels
Generating Model	1	PTDCM
Sample Size	1	1,000
Item Type	1	Simple structure
Base Rate at T1	1	.6, .25, .15 (three proficiency statuses) .45, .25, .15, .1, .05 (five statuses)
Number of Testing Occasions	1	Two testing occasions: pre-/post-test
Item Quality	2	Lower or higher
Test Length	2	8 or 12 items per attribute at each testing occasion
Attribute Proficiency Growth	3	No growth, moderate growth, large growth
Attribute Correlations	4	0, .25, .5, or .75 (between and within)
Number of Proficiency Statuses Per Attribute	2	3 or 5 proficiency statuses
Number of Attributes	2	1 or 3* attributes
Estimating DGP Model	4	Basic, CW, CWP, and PW DGPs
Estimating Model	1	PTDCM
Reliability Metric	3	Polychoric, average maximum transition, and PB metrics
TOTAL	144 generation; 4+3 evaluation; 1008 conditions	

Note. *The conditions with three attributes only have three proficiency statuses due to the computational demand of having multiple attributes each with five proficiency statuses. Therefore, the manipulated conditions are not fully crossed.

Table 9

Summary Statistics for Generated Item Discriminations

Number of Attributes	Number of Proficiency Statuses	Item Quality	IRP Disc. Type	Min	Mean	Median	Max
1	3	Lower	$d_{3,10}$	.099	.202	.201	.294
			$d_{3,21}$	.102	.202	.202	.292
			$d_{3,20}$	.271	.403	.404	.536
		Higher	$d_{3,10}$	.212	.299	.300	.405
			$d_{3,21}$	.210	.299	.299	.413
			$d_{3,20}$	.466	.599	.599	.727
	5	Lower	$d_{5,10}$	.026	.124	.124	.232
			$d_{5,21}$	.026	.125	.125	.235
			$d_{5,32}$	.016	.125	.125	.226
			$d_{5,43}$	.026	.125	.125	.219
			$d_{5,40}$	.322	.499	.499	.689
		Higher	$d_{5,10}$	.079	.173	.173	.275
			$d_{5,21}$	.090	.172	.172	.262
			$d_{5,32}$	.071	.173	.173	.268
			$d_{5,43}$	.078	.172	.173	.261
			$d_{5,40}$	.504	.690	.690	.868
3	3	Lower	$d_{3,10}$	.103	.202	.202	.301
			$d_{3,21}$	.102	.202	.202	.305
			$d_{3,20}$	.264	.403	.403	.550
		Higher	$d_{3,10}$	.191	.299	.299	.411
			$d_{3,21}$	.196	.299	.299	.398
			$d_{3,20}$	.451	.598	.598	.736

Note. For the conditions with three proficiency statuses, $d_{3,10}$ is Proficient vs Beginning; $d_{3,21}$ is Advanced vs Proficient; $d_{3,20}$ is Advanced vs Beginning. For the conditions with five proficiency statuses, $d_{5,10}$ is Developing vs Beginning; $d_{5,21}$ is Proficient vs Developing; $d_{5,32}$ is Distinguished vs Proficient; $d_{5,43}$ is Advanced vs Distinguished; $d_{5,40}$ is Advanced vs Beginning.

Table 10

Base Rates for the Different Levels of Growth for Attributes with Five Proficiency Statuses in the Simulation Study

Proficiency Status	Base Rate at T1	Base Rate at T2		
		No Growth	Moderate Growth	Large Growth
Beginning	.45	.45	.05	.05
Developing	.25	.25	.1	.1
Proficient	.15	.15	.45	.15
Distinguished	.1	.1	.25	.25
Advanced	.05	.05	.15	.45

Note. "T1" = first testing occasion, and "T2" = second testing occasion.

Table 11

Summary Statistics for Generated Attribute Correlations

Target Attribute Correlation	Minimum	Mean	Median	Maximum
0	-.161	-.001	-.001	.188
.25	.039	.250	.251	.417
.5	.328	.499	.500	.624
.75	.657	.750	.750	.858

Table 12

Convergence Rates for the Simulation Study

Number of Attributes	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
1	Lower	0	1	.740	.980	.940	.940	.860	1	.820	.920	.940	.960	.860
		.25	.960	.920	.980	.860	1	.800	1	.880	.980	.880	.980	.840
		.5	.880	.880	.980	.960	.980	.840	1	.860	.940	.880	.980	.860
		.75	.960	.880	1	.860	1	.860	1	.940	1	.820	.960	.820
	Higher	0	.960	.860	1	.720	1	.920	1	.800	1	.760	1	.840
		.25	1	.880	1	.860	1	.900	1	.880	.980	.900	1	.780
		.5	1	.980	1	.800	.980	.940	1	.840	1	.840	1	.920
		.75	.980	.920	1	.900	1	.940	1	.900	1	.840	1	.760
3	Lower	0	.760	-	.720	-	.720	-	.840	-	.740	-	.840	-
		.25	.780	-	.760	-	.680	-	.860	-	.640	-	.780	-
		.5	.780	-	.880	-	.680	-	.780	-	.640	-	.780	-
		.75	.800	-	.820	-	.800	-	.880	-	.740	-	.780	-
	Higher	0	.780	-	.940	-	.800	-	.940	-	.740	-	.940	-
		.25	.680	-	.880	-	.720	-	.980	-	.760	-	.880	-
		.5	.800	-	.960	-	.800	-	1	-	.920	-	.980	-
		.75	.960	-	.960	-	.840	-	.920	-	.880	-	.960	-

Note. "I8_cat3" = conditions with eight items and three proficiency statuses per attribute. "I8_cat5" = conditions with eight items and five proficiency statuses per attribute. "I12_cat3" = conditions with 12 items and three proficiency statuses per attribute. "I12_cat5" = conditions with 12 items and five proficiency statuses per attribute.

Table 13

Item Parameter Estimation Accuracy Results: Average Mean Absolute Difference Between True and Estimated IRPs for the One-Attribute Conditions

IRP Type	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
All	Lower	0	.056	.084	.053	.073	.051	.081	.034	.070	.062	.095	.056	.087
		.25	.059	.081	.053	.069	.049	.080	.035	.075	.057	.086	.051	.087
		.5	.064	.084	.053	.071	.048	.075	.031	.070	.060	.088	.044	.088
		.75	.057	.073	.048	.066	.036	.069	.027	.064	.046	.080	.038	.083
	Higher	0	.047	.076	.025	.064	.023	.073	.017	.074	.030	.080	.019	.097
		.25	.039	.072	.024	.065	.023	.070	.018	.072	.036	.085	.020	.096
		.5	.034	.071	.025	.067	.022	.068	.019	.069	.025	.082	.018	.090
		.75	.032	.069	.023	.063	.020	.065	.018	.068	.023	.081	.017	.084
IRP0	Lower	0	.020	.027	.016	.023	.022	.030	.017	.025	.026	.036	.022	.027
		.25	.018	.027	.016	.021	.020	.028	.018	.027	.022	.033	.022	.026
		.5	.024	.025	.017	.023	.022	.029	.018	.024	.025	.030	.022	.027
		.75	.018	.023	.016	.020	.019	.025	.017	.023	.023	.027	.019	.022
	Higher	0	.019	.026	.012	.019	.016	.028	.015	.026	.019	.030	.014	.026
		.25	.015	.024	.013	.019	.016	.028	.015	.026	.019	.027	.015	.027
		.5	.014	.023	.012	.019	.017	.027	.015	.025	.017	.025	.014	.022
		.75	.014	.021	.011	.018	.015	.026	.015	.024	.015	.023	.014	.023
IRP1	Lower	0	.089	.094	.080	.089	.051	.111	.037	.091	.106	.120	.095	.106
		.25	.088	.091	.078	.077	.053	.104	.038	.098	.097	.112	.087	.098
		.5	.101	.098	.076	.089	.054	.096	.034	.095	.100	.109	.075	.102
		.75	.091	.085	.071	.080	.043	.095	.027	.088	.076	.093	.068	.090

	Higher	0	.079	.091	.037	.084	.027	.117	.018	.121	.050	.104	.030	.123
		.25	.062	.088	.034	.082	.026	.116	.019	.118	.066	.121	.031	.123
		.5	.053	.079	.037	.083	.024	.109	.021	.112	.041	.107	.027	.106
		.75	.051	.080	.031	.077	.021	.103	.019	.109	.038	.104	.026	.094
IRP2	Lower	0	.058	.086	.065	.074	.081	.079	.048	.080	.053	.126	.050	.129
		.25	.070	.089	.065	.074	.075	.078	.048	.077	.052	.117	.043	.132
		.5	.068	.085	.066	.072	.067	.078	.042	.073	.055	.123	.037	.128
		.75	.062	.075	.056	.071	.045	.071	.036	.068	.037	.113	.027	.129
	Higher	0	.042	.091	.026	.074	.027	.075	.020	.087	.019	.107	.013	.170
		.25	.040	.084	.024	.079	.026	.071	.020	.085	.021	.122	.014	.163
		.5	.034	.083	.027	.086	.025	.075	.021	.082	.016	.122	.012	.159
		.75	.031	.082	.025	.081	.022	.074	.020	.087	.016	.118	.012	.150
IRP3	Lower	0	-	.092	-	.076	-	.083	-	.072	-	.094	-	.093
		.25	-	.092	-	.075	-	.083	-	.076	-	.087	-	.098
		.5	-	.092	-	.077	-	.076	-	.071	-	.093	-	.093
		.75	-	.088	-	.076	-	.076	-	.070	-	.093	-	.096
	Higher	0	-	.107	-	.088	-	.089	-	.088	-	.117	-	.125
		.25	-	.111	-	.088	-	.087	-	.086	-	.114	-	.125
		.5	-	.106	-	.088	-	.083	-	.082	-	.110	-	.125
		.75	-	.104	-	.083	-	.077	-	.076	-	.119	-	.120
IRP4	Lower	0	-	.121	-	.102	-	.102	-	.083	-	.097	-	.080
		.25	-	.107	-	.098	-	.109	-	.097	-	.083	-	.080
		.5	-	.118	-	.094	-	.095	-	.087	-	.085	-	.087
		.75	-	.092	-	.083	-	.075	-	.070	-	.074	-	.076
	Higher	0	-	.066	-	.053	-	.054	-	.049	-	.043	-	.040
		.25	-	.054	-	.056	-	.049	-	.046	-	.043	-	.041
		.5	-	.063	-	.057	-	.045	-	.044	-	.044	-	.037
		.75	-	.057	-	.056	-	.043	-	.042	-	.039	-	.033

Note. “All” = average mean absolute difference (MAD) across all IRP types for the condition. “IRP0” = IRP for students with the proficiency status [0]. “IRP1” = IRP for students with the proficiency status [1]. “IRP2” = IRP for students with the proficiency status [2]. “IRP3” = IRP for students with the proficiency status [3]. “IRP4” = IRP for students with the proficiency status [4]. The hyphen highlight that only the conditions with five proficiency statuses have IRP3 and IRP4.

Table 14

Student Classification Accuracy Results for the One-Attribute Conditions

CCR Type	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
ACCR Pre	Lower	0	.673	.481	.720	.529	.687	.502	.731	.533	.676	.496	.705	.530
		.25	.681	.492	.717	.530	.690	.505	.728	.532	.684	.497	.719	.525
		.5	.661	.501	.725	.538	.691	.503	.743	.539	.677	.497	.728	.531
		.75	.698	.516	.734	.550	.700	.515	.744	.548	.687	.519	.738	.543
	Higher	0	.771	.542	**.856**	.603	**.800**	.562	**.861**	.596	.795	.550	**.858**	.585
		.25	.795	.548	**.860**	.608	**.802**	.566	**.861**	.598	.788	.558	**.858**	.581
		.5	.797	.555	**.862**	.611	**.807**	.568	**.861**	.603	**.804**	.562	**.865**	.595
		.75	**.816**	.569	**.871**	.622	**.819**	.578	**.873**	.614	**.819**	.583	**.872**	.603
ACCR Post	Lower	0	.676	.488	.725	.529	.609	.341	.680	.331	.581	.328	.620	.341
		.25	.679	.495	.714	.531	.615	.335	.671	.338	.591	.353	.635	.335
		.5	.664	.506	.728	.538	.616	.346	.691	.343	.568	.339	.661	.319
		.75	.694	.519	.737	.548	.655	.335	.725	.361	.638	.345	.704	.330
	Higher	0	.777	.539	**.852**	.610	.769	.414	**.836**	.405	.797	.473	**.865**	.448
		.25	.791	.553	**.860**	.605	.774	.420	**.838**	.404	.783	.465	**.865**	.445
		.5	.798	.555	**.859**	.614	.781	.418	**.838**	.427	**.813**	.459	**.872**	.469
		.75	**.814**	.573	**.870**	.622	**.802**	.420	**.855**	.429	**.820**	.462	**.877**	.493
CCCR	Lower	0	.455	.237	.523	.277	.417	.173	.497	.176	.394	.164	.437	.180
		.25	.478	.263	.520	.299	.429	.173	.493	.183	.404	.175	.460	.181
		.5	.480	.299	.551	.327	.442	.185	.526	.195	.388	.171	.485	.177
		.75	.545	.348	.589	.373	.492	.190	.568	.218	.442	.188	.520	.194
	Higher	0	.598	.295	.730	.368	.615	.231	.719	.245	.634	.262	.742	.265

.25	.634	.316	.741	.378	.622	.239	.722	.244	.617	.259	.742	.259
.5	.648	.343	.746	.403	.636	.243	.726	.262	.655	.252	.754	.277
.75	.691	.383	.771	.437	.672	.258	.757	.276	.673	.267	.767	.293

Note. "ACCR Pre" is the attribute-level correct classification rate for the pre-test only. "ACCR Post" is the attribute-level correct classification rate for the post-test only. For conditions with multiple attributes, the ACCRs are averaged across the attributes within the same testing occasion. "CCCR" is the class-level correct classification rate, which refers to students' entire profiles of proficiency statuses across both testing occasions. The bold font shows the conditions with high attribute-level classification accuracy (greater than .8).

Table 15

PTDCM Reliability Results for the One-Attribute Conditions

Reliability Metric	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
Polychoric	Lower	0	.585	.568	.716	.678	.536	.521	.626	.623	.573	.552	.666	.638
		.25	.625	.592	.721	.699	.566	.549	.652	.640	.602	.585	.700	.670
		.5	.674	.642	.766	.743	.593	.575	.681	.665	.640	.617	.726	.715
		.75	.730	.686	*.803*	*.775*	.615	.599	.704	.698	.666	.654	.760	.742
	Higher	0	*.774*	.713	*.855**	*.796*	.731	.683	*.820*	.763	.745	.695	*.826*	*.772*
		.25	*.809*	.742	*.875**	*.823*	.757	.701	*.837**	*.787*	*.780*	.731	*.849**	*.806*
		.5	*.830**	.767	**.899**	*.845**	*.785*	.732	*.859**	*.808*	*.810*	.762	*.878**	*.836**
		.75	*.869**	*.804*	**.919**	*.873**	*.807*	.751	*.883**	*.827*	*.850**	*.807*	**.906**	*.851**
Average Maximum Transition	Lower	0	.600	.515	.686	.563	.584	.468	.599	.505	.633	.497	.657	.512
		.25	.628	.531	.670	.559	.585	.480	.600	.505	.599	.490	.647	.511
		.5	.659	.556	.705	.601	.596	.472	.609	.516	.615	.482	.639	.515
		.75	.691	.575	.712	.619	.607	.493	.625	.529	.597	.483	.640	.531
	Higher	0	.681	.553	.759	.612	.650	.522	.733	.575	.680	.541	.757	.579
		.25	.690	.554	.763	.625	.652	.519	.738	.574	.678	.532	.760	.589
		.5	.690	.566	.779	.647	.666	.521	.746	.579	.685	.541	.767	.591
		.75	.726	.592	.793	.663	.695	.526	.768	.589	.698	.553	.779	.597
Point Biserial	Lower	0	.276	.205	.362	.263	.243	.185	.285	.219	.287	.199	.345	.243
		.25	.278	.200	.348	.253	.239	.192	.286	.234	.279	.196	.332	.252
		.5	.279	.198	.351	.251	.242	.177	.279	.224	.281	.195	.326	.237
		.75	.273	.191	.326	.241	.217	.175	.267	.225	.277	.193	.341	.233
	Higher	0	.421	.264	.545	.351	.380	.253	.511	.310	.409	.273	.537	.330

.25	.420	.260	.546	.340	.372	.249	.508	.309	.403	.266	.530	.337
.5	.404	.252	.551	.339	.372	.240	.498	.303	.401	.274	.517	.335
.75	.392	.242	.520	.328	.391	.237	.503	.303	.415	.269	.534	.335

Note. Based on the suggested reliability levels for what can be considered acceptable, good, very good, or excellent reliability from Schellman and Madison (in press) for the PB, PF, IG, polychoric, and average maximum transition metrics, I used italics to mark "Acceptable" reliabilities, italics and * to mark "Good" reliabilities, bold to mark "Very good" reliabilities, and bold font and * to mark "Excellent" reliabilities.

Table 16

Average Proportion of Maximum Posterior Probabilities for the One-Attribute Conditions

Greater Than	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
>.1	Lower	0	1	1	1	1	1	1	1	1	1	1	1	1
		.25	1	1	1	1	1	1	1	1	1	1	1	1
		.5	1	1	1	1	1	1	1	1	1	1	1	1
		.75	1	1	1	1	1	1	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1	1	1	1	1	1	1
		.25	1	1	1	1	1	1	1	1	1	1	1	1
		.5	1	1	1	1	1	1	1	1	1	1	1	1
		.75	1	1	1	1	1	1	1	1	1	1	1	1
>.2	Lower	0	.999	.992	.999	.994	1	.989	1	.992	1	.992	1	.996
		.25	.999	.989	.999	.994	1	.994	1	.995	1	.993	1	.996
		.5	1	.989	1	.994	1	.993	1	.995	1	.994	1	.998
		.75	1	.993	1	.996	1	.995	1	.997	1	.997	1	.999
	Higher	0	1	.995	1	.999	1	.996	1	.998	1	.998	1	.999
		.25	1	.994	1	.999	1	.997	1	.998	1	.998	1	.999
		.5	1	.995	1	.999	1	.997	1	.999	1	.999	1	1
		.75	1	.997	1	.999	1	.997	1	.999	1	.999	1	1
>.3	Lower	0	.963	.893	.984	.920	.977	.865	.981	.891	.980	.889	.984	.915
		.25	.972	.887	.979	.916	.971	.881	.979	.904	.976	.891	.985	.915
		.5	.976	.888	.984	.923	.978	.874	.984	.914	.983	.886	.986	.922
		.75	.979	.899	.986	.929	.985	.897	.990	.925	.988	.893	.993	.938
	Higher	0	.991	.918	.998	.965	.993	.924	.998	.953	.993	.941	.998	.967

		.25	.991	.915	.998	.964	.994	.921	.998	.952	.994	.942	.998	.970
		.5	.992	.919	.999	.967	.997	.921	.999	.955	.997	.948	.999	.971
		.75	.997	.928	1	.968	.999	.926	1	.964	.999	.956	1	.976
>.4	Lower	0	.839	.690	.919	.756	.863	.620	.877	.694	.890	.672	.914	.718
		.25	.867	.704	.899	.740	.852	.649	.878	.703	.868	.668	.915	.716
		.5	.877	.720	.912	.779	.874	.632	.891	.720	.891	.648	.908	.720
		.75	.893	.739	.911	.792	.890	.681	.907	.747	.893	.651	.927	.749
	Higher	0	.934	.747	.977	.848	.932	.737	.972	.819	.939	.769	.977	.840
		.25	.933	.743	.976	.848	.934	.728	.974	.816	.948	.767	.980	.845
		.5	.928	.748	.977	.860	.943	.735	.976	.819	.959	.775	.984	.849
		.75	.942	.771	.980	.866	.962	.748	.984	.834	.974	.796	.989	.859
>.5	Lower	0	.655	.476	.791	.563	.658	.379	.693	.469	.725	.441	.775	.481
		.25	.698	.511	.761	.551	.657	.409	.692	.469	.670	.426	.765	.478
		.5	.724	.548	.788	.620	.684	.387	.712	.492	.711	.406	.744	.485
		.75	.765	.577	.788	.639	.713	.447	.735	.527	.687	.411	.762	.522
	Higher	0	.803	.547	.912	.668	.786	.512	.895	.622	.814	.536	.913	.635
		.25	.812	.546	.910	.680	.793	.500	.901	.618	.830	.529	.921	.648
		.5	.798	.561	.916	.708	.811	.507	.906	.628	.852	.552	.933	.654
		.75	.826	.603	.916	.716	.849	.519	.923	.646	.885	.579	.947	.673
>.6	Lower	0	.475	.309	.637	.402	.442	.203	.483	.287	.535	.262	.588	.288
		.25	.530	.349	.603	.394	.445	.225	.483	.279	.462	.244	.569	.285
		.5	.577	.400	.652	.478	.471	.209	.502	.301	.503	.227	.543	.292
		.75	.634	.442	.661	.509	.503	.254	.534	.330	.454	.227	.541	.330
	Higher	0	.628	.382	.778	.493	.591	.317	.746	.429	.634	.349	.782	.436
		.25	.644	.383	.779	.517	.597	.307	.756	.424	.639	.320	.788	.451
		.5	.636	.406	.796	.557	.625	.312	.764	.437	.662	.349	.803	.456
		.75	.691	.459	.808	.571	.676	.325	.799	.455	.695	.375	.824	.475
>.7	Lower	0	.322	.182	.493	.268	.261	.091	.297	.152	.373	.137	.420	.150

		.25	.380	.222	.461	.270	.266	.103	.299	.142	.294	.121	.393	.149
		.5	.443	.278	.530	.358	.285	.091	.316	.161	.323	.111	.371	.154
		.75	.515	.322	.545	.394	.308	.117	.351	.177	.269	.108	.359	.184
	Higher	0	.463	.246	.640	.342	.406	.167	.592	.263	.466	.200	.637	.257
		.25	.485	.254	.645	.374	.409	.165	.603	.263	.456	.175	.645	.282
		.5	.491	.284	.671	.420	.443	.165	.619	.273	.472	.195	.664	.288
		.75	.563	.336	.693	.450	.511	.170	.663	.291	.499	.211	.690	.298
>.8	Lower	0	.185	.087	.345	.159	.113	.029	.136	.060	.227	.058	.265	.062
		.25	.235	.119	.320	.163	.126	.033	.140	.049	.161	.046	.237	.058
		.5	.312	.169	.403	.241	.133	.027	.150	.063	.174	.040	.228	.062
		.75	.387	.204	.423	.279	.142	.034	.184	.070	.135	.037	.211	.076
	Higher	0	.314	.139	.475	.209	.230	.059	.420	.128	.307	.092	.475	.124
		.25	.340	.146	.490	.244	.235	.059	.430	.129	.286	.066	.480	.147
		.5	.344	.174	.530	.295	.264	.065	.450	.142	.288	.080	.496	.150
		.75	.436	.222	.565	.334	.332	.067	.506	.156	.302	.091	.525	.148
>.9	Lower	0	.063	.021	.180	.063	.023	.003	.027	.011	.097	.015	.123	.015
		.25	.096	.036	.170	.065	.031	.004	.027	.007	.059	.009	.104	.012
		.5	.168	.064	.251	.114	.031	.002	.033	.012	.061	.007	.104	.014
		.75	.225	.078	.272	.148	.033	.005	.046	.010	.044	.006	.098	.013
	Higher	0	.162	.054	.278	.091	.063	.009	.204	.035	.143	.024	.269	.035
		.25	.184	.055	.297	.117	.061	.007	.217	.035	.126	.014	.270	.046
		.5	.197	.071	.348	.160	.081	.011	.237	.043	.120	.015	.274	.049
		.75	.280	.101	.396	.205	.126	.010	.294	.047	.125	.018	.290	.042

Table 17

Average Minimum DGPs for the One-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
Basic	Lower	0	.152	.012	.008	.021	.005	.008	.004	.002	.006	.007	.017	.006
		.25	.037	.009	.023	.004	.003	.003	.003	.005	.004	.004	.006	.003
		.5	.025	.005	.009	.011	.005	.003	.003	.005	.003	.004	.003	.003
		.75	.005	.006	.011	.005	.004	.003	.004	.003	.003	.007	.003	.002
	Higher	0	.211	.008	.458	.033	.005	.004	.013	.003	.004	.002	.021	.011
		.25	.088	.016	.277	.023	.005	.004	.006	.005	.005	.003	.009	.003
		.5	.059	.011	.109	.008	.004	.004	.004	.004	.004	.003	.004	.003
		.75	.006	.005	.010	.006	.004	.005	.004	.004	.004	.002	.003	.003
CW	Lower	0	.346	.318	.295	.289	.154	.106	.118	.093	.114	.087	.092	.057
		.25	.281	.195	.193	.194	.147	.123	.084	.069	.088	.077	.061	.039
		.5	.239	.232	.210	.158	.109	.081	.097	.049	.056	.064	.050	.030
		.75	.143	.194	.110	.070	.107	.072	.075	.034	.079	.081	.069	.045
	Higher	0	.338	.335	.472	.237	.109	.082	.078	.048	.066	.040	.046	.047
		.25	.227	.175	.291	.173	.082	.040	.047	.042	.054	.047	.032	.019
		.5	.184	.093	.165	.133	.064	.063	.030	.026	.029	.025	.011	.016
		.75	.077	.092	.076	.076	.040	.037	.022	.028	.007	.028	.008	.021
CWP	Lower	0	0	0	0	0	0	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0	0	0	0	0	0	0
		.5	0	0	0	0	0	0	0	0	0	0	0	0
		.75	0	0	0	0	0	0	0	0	0	0	0	0
	Higher	0	0	0	0	0	0	0	0	0	0	0	0	0

		.25	0	0	0	0	0	0	0	0	0	0	0	0
		.5	0	0	0	0	0	0	0	0	0	0	0	0
		.75	0	0	0	0	0	0	0	0	0	0	0	0
PW	Lower	0	0	0	0	0	.001	0	0	0	.002	.001	.001	0
		.25	.002	.001	0	0	.001	.004	.001	.001	.005	.001	.001	0
		.5	.005	.002	.001	0	.008	.009	.005	0	.013	.006	.001	.001
		.75	.014	.003	.002	0	.021	.010	.032	.005	.022	.025	.031	.006
	Higher	0	0	0	0	0	.001	0	0	0	0	0	0	0
		.25	0	0	0	0	.001	.001	0	0	0	.001	0	0
		.5	0	0	0	0	.001	.002	0	0	0	0	0	0
		.75	.001	.001	0	0	.007	.002	.001	.001	0	.001	.001	.001

Note. “CW” = adjusted DGP with complete weighting. “CWP” = adjusted DGP with complete weighting and a penalty for forgetting.

“PW” = adjusted DGP with partial weighting.

Table 18

Average Mean DGPs for the One-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
Basic	Lower	0	.783	.779	.780	.762	.824	.749	.786	.740	.789	.731	.764	.707
		.25	.808	.809	.784	.779	.810	.771	.790	.745	.768	.744	.759	.711
		.5	.841	.834	.826	.812	.823	.771	.799	.773	.785	.747	.767	.724
		.75	.893	.879	.867	.872	.863	.807	.837	.786	.806	.772	.779	.760
	Higher	0	.745	.741	.732	.724	.751	.710	.735	.702	.739	.701	.732	.672
		.25	.754	.749	.742	.731	.757	.720	.742	.708	.741	.701	.738	.681
		.5	.778	.779	.765	.763	.769	.740	.753	.722	.757	.733	.753	.698
		.75	.829	.829	.807	.816	.809	.772	.789	.754	.786	.756	.778	.723
CW	Lower	0	.783	.772	.782	.755	.786	.708	.760	.710	.757	.681	.742	.670
		.25	.800	.790	.782	.762	.770	.720	.759	.717	.741	.696	.737	.673
		.5	.818	.804	.812	.788	.787	.723	.768	.735	.760	.700	.745	.686
		.75	.858	.836	.839	.832	.810	.762	.793	.746	.785	.726	.764	.728
	Higher	0	.748	.735	.736	.720	.733	.678	.726	.682	.715	.657	.718	.638
		.25	.753	.736	.743	.723	.735	.688	.731	.687	.718	.652	.724	.646
		.5	.768	.753	.762	.750	.747	.706	.740	.695	.734	.687	.740	.663
		.75	.811	.791	.797	.794	.782	.737	.773	.726	.767	.711	.766	.686
CWP	Lower	0	.474	.419	.526	.444	.665	.554	.611	.571	.687	.618	.692	.611
		.25	.556	.475	.547	.492	.663	.595	.653	.598	.710	.645	.700	.625
		.5	.664	.538	.651	.556	.726	.639	.708	.654	.726	.673	.728	.664
		.75	.745	.622	.746	.652	.775	.721	.779	.701	.768	.706	.753	.704
	Higher	0	.482	.422	.497	.427	.612	.553	.616	.580	.674	.597	.678	.582

		.25	.552	.449	.551	.495	.644	.597	.650	.603	.694	.610	.707	.617
		.5	.613	.523	.616	.549	.693	.653	.692	.646	.729	.658	.734	.641
		.75	.714	.605	.707	.635	.766	.710	.758	.699	.767	.699	.766	.672
PW	Lower	0	.587	.544	.607	.542	.716	.616	.668	.622	.717	.644	.713	.634
		.25	.653	.598	.632	.587	.708	.648	.697	.645	.724	.667	.716	.644
		.5	.731	.649	.716	.645	.754	.675	.735	.687	.742	.686	.736	.673
		.75	.797	.717	.788	.727	.791	.740	.786	.720	.776	.716	.758	.714
	Higher	0	.558	.515	.553	.500	.654	.598	.648	.612	.689	.619	.691	.599
		.25	.618	.543	.605	.562	.679	.632	.677	.632	.704	.627	.713	.627
		.5	.672	.609	.664	.618	.715	.675	.710	.665	.731	.670	.737	.650
		.75	.757	.683	.744	.697	.774	.722	.765	.710	.767	.705	.766	.678

Note. “CW” = adjusted DGP with complete weighting. “CWP” = adjusted DGP with complete weighting and a penalty for forgetting.

“PW” = adjusted DGP with partial weighting.

Table 19

Average Median DGPs for the One-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
Basic	Lower	0	.762	.774	.735	.759	.951	.886	.876	.805	1	.866	.991	.821
		.25	.806	.819	.772	.786	.938	.912	.901	.830	1	.883	1	.848
		.5	.875	.877	.853	.831	.980	.939	.924	.876	.974	.902	1	.808
		.75	.945	.937	.911	.923	.998	.928	.984	.903	1	.881	1	.857
	Higher	0	.710	.735	.665	.697	.784	.758	.768	.728	1	.799	1	.725
		.25	.742	.767	.706	.708	.837	.821	.803	.763	1	.794	1	.738
		.5	.804	.812	.782	.782	.901	.827	.865	.783	1	.918	1	.746
		.75	.878	.898	.846	.860	.961	.898	.943	.845	1	.933	1	.807
CW	Lower	0	.772	.773	.755	.748	.844	.739	.811	.743	.844	.731	.846	.726
		.25	.803	.800	.772	.768	.820	.753	.816	.755	.816	.750	.836	.731
		.5	.847	.823	.838	.806	.842	.761	.822	.777	.842	.752	.832	.737
		.75	.899	.867	.884	.868	.859	.794	.853	.786	.863	.772	.855	.771
	Higher	0	.716	.724	.685	.692	.773	.706	.763	.697	.861	.720	.910	.690
		.25	.739	.735	.711	.708	.783	.715	.781	.715	.847	.712	.915	.700
		.5	.787	.773	.777	.768	.813	.742	.812	.727	.878	.754	.921	.717
		.75	.859	.829	.838	.836	.856	.781	.873	.773	.911	.775	.936	.729
CWP	Lower	0	.560	.467	.644	.519	.765	.613	.712	.646	.802	.688	.823	.686
		.25	.668	.552	.669	.592	.748	.656	.746	.670	.796	.716	.810	.691
		.5	.786	.638	.784	.686	.799	.696	.788	.723	.812	.731	.820	.718
		.75	.852	.731	.854	.786	.839	.766	.847	.756	.844	.754	.844	.752
	Higher	0	.576	.465	.588	.483	.715	.615	.731	.642	.840	.664	.895	.646

		.25	.667	.515	.666	.601	.748	.655	.766	.676	.832	.674	.909	.670
		.5	.734	.637	.749	.684	.793	.709	.801	.706	.874	.722	.919	.694
		.75	.831	.730	.824	.784	.849	.766	.870	.762	.911	.759	.936	.715
PW	Lower	0	.648	.596	.682	.610	.792	.653	.742	.673	.817	.703	.831	.699
		.25	.722	.661	.714	.665	.772	.689	.770	.696	.803	.728	.819	.703
		.5	.813	.720	.808	.740	.813	.716	.801	.740	.822	.738	.824	.724
		.75	.871	.791	.867	.820	.846	.775	.850	.766	.850	.761	.848	.757
	Higher	0	.624	.566	.614	.555	.730	.638	.738	.653	.848	.682	.902	.657
		.25	.696	.612	.681	.640	.760	.673	.772	.686	.836	.685	.912	.679
		.5	.757	.696	.759	.718	.801	.720	.806	.713	.876	.733	.920	.701
		.75	.845	.774	.832	.807	.852	.771	.871	.766	.911	.765	.936	.719

Note. “CW” = adjusted DGP with complete weighting. “CWP” = adjusted DGP with complete weighting and a penalty for forgetting.

“PW” = adjusted DGP with partial weighting.

Table 20

Average Maximum DGPs for the One-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth				Moderate Growth				Large Growth			
			I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5	I8 cat3	I8 cat5	I12 cat3	I12 cat5
Basic	Lower	0	1	1	1	1	1	1	1	1	1	1	1	1
		.25	1	1	1	1	1	1	1	1	1	1	1	1
		.5	1	1	1	1	1	1	1	1	1	1	1	1
		.75	1	1	1	1	1	1	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1	1	1	1	1	1	1
		.25	1	1	1	1	1	1	1	1	1	1	1	1
		.5	1	1	1	1	1	1	1	1	1	1	1	1
		.75	1	1	1	1	1	1	1	1	1	1	1	1
CW	Lower	0	1	.999	1	1	1	1	1	.999	1	1	1	1
		.25	1	.999	1	1	1	.999	.999	.999	1	1	1	1
		.5	1	1	1	1	1	.999	.999	1	1	1	1	1
		.75	1	1	1	1	1	1	1	.999	1	1	1	1
	Higher	0	1	1	1	1	.996	.999	.999	.999	.999	.999	1	1
		.25	.999	1	1	1	.995	.997	.999	1	1	.999	1	1
		.5	1	1	1	1	.995	1	.999	.999	1	1	1	1
		.75	.999	.999	1	1	.999	1	1	1	1	.999	1	1
CWP	Lower	0	.999	.998	1	.999	1	.997	1	.997	1	.999	1	1
		.25	.999	.998	1	1	1	.999	.999	.998	1	1	1	1
		.5	.999	.997	1	1	1	.996	.999	.999	1	1	1	.999
		.75	1	.998	1	1	1	.999	1	.999	1	1	1	1
	Higher	0	1	.999	1	1	.995	.999	.999	.998	.999	.999	1	1

		.25	.999	.999	1	1	.994	.997	.999	.998	1	.998	1	.999
		.5	1	.999	1	1	.995	1	.999	.999	1	1	1	1
		.75	.999	.998	1	1	.999	1	.999	1	1	.999	1	1
PW	Lower	0	.999	.999	1	.999	1	.997	1	.997	1	.999	1	1
		.25	1	.998	1	1	1	.999	.999	.999	1	1	1	1
		.5	.999	.997	1	1	1	.996	.999	.999	1	1	1	.999
		.75	1	.998	1	1	1	.999	1	.999	1	1	1	1
	Higher	0	1	.999	1	1	.995	.999	.999	.998	.999	.999	1	1
		.25	.999	.999	1	1	.994	.997	.999	.999	1	.998	1	.999
		.5	1	.999	1	1	.995	1	.999	.999	1	1	1	1
		.75	.999	.999	1	1	.999	1	1	1	1	.999	1	1

Note. “CW” = adjusted DGP with complete weighting. “CWP” = adjusted DGP with complete weighting and a penalty for forgetting.

“PW” = adjusted DGP with partial weighting.

Table 21

Item Parameter Estimation Accuracy Results: Average Mean Absolute Difference Between True and Estimated IRPs for the Three-Attribute Conditions

IRP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
All	Lower	0	.066	.061	.068	.061	.067	.060
		.25	.066	.061	.069	.059	.067	.061
		.5	.067	.058	.061	.057	.063	.059
		.75	.063	.052	.060	.055	.061	.055
	Higher	0	.058	.034	.058	.033	.060	.037
		.25	.055	.035	.054	.035	.053	.035
		.5	.047	.032	.052	.032	.052	.032
		.75	.040	.028	.042	.029	.042	.028
IRP0	Lower	0	.028	.021	.027	.020	.028	.020
		.25	.026	.019	.027	.020	.027	.020
		.5	.027	.020	.024	.020	.026	.019
		.75	.024	.019	.024	.021	.023	.019
	Higher	0	.027	.016	.025	.017	.027	.018
		.25	.023	.017	.023	.017	.022	.017
		.5	.020	.017	.023	.016	.022	.016
		.75	.019	.015	.019	.016	.021	.016
IRP1	Lower	0	.087	.086	.087	.084	.088	.082
		.25	.088	.081	.092	.082	.092	.081
		.5	.096	.079	.085	.079	.090	.081
		.75	.092	.073	.091	.078	.088	.077
	Higher	0	.095	.049	.094	.046	.096	.055
		.25	.084	.050	.086	.050	.084	.051
		.5	.072	.044	.082	.045	.080	.044
		.75	.060	.038	.064	.039	.065	.037
IRP2	Lower	0	.082	.076	.089	.079	.086	.077
		.25	.084	.082	.089	.074	.081	.082
		.5	.077	.076	.074	.073	.075	.078
		.75	.072	.064	.066	.067	.071	.068
	Higher	0	.052	.038	.054	.036	.057	.038
		.25	.059	.038	.053	.037	.053	.037
		.5	.050	.034	.051	.034	.053	.036
		.75	.042	.032	.043	.031	.041	.032

Note. “All” = average mean absolute difference (MAD) across all IRP types for the condition. “IRP0” = IRP for students with the proficiency status [0]. “IRP1” = IRP for students with the proficiency status [1]. “IRP2” = IRP for students with the proficiency status [2]. The MAD IRP values here were averaged across all attributes.

Table 22

Student Classification Accuracy Results for the Three-Attribute Conditions

CCR Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
ACCR Pre	Lower	0	.659	.713	.659	.714	.653	.712
		.25	.668	.714	.666	.715	.663	.720
		.5	.675	.727	.683	.728	.676	.728
		.75	.702	.749	.705	.745	.710	.751
	Higher	0	.756	.847	.760	.846	.754	.840
		.25	.774	.843	.769	.848	.766	.844
		.5	.793	.855	.781	.855	.775	.856
		.75	.815	.873	.812	.873	.808	.874
ACCR Post	Lower	0	.659	.710	.554	.589	.581	.610
		.25	.669	.713	.548	.599	.603	.599
		.5	.675	.726	.557	.610	.630	.611
		.75	.701	.749	.558	.627	.644	.664
	Higher	0	.753	.843	.681	.815	.740	.831
		.25	.769	.845	.705	.813	.759	.836
		.5	.789	.855	.723	.829	.762	.852
		.75	.813	.874	.773	.857	.803	.875
CCR Pre	Lower	0	.288	.365	.288	.365	.284	.360
		.25	.319	.386	.317	.384	.314	.392
		.5	.354	.429	.372	.428	.355	.431
		.75	.438	.499	.442	.486	.454	.499
	Higher	0	.432	.608	.442	.606	.429	.592
		.25	.475	.605	.466	.614	.458	.607
		.5	.522	.635	.502	.636	.485	.643
		.75	.586	.689	.579	.690	.572	.692
CCR Post	Lower	0	.288	.360	.171	.204	.192	.228
		.25	.320	.384	.171	.219	.227	.212
		.5	.354	.428	.181	.243	.277	.233
		.75	.436	.495	.218	.290	.319	.332
	Higher	0	.426	.601	.316	.538	.407	.574
		.25	.465	.607	.356	.538	.441	.586
		.5	.515	.637	.390	.581	.459	.628
		.75	.581	.689	.504	.652	.553	.689
CCCR	Lower	0	.084	.131	.050	.075	.056	.081
		.25	.117	.169	.056	.089	.066	.084
		.5	.169	.234	.068	.114	.083	.102
		.75	.275	.328	.103	.155	.122	.154

Higher	0	.188	.364	.142	.326	.175	.341
	.25	.234	.374	.168	.331	.195	.351
	.5	.302	.423	.201	.368	.212	.395
	.75	.392	.505	.296	.449	.291	.459

Note. "ACCR Pre" is the attribute-level correct classification rate for the pre-test only. "ACCR Post" is the attribute-level correct classification rate for the post-test only. For conditions with multiple attributes, the ACCRs are averaged across the attributes within the same testing occasion. "CCR Pre" is the class-level correct classification rate for the pre-test only. "CCR Post" is the class-level correct classification rate for the post-test only. "CCCR" is the overall, longitudinal class-level correct classification rate, which refers to students' entire profiles of proficiency statuses across both testing occasions.

Table 23

PTDCM Reliability Results for the Three-Attribute Conditions

Reliability Metric	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Polychoric	Lower	0	.611	.724	.570	.673	.581	.682
		.25	.630	.747	.597	.697	.611	.709
		.5	.688	*.786*	.631	.733	.648	.749
		.75	.760	*.836**	.693	*.774*	.719	*.797*
	Higher	0	*.793*	*.861**	.754	*.826*	.768	*.833**
		.25	*.817*	*.876**	*.774*	*.844**	*.788*	*.851**
		.5	*.853**	**.899**	*.802*	*.864**	*.809*	*.874**
		.75	*.886**	**.926**	*.836**	**.893**	*.850**	**.903**
Average Maximum Transition	Lower	0	.605	.686	.574	.636	.612	.663
		.25	.604	.690	.584	.638	.620	.672
		.5	.626	.701	.576	.645	.621	.681
		.75	.661	.709	.590	.653	.652	.694
	Higher	0	.701	.764	.678	.744	.717	.763
		.25	.706	.764	.682	.750	.710	.768
		.5	.727	.776	.693	.757	.713	.778
		.75	.737	.796	.709	.781	.732	.798
Point Biserial	Lower	0	.336	.412	.315	.383	.340	.407
		.25	.334	.420	.331	.394	.356	.425
		.5	.353	.425	.328	.406	.373	.444
		.75	.375	.431	.352	.410	.424	.470
	Higher	0	.471	.569	.450	.555	.488	.570
		.25	.476	.570	.456	.562	.485	.580
		.5	.486	.582	.477	.580	.494	.598
		.75	.494	.596	.502	.602	.519	.617

Note. Based on the suggested reliability levels for what can be considered acceptable, good, very good, or excellent reliability from Schellman and Madison (in press) for the PB, PF, IG, polychoric, and average maximum transition metrics, I used italics to mark "Acceptable" reliabilities, italics and * to mark "Good" reliabilities, bold to mark "Very good" reliabilities, and bold font and * to mark "Excellent" reliabilities.

Table 24

Average Proportion of Maximum Posterior Probabilities for the Three-Attribute Conditions

Greater Than	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
>.1	Lower	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
>.2	Lower	0	.998	.999	.997	.999	.998	.999
		.25	.997	.999	.997	.999	.998	.999
		.5	.998	.999	.998	.999	.998	.999
		.75	.998	.999	.997	.999	.999	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
>.3	Lower	0	.960	.979	.951	.973	.964	.979
		.25	.956	.981	.956	.975	.965	.982
		.5	.961	.981	.954	.976	.968	.985
		.75	.967	.980	.957	.981	.977	.988
	Higher	0	.989	.997	.988	.997	.992	.997
		.25	.990	.997	.991	.997	.991	.997
		.5	.992	.997	.992	.997	.992	.998
		.75	.993	.998	.994	.998	.996	.999
>.4	Lower	0	.851	.916	.822	.890	.863	.911
		.25	.845	.919	.839	.897	.870	.918
		.5	.861	.920	.830	.900	.874	.928
		.75	.879	.922	.843	.911	.903	.939
	Higher	0	.939	.976	.937	.974	.953	.977
		.25	.942	.975	.942	.975	.950	.978
		.5	.951	.978	.949	.976	.955	.982
		.75	.955	.982	.958	.983	.969	.988
>.5	Lower	0	.673	.792	.623	.740	.692	.775
		.25	.669	.797	.645	.746	.706	.789

		.5	.698	.803	.629	.754	.705	.807
		.75	.739	.809	.655	.770	.757	.826
	Higher	0	.826	.914	.817	.903	.861	.916
		.25	.832	.911	.826	.908	.848	.922
		.5	.852	.920	.841	.914	.859	.932
		.75	.863	.931	.864	.933	.892	.951
>.6	Lower	0	.488	.636	.420	.550	.503	.600
		.25	.487	.643	.446	.555	.522	.617
		.5	.525	.656	.425	.567	.520	.636
		.75	.586	.666	.454	.584	.580	.659
	Higher	0	.665	.783	.636	.761	.705	.785
		.25	.670	.781	.648	.770	.691	.794
		.5	.706	.797	.667	.781	.699	.809
		.75	.722	.821	.701	.813	.738	.841
>.7	Lower	0	.329	.493	.258	.383	.339	.440
		.25	.330	.499	.277	.386	.358	.458
		.5	.375	.521	.259	.400	.356	.476
		.75	.449	.537	.290	.414	.420	.499
	Higher	0	.512	.648	.465	.614	.550	.646
		.25	.521	.645	.475	.625	.536	.658
		.5	.565	.669	.498	.640	.542	.681
		.75	.584	.707	.536	.688	.581	.722
>.8	Lower	0	.189	.347	.131	.230	.192	.285
		.25	.191	.353	.142	.231	.207	.302
		.5	.238	.382	.127	.245	.208	.317
		.75	.317	.404	.152	.258	.264	.340
	Higher	0	.362	.490	.296	.445	.388	.489
		.25	.370	.493	.305	.462	.376	.501
		.5	.420	.520	.326	.479	.377	.528
		.75	.440	.572	.362	.540	.412	.577
>.9	Lower	0	.066	.187	.039	.095	.069	.133
		.25	.068	.193	.043	.092	.075	.145
		.5	.105	.224	.035	.103	.077	.154
		.75	.168	.249	.050	.113	.113	.174
	Higher	0	.201	.300	.132	.238	.208	.295
		.25	.211	.305	.131	.253	.195	.303
		.5	.257	.333	.148	.270	.196	.324
		.75	.276	.391	.168	.332	.220	.366

Table 25

Attribute 1 Average Minimum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.333	.331	.012	.020	.012	.041
		.25	.364	.390	.042	.010	.013	.019
		.5	.175	.198	.006	.007	.007	.011
		.75	.005	.121	.010	.010	.005	.004
	Higher	0	.201	.457	.002	.038	.004	.049
		.25	.341	.330	.002	.009	.021	.008
		.5	.200	.217	.003	.006	.004	.007
		.75	.065	.081	.006	.005	.004	.006
CW	Lower	0	.384	.384	.103	.099	.072	.078
		.25	.404	.402	.135	.085	.067	.076
		.5	.271	.262	.138	.102	.066	.049
		.75	.148	.153	.093	.047	.039	.021
	Higher	0	.287	.461	.121	.077	.048	.067
		.25	.380	.341	.055	.054	.045	.039
		.5	.230	.221	.062	.034	.027	.021
		.75	.080	.087	.020	.034	.019	.014
CWP	Lower	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
PW	Lower	0	0	0	.002	0	.006	0
		.25	.001	0	.005	.001	.003	0
		.5	.001	0	.005	.001	.005	.001
		.75	.001	0	.006	0	.009	.001
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	.001	0	.001	0

Table 26

Attribute 2 Average Minimum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.068	.384	.007	.032	.008	.028
		.25	.312	.404	.013	.012	.013	.017
		.5	.078	.261	.012	.008	.007	.011
		.75	.140	.105	.004	.006	.005	.006
	Higher	0	.286	.487	.007	.005	.009	.014
		.25	.381	.418	.005	.017	.004	.041
		.5	.175	.233	.006	.005	.004	.008
		.75	.081	.099	.005	.005	.005	.005
CW	Lower	0	.092	.407	.124	.088	.042	.067
		.25	.371	.431	.143	.104	.058	.069
		.5	.245	.291	.114	.082	.068	.026
		.75	.217	.128	.082	.049	.032	.030
	Higher	0	.358	.491	.099	.076	.069	.040
		.25	.393	.422	.077	.058	.062	.045
		.5	.219	.248	.054	.029	.017	.025
		.75	.133	.104	.038	.018	.019	.012
CWP	Lower	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
PW	Lower	0	0	0	.002	0	.002	0
		.25	0	0	.001	0	.002	.001
		.5	0	0	.002	0	.004	.001
		.75	.001	0	.002	.002	.001	.002
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	.001	0	0	0

Table 27

Attribute 3 Average Minimum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.268	.341	.020	.013	.002	.017
		.25	.119	.200	.004	.010	.003	.017
		.5	.225	.172	.005	.019	.009	.012
		.75	.055	.101	.005	.006	.007	.004
	Higher	0	.046	.473	.003	.041	.020	.058
		.25	.217	.346	.007	.025	.006	.018
		.5	.198	.205	.005	.006	.009	.006
		.75	.096	.060	.004	.005	.005	.005
CW	Lower	0	.331	.388	.143	.126	.076	.094
		.25	.257	.268	.130	.060	.067	.074
		.5	.278	.230	.087	.064	.041	.038
		.75	.171	.142	.079	.060	.038	.044
	Higher	0	.260	.482	.084	.072	.047	.071
		.25	.242	.352	.091	.056	.049	.031
		.5	.212	.221	.039	.030	.024	.023
		.75	.123	.082	.030	.020	.023	.013
CWP	Lower	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
PW	Lower	0	.001	0	.001	0	.001	0
		.25	0	0	.002	0	.003	.001
		.5	0	0	.005	.001	.002	.001
		.75	0	0	.004	.001	.010	.001
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	.001	0

Table 28

Profile Average Minimum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.404	.542	.077	.129	.079	.099
		.25	.424	.475	.082	.083	.055	.094
		.5	.257	.342	.052	.056	.027	.055
		.75	.189	.208	.053	.034	.021	.017
	Higher	0	.351	.528	.036	.092	.087	.096
		.25	.426	.431	.076	.068	.092	.090
		.5	.284	.269	.033	.041	.043	.048
		.75	.129	.126	.019	.033	.033	.029
CW	Lower	0	.471	.578	.198	.226	.156	.128
		.25	.517	.484	.196	.141	.115	.138
		.5	.390	.376	.163	.195	.090	.082
		.75	.284	.254	.114	.092	.042	.073
	Higher	0	.394	.533	.182	.112	.117	.106
		.25	.462	.455	.131	.109	.123	.109
		.5	.341	.304	.108	.110	.059	.080
		.75	.185	.186	.089	.099	.082	.065
CWP	Lower	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
	Higher	0	0	0	0	0	0	0
		.25	0	0	0	0	0	0
		.5	0	0	0	0	0	0
		.75	0	0	0	0	0	0
PW	Lower	0	.022	.002	.018	.017	.076	.037
		.25	.009	.001	.037	.011	.055	.044
		.5	.004	.001	.028	.012	.041	.025
		.75	.005	.001	.020	.013	.012	.027
	Higher	0	.002	0	.010	.003	.012	.003
		.25	.001	0	.019	.002	.021	.007
		.5	0	0	.016	.003	.010	.023
		.75	.001	0	.021	.009	.040	.041

Table 29

Attribute 1 Average Mean DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.753	.769	.722	.739	.764	.739
		.25	.764	.767	.729	.728	.763	.744
		.5	.766	.778	.716	.725	.763	.739
		.75	.787	.797	.718	.738	.753	.756
	Higher	0	.734	.731	.731	.737	.739	.726
		.25	.740	.739	.734	.738	.740	.726
		.5	.757	.750	.744	.742	.750	.742
		.75	.780	.780	.754	.764	.760	.758
CW	Lower	0	.766	.776	.700	.722	.726	.715
		.25	.770	.771	.704	.715	.720	.723
		.5	.769	.778	.694	.708	.719	.717
		.75	.778	.788	.696	.720	.715	.732
	Higher	0	.738	.736	.716	.729	.721	.714
		.25	.744	.741	.720	.729	.715	.713
		.5	.756	.750	.728	.731	.729	.727
		.75	.774	.774	.736	.750	.739	.742
CWP	Lower	0	.473	.508	.592	.616	.671	.660
		.25	.499	.531	.614	.621	.683	.672
		.5	.533	.571	.611	.632	.691	.677
		.75	.591	.630	.629	.657	.689	.707
	Higher	0	.490	.503	.612	.621	.680	.675
		.25	.518	.540	.630	.641	.686	.688
		.5	.575	.581	.656	.667	.711	.716
		.75	.630	.649	.693	.719	.733	.739
PW	Lower	0	.584	.593	.638	.658	.695	.683
		.25	.605	.615	.653	.660	.700	.693
		.5	.631	.650	.649	.665	.704	.694
		.75	.672	.696	.660	.685	.701	.718
	Higher	0	.562	.558	.649	.652	.696	.688
		.25	.589	.594	.663	.669	.698	.697
		.5	.638	.634	.684	.689	.719	.721
		.75	.687	.695	.712	.731	.736	.740

Table 30

Attribute 2 Average Mean DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.767	.764	.746	.729	.768	.745
		.25	.747	.772	.728	.727	.761	.744
		.5	.771	.777	.730	.731	.752	.755
		.75	.794	.793	.720	.737	.769	.749
	Higher	0	.742	.732	.732	.739	.740	.729
		.25	.746	.734	.735	.739	.746	.735
		.5	.760	.749	.738	.743	.743	.735
		.75	.775	.779	.755	.762	.758	.752
CW	Lower	0	.776	.772	.719	.716	.731	.720
		.25	.757	.776	.700	.713	.725	.717
		.5	.770	.777	.704	.713	.709	.731
		.75	.787	.787	.695	.720	.728	.724
	Higher	0	.747	.737	.716	.730	.721	.715
		.25	.749	.737	.721	.729	.724	.722
		.5	.759	.749	.722	.731	.720	.719
		.75	.770	.774	.736	.749	.734	.735
CWP	Lower	0	.489	.505	.616	.608	.679	.667
		.25	.477	.543	.607	.619	.680	.677
		.5	.540	.571	.621	.636	.674	.696
		.75	.596	.624	.631	.666	.708	.700
	Higher	0	.498	.504	.614	.624	.677	.678
		.25	.530	.529	.626	.641	.696	.699
		.5	.574	.584	.651	.669	.702	.707
		.75	.623	.646	.691	.714	.728	.731
PW	Lower	0	.594	.592	.660	.651	.702	.689
		.25	.587	.624	.647	.658	.699	.695
		.5	.634	.649	.658	.669	.690	.711
		.75	.679	.692	.660	.690	.717	.711
	Higher	0	.568	.559	.649	.655	.693	.691
		.25	.598	.586	.661	.668	.707	.708
		.5	.638	.636	.680	.690	.710	.712
		.75	.683	.692	.711	.728	.731	.733

Table 31

Attribute 3 Average Mean DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.752	.773	.732	.736	.746	.752
		.25	.762	.765	.750	.747	.757	.753
		.5	.762	.777	.717	.742	.773	.738
		.75	.795	.792	.718	.737	.756	.752
	Higher	0	.746	.733	.725	.733	.745	.726
		.25	.739	.735	.743	.734	.737	.727
		.5	.760	.746	.741	.743	.746	.737
		.75	.783	.780	.762	.763	.753	.757
CW	Lower	0	.762	.780	.709	.723	.709	.725
		.25	.769	.769	.726	.730	.717	.730
		.5	.767	.777	.698	.725	.738	.716
		.75	.787	.784	.695	.718	.722	.723
	Higher	0	.750	.737	.712	.724	.726	.713
		.25	.742	.737	.726	.724	.715	.714
		.5	.760	.746	.724	.732	.723	.722
		.75	.777	.776	.743	.750	.730	.741
CWP	Lower	0	.475	.511	.602	.616	.655	.675
		.25	.504	.530	.630	.637	.675	.683
		.5	.523	.573	.608	.644	.709	.680
		.75	.600	.618	.629	.664	.694	.701
	Higher	0	.508	.504	.603	.620	.687	.675
		.25	.519	.530	.629	.638	.685	.689
		.5	.579	.578	.653	.670	.705	.710
		.75	.633	.650	.696	.717	.723	.738
PW	Lower	0	.582	.597	.647	.658	.679	.696
		.25	.609	.614	.672	.674	.694	.703
		.5	.623	.650	.648	.677	.723	.696
		.75	.680	.688	.659	.688	.706	.711
	Higher	0	.575	.559	.641	.650	.702	.688
		.25	.589	.586	.665	.665	.697	.699
		.5	.643	.631	.681	.692	.713	.715
		.75	.691	.696	.716	.730	.726	.739

Table 32

Profile Average Mean DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.757	.769	.733	.735	.759	.745
		.25	.757	.768	.736	.734	.760	.747
		.5	.766	.777	.721	.733	.763	.744
		.75	.792	.794	.719	.738	.759	.752
	Higher	0	.740	.732	.729	.736	.742	.727
		.25	.742	.736	.738	.737	.741	.730
		.5	.759	.748	.741	.743	.746	.738
		.75	.779	.780	.757	.763	.757	.756
CW	Lower	0	.768	.776	.709	.720	.722	.720
		.25	.766	.772	.710	.719	.721	.723
		.5	.769	.777	.699	.715	.722	.721
		.75	.784	.787	.695	.719	.722	.727
	Higher	0	.745	.736	.715	.728	.723	.714
		.25	.745	.738	.722	.728	.718	.716
		.5	.758	.748	.725	.731	.724	.723
		.75	.773	.775	.738	.749	.734	.739
CWP	Lower	0	.479	.508	.603	.613	.668	.667
		.25	.493	.535	.617	.625	.679	.678
		.5	.532	.572	.613	.637	.692	.684
		.75	.596	.624	.630	.662	.697	.703
	Higher	0	.499	.504	.610	.621	.681	.676
		.25	.522	.533	.629	.640	.689	.692
		.5	.576	.581	.653	.669	.706	.711
		.75	.629	.648	.693	.717	.728	.736
PW	Lower	0	.587	.594	.648	.656	.692	.689
		.25	.600	.618	.657	.664	.697	.697
		.5	.629	.650	.652	.670	.706	.700
		.75	.677	.692	.660	.687	.708	.713
	Higher	0	.568	.559	.646	.652	.697	.689
		.25	.592	.589	.663	.667	.701	.701
		.5	.640	.634	.681	.691	.714	.716
		.75	.687	.694	.713	.730	.731	.737

Table 33

Attribute 1 Average Median DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.747	.728	.809	.868	.904	.931
		.25	.746	.747	.858	.823	1	.855
		.5	.777	.774	.803	.826	1	.843
		.75	.826	.834	.802	.856	1	1
	Higher	0	.700	.647	.807	.780	1	1
		.25	.715	.704	.793	.801	1	1
		.5	.760	.742	.833	.832	1	1
		.75	.810	.804	.875	.884	1	1
CW	Lower	0	.768	.748	.728	.785	.779	.789
		.25	.762	.754	.735	.768	.791	.807
		.5	.774	.777	.721	.759	.800	.802
		.75	.803	.815	.727	.777	.796	.831
	Higher	0	.707	.679	.768	.774	.859	.886
		.25	.719	.708	.763	.785	.857	.869
		.5	.760	.743	.786	.797	.871	.910
		.75	.801	.801	.800	.838	.892	.915
CWP	Lower	0	.548	.630	.650	.705	.748	.753
		.25	.586	.651	.665	.696	.766	.779
		.5	.638	.703	.654	.698	.781	.774
		.75	.707	.761	.674	.722	.770	.812
	Higher	0	.581	.602	.710	.744	.837	.867
		.25	.626	.657	.708	.761	.840	.855
		.5	.704	.708	.748	.784	.858	.903
		.75	.760	.782	.783	.832	.887	.912
PW	Lower	0	.638	.677	.680	.728	.758	.767
		.25	.665	.697	.690	.721	.777	.788
		.5	.703	.738	.680	.721	.791	.783
		.75	.754	.786	.695	.740	.782	.819
	Higher	0	.635	.622	.726	.750	.846	.876
		.25	.668	.675	.723	.767	.849	.861
		.5	.727	.723	.760	.789	.864	.906
		.75	.779	.791	.789	.835	.890	.913

Table 34

Attribute 2 Average Median DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.739	.743	.841	.883	1	.939
		.25	.717	.737	.859	.818	.920	1
		.5	.773	.786	.793	.845	1	.905
		.75	.827	.830	.801	.828	1	1
	Higher	0	.690	.671	.785	.783	1	1
		.25	.718	.689	.817	.800	1	1
		.5	.768	.742	.810	.837	1	1
		.75	.806	.802	.865	.896	1	1
CW	Lower	0	.763	.755	.765	.776	.802	.797
		.25	.743	.755	.736	.760	.800	.801
		.5	.772	.782	.735	.765	.778	.819
		.75	.807	.816	.724	.769	.824	.801
	Higher	0	.715	.688	.768	.770	.853	.893
		.25	.723	.699	.786	.785	.876	.910
		.5	.767	.745	.781	.794	.844	.890
		.75	.799	.799	.805	.843	.878	.896
CWP	Lower	0	.575	.611	.690	.692	.767	.766
		.25	.549	.667	.657	.691	.774	.776
		.5	.644	.701	.673	.705	.749	.798
		.75	.715	.759	.669	.727	.810	.785
	Higher	0	.598	.601	.710	.735	.829	.876
		.25	.646	.641	.732	.761	.859	.899
		.5	.707	.712	.748	.782	.831	.881
		.75	.754	.778	.787	.836	.872	.892
PW	Lower	0	.655	.666	.716	.716	.780	.777
		.25	.638	.703	.686	.713	.784	.785
		.5	.706	.735	.696	.726	.762	.805
		.75	.762	.788	.690	.742	.817	.791
	Higher	0	.639	.622	.727	.743	.839	.883
		.25	.677	.661	.747	.767	.867	.904
		.5	.735	.727	.761	.787	.836	.886
		.75	.776	.787	.795	.839	.875	.894

Table 35

Attribute 3 Average Median DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.743	.738	.819	.857	1	.961
		.25	.739	.744	.866	.885	1	.896
		.5	.779	.787	.755	.868	1	.846
		.75	.822	.825	.788	.864	1	1
	Higher	0	.728	.656	.751	.776	1	1
		.25	.725	.688	.815	.783	1	1
		.5	.757	.736	.836	.827	1	1
		.75	.815	.807	.884	.896	1	1
CW	Lower	0	.758	.755	.744	.782	.762	.806
		.25	.758	.752	.776	.797	.794	.816
		.5	.770	.783	.721	.789	.831	.788
		.75	.809	.809	.726	.772	.809	.812
	Higher	0	.728	.680	.744	.768	.876	.886
		.25	.725	.696	.780	.767	.849	.884
		.5	.759	.738	.785	.794	.862	.888
		.75	.805	.803	.813	.845	.864	.918
CWP	Lower	0	.549	.629	.668	.705	.726	.782
		.25	.595	.650	.706	.724	.766	.784
		.5	.619	.705	.657	.725	.812	.765
		.75	.724	.749	.672	.730	.785	.795
	Higher	0	.620	.600	.683	.735	.857	.870
		.25	.626	.643	.725	.746	.831	.870
		.5	.706	.704	.744	.779	.851	.880
		.75	.766	.783	.795	.836	.859	.914
PW	Lower	0	.639	.677	.695	.730	.742	.791
		.25	.668	.693	.731	.748	.779	.792
		.5	.691	.741	.681	.746	.820	.774
		.75	.765	.777	.694	.746	.794	.802
	Higher	0	.658	.618	.699	.742	.864	.877
		.25	.667	.664	.743	.752	.840	.876
		.5	.728	.718	.759	.785	.857	.883
		.75	.782	.793	.802	.840	.861	.916

Table 36

Profile Average Median DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	.759	.768	.744	.742	.772	.753
		.25	.750	.757	.763	.754	.793	.779
		.5	.768	.777	.756	.758	.826	.803
		.75	.813	.821	.763	.788	.841	.838
	Higher	0	.740	.727	.740	.777	.745	.729
		.25	.731	.727	.775	.783	.771	.766
		.5	.755	.742	.786	.797	.797	.773
		.75	.798	.801	.833	.834	.847	.850
CW	Lower	0	.769	.774	.717	.729	.733	.732
		.25	.761	.763	.725	.735	.747	.754
		.5	.767	.776	.716	.739	.769	.771
		.75	.797	.805	.720	.754	.787	.799
	Higher	0	.743	.733	.721	.748	.730	.723
		.25	.734	.727	.745	.757	.741	.741
		.5	.756	.743	.753	.767	.770	.763
		.75	.790	.794	.784	.801	.812	.814
CWP	Lower	0	.486	.517	.616	.625	.683	.681
		.25	.506	.555	.632	.638	.704	.704
		.5	.564	.616	.628	.652	.735	.727
		.75	.657	.700	.649	.692	.760	.770
	Higher	0	.506	.511	.620	.629	.692	.686
		.25	.540	.550	.641	.652	.712	.715
		.5	.617	.621	.673	.696	.749	.750
		.75	.700	.727	.748	.779	.804	.810
PW	Lower	0	.596	.605	.660	.667	.704	.703
		.25	.617	.643	.672	.678	.721	.724
		.5	.660	.693	.664	.687	.749	.743
		.75	.727	.753	.680	.717	.771	.780
	Higher	0	.574	.565	.656	.664	.707	.701
		.25	.614	.613	.679	.683	.724	.725
		.5	.681	.673	.703	.723	.757	.755
		.75	.746	.762	.763	.788	.807	.811

Table 37

Attribute 1 Average Maximum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CW	Lower	0	1	1	.999	1	1	1
		.25	1	1	.999	1	1	1
		.5	1	1	.999	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CWP	Lower	0	1	1	.999	1	1	1
		.25	1	1	.999	1	1	1
		.5	1	1	.999	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
PW	Lower	0	1	1	.999	1	1	1
		.25	1	1	.999	1	1	1
		.5	1	1	.999	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1

Table 38

Attribute 2 Average Maximum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CW	Lower	0	1	1	.999	1	1	1
		.25	.999	1	.998	1	1	1
		.5	1	1	.999	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CWP	Lower	0	1	1	.999	1	1	1
		.25	.999	1	.998	1	1	1
		.5	.999	1	.999	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
PW	Lower	0	1	1	.999	1	1	1
		.25	.999	1	.998	1	1	1
		.5	.999	1	.999	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1

Table 39

Attribute 3 Average Maximum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CW	Lower	0	1	1	.999	1	1	1
		.25	1	1	.999	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CWP	Lower	0	1	1	.999	1	1	1
		.25	1	1	.999	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
PW	Lower	0	1	1	.999	1	1	1
		.25	1	1	.999	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1

Table 40

Profile Average Maximum DGPs for the Three-Attribute Conditions

DGP Type	Item Quality	Attribute Correlation	No Growth		Moderate Growth		Large Growth	
			I8 cat3	I12 cat3	I8 cat3	I12 cat3	I8 cat3	I12 cat3
Basic	Lower	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
	Higher	0	1	1	1	1	1	1
		.25	1	1	1	1	1	1
		.5	1	1	1	1	1	1
		.75	1	1	1	1	1	1
CW	Lower	0	.995	.999	.990	.996	.996	.998
		.25	.997	.999	.995	.997	.999	.997
		.5	.996	.999	.997	.998	.999	.999
		.75	.998	.999	.999	.999	.998	1
	Higher	0	.998	1	.997	.999	.999	1
		.25	.999	.999	.999	.999	.999	1
		.5	.998	1	.998	1	.999	1
		.75	.999	1	.999	1	.999	1
CWP	Lower	0	.989	.999	.989	.993	.996	.998
		.25	.995	.996	.995	.996	.999	.997
		.5	.994	.999	.995	.995	.998	.999
		.75	.996	.999	.998	.999	.998	1
	Higher	0	.998	.999	.996	.998	.999	1
		.25	.997	.999	.999	.999	.999	1
		.5	.997	1	.998	1	.999	1
		.75	.999	.999	.999	1	.999	1
PW	Lower	0	.991	.999	.989	.994	.996	.998
		.25	.995	.997	.995	.996	.999	.997
		.5	.995	.999	.995	.996	.999	.999
		.75	.997	.999	.998	.999	.998	1
	Higher	0	.998	.999	.996	.998	.999	1
		.25	.998	.999	.999	.999	.999	1
		.5	.997	1	.998	1	.999	1
		.75	.999	1	.999	1	.999	1

Table 41

Summary of the Relative Impact of Manipulated Factors on the Key Evaluation Metrics

Factor	Level Change	Average change in the average values for…				
		IRP	CCR	Reliability	Mean DGPs	Median DGPs
One-Attribute Conditions						
Proficiency Statuses	3→5	+.040	-.283	-.097	-.052	-.089
Item Quality	Lower→Higher	-.011	+.121	+.115	-.028	-.021
Items Per Attribute	8→12	-.005	+.045	+.070	-.008	-.007
Growth	No→Moderate	-.005	-.054	-.048	+.021	+.042
	Moderate→Large	+.012	-.001	+.027	-.005	+.035
Attribute Correlation	0→.25	-.001	+.005	+.009	+.018	+.025
	.25→.5	-.001	+.008	+.012	+.033	+.042
	.5→.75	-.005	+.020	+.014	+.045	+.047
Three-Attribute Conditions*						
Item Quality	Lower→Higher	-.019	+.180	+.131	+.007	+.037
Items Per Attribute	8→12	-.013	+.084	+.074	+.006	+.016
Growth	No→Moderate	0	-.058	-.032	+.015	+.044
	Moderate→Large	+.001	+.018	+.025	+.027	+.102
Attribute Correlation	0→.25	-.001	+.014	+.012	+.009	+.011
	.25→.5	-.003	+.027	+.018	+.014	+.021
	.5→.75	-.005	+.052	+.028	+.023	+.034

Note. *The three-attribute condition results are for Attribute 1 only, without loss of generality.

Table 42

Q-Matrix for the Empirical Data Analysis with Four Two-Category Attributes

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	1	0
6	0	0	1	0
7	0	0	1	0
8	0	0	1	0
9	0	1	0	0
10	0	1	0	0
11	0	1	0	0
12	0	1	0	0
13	1	0	0	0
14	1	0	0	0
15	0	0	0	1
16	0	0	0	1
17	1	0	0	0
18	0	0	0	1
19	0	0	0	1
20	0	0	0	1
21	0	0	0	1

Table 43

Attribute Correlations for the Empirical Data Analysis with Four Two-Category Attributes

	Att1 Pre	Att2 Pre	Att3 Pre	Att4 Pre	Att1 Post	Att2 Post	Att3 Post	Att4 Post
Att1 Pre	1	.496	.648	.501	.339	.315	.313	.264
Att2 Pre		1	.461	.382	.407	.439	.377	.275
Att3 Pre			1	.448	.261	.266	.258	.220
Att4 Pre				1	.224	.300	.274	.281
Att1 Post					1	.660	.660	.680
Att2 Post						1	.615	.589
Att3 Post							1	.481
Att4 Post								1

Table 44

Conditional Transition Matrices for the Empirical Data Analysis with Four Two-Category Attributes

Attribute	Pre-Test Proficiency Status	Post-Test Proficiency Status	
		Non-Proficient [0]	Proficient [1]
1	Non-Proficient [0]	.522	.478
	Proficient [1]	.178	.822
2	Non-Proficient [0]	.573	.427
	Proficient [1]	.154	.846
3	Non-Proficient [0]	.585	.415
	Proficient [1]	.314	.685
4	Non-Proficient [0]	.427	.573
	Proficient [1]	.172	.828

Note. The proficiency statuses show the corresponding labels and indices.

Table 45

PTDCM Reliability for the Empirical Data Analysis with Four Two-Category Attributes

Reliability Metric	Attribute 1	Attribute 2	Attribute 3	Attribute 4
PB	.759**	.865***	.790**	.780**
PBW	.779	.886	.803	.790
PF	.808**	.875**	.817**	.808**
PFW	.826	.897	.831	.817
IG	.534**	.563***	.557**	.542**
IGW	.585	.627	.592	.587
Polychoric	.918***	.968****	.927***	.911***
AvgMax	.890**	.947***	.901***	.898**

Note. PB = point biserial reliability metric; PBW weighted point biserial reliability metric; PF = parallel forms reliability metric; PFW = weighted parallel forms reliability metric; IG = information gain reliability metric; IGW = weighted information gain reliability metric; Polychoric = polychoric reliability metric; AvgMax = average maximum transition reliability metric. Based on the suggested reliability levels for what can be considered acceptable, good, very good, or excellent reliability from Schellman and Madison (in press) for the PB, PF, IG, polychoric, and average maximum transition metrics, I used * to mark “Acceptable” reliabilities, ** to mark “Good” reliabilities, *** to mark “Very good” reliabilities, and **** to mark “Excellent” reliabilities.

Table 46

Analysis of Average Maximum Posterior Probability for the Empirical Data Analysis with Four Two-Category Attributes

Proportion Maximum Posterior Probabilities Greater Than	Attribute 1	Attribute 2	Attribute 3	Attribute 4
.1	1	1	1	1
.2	1	1	1	1
.3	1	1	1	1
.4	.991	1	.997	.999
.5	.976	.992	.980	.990
.6	.929	.973	.933	.935
.7	.869	.941	.878	.870
.8	.777	.898	.786	.787
.9	.661	.821	.694	.686

Table 47

Item Response Probabilities for the 25 Items in the Empirical Data Analysis with One Three-Category Attribute

Item	IRP for Beginning Students	IRP for Proficient Students	IRP for Advanced Students
1	.225	.386	.590
2	.288	.635	.903
3	.322	.499	.664
4	.281	.742	.961
5	.300	.499	.619
6	.220	.220	.293
7	.338	.667	.913
8	.198	.198	.520
9	.183	.183	.345
10	.250	.250	.286
11	.292	.529	.670
12	.455	.865	.971
13	.266	.405	.805
14	.325	.571	.818
15	.367	.415	.810
16	.312	.418	.607
17	.306	.742	.944
18	.184	.445	.646
19	.279	.498	.897
20	.318	.524	.928
21	.220	.454	.802
22	.293	.536	.868
23	.236	.469	.890
24	.231	.688	.970
25	.254	.415	.608

Table 48

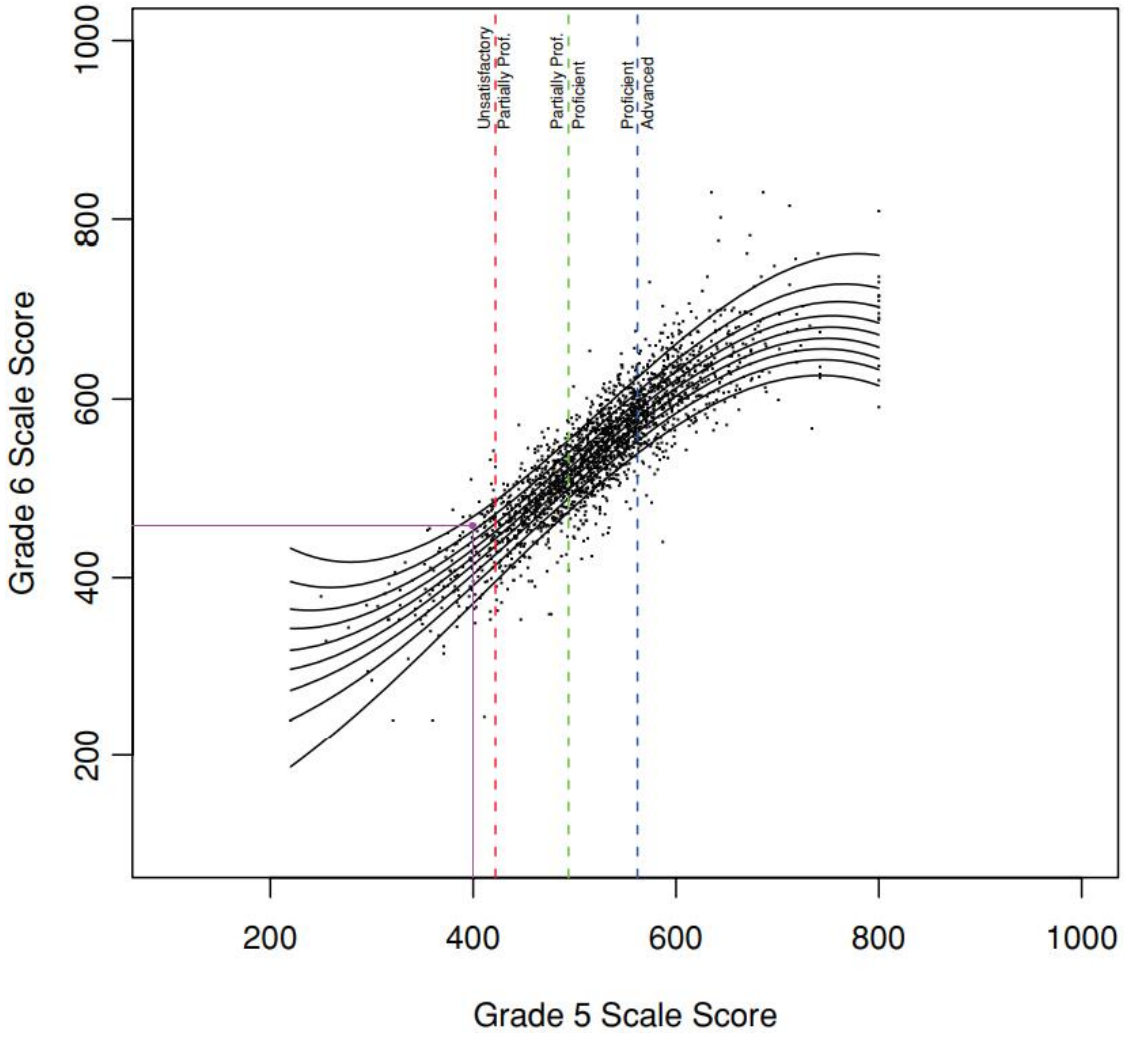
Conditional Transition Matrix for the Empirical Data Analysis with One Three-Category Attribute

Pre-Test Proficiency Status	Post-Test Proficiency Status		
	Beginning [0]	Proficient [1]	Advanced [2]
Beginning [0]	.707	.289	.004
Proficient [1]	.009	.402	.589
Advanced [2]	.013	.007	.980

Note. The proficiency statuses show the corresponding labels and indices.

Figure 1

B-Spline Conditional Deciles for Grade 5 and 6 Scaled Scores



Note. Copied from Betebenner's Figure 3 (2009, p. 47). I added the purple point and lines.

Figure 2

True Attribute Correlation Distributions for the Simulation Study

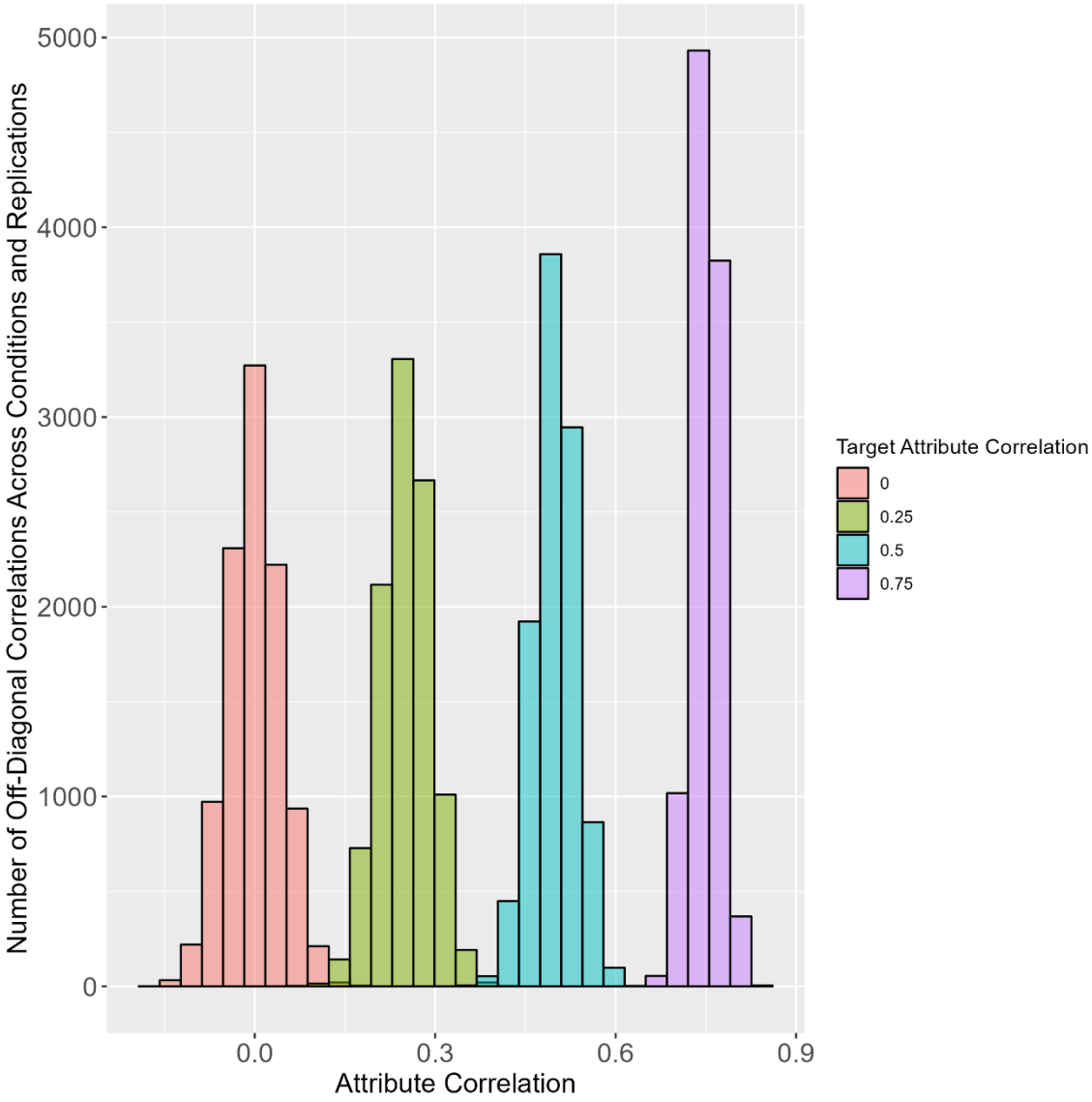


Figure 3

Convergence Rates for the Simulation Study

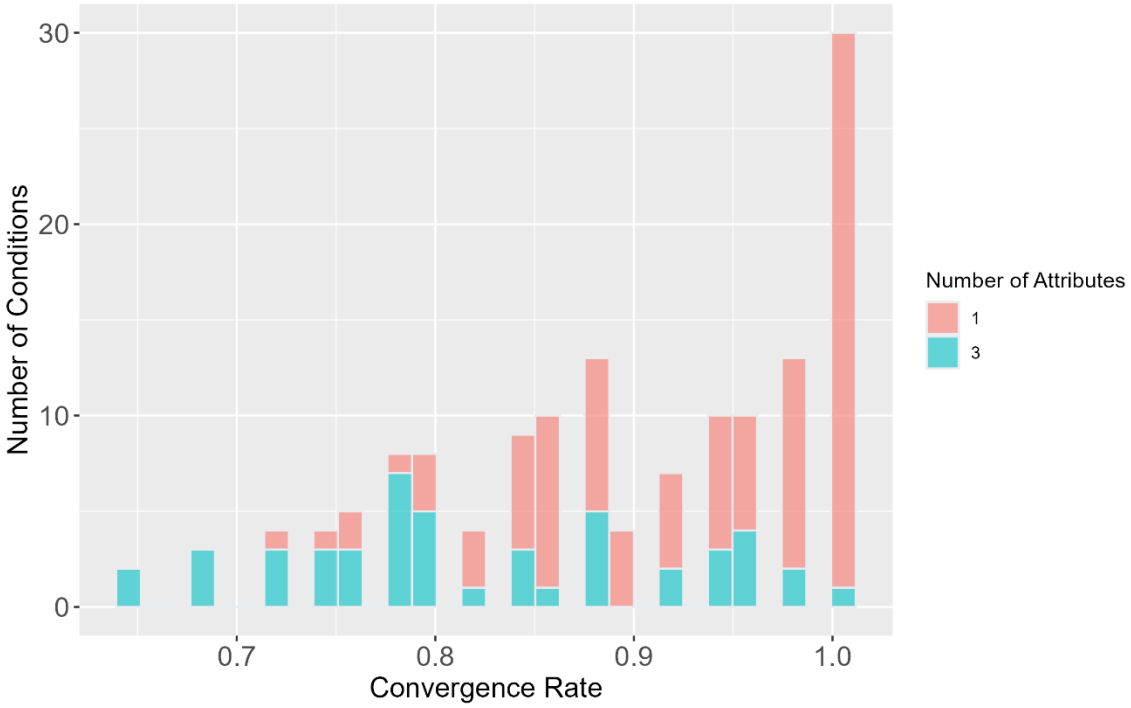
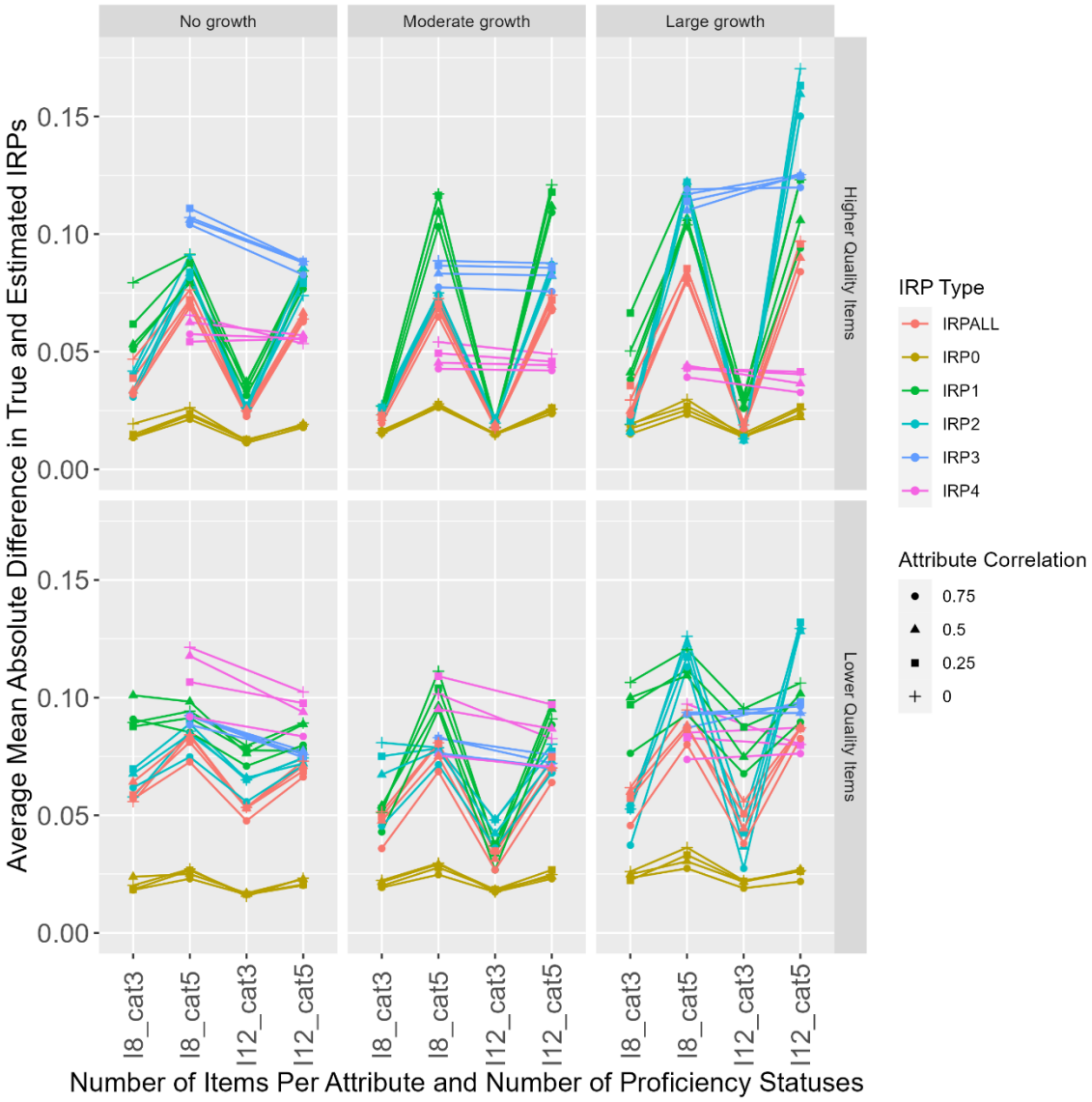


Figure 4

Item Parameter Estimation Accuracy Results: Mean Absolute Difference Between True and Estimated IRPs for the One-Attribute Conditions

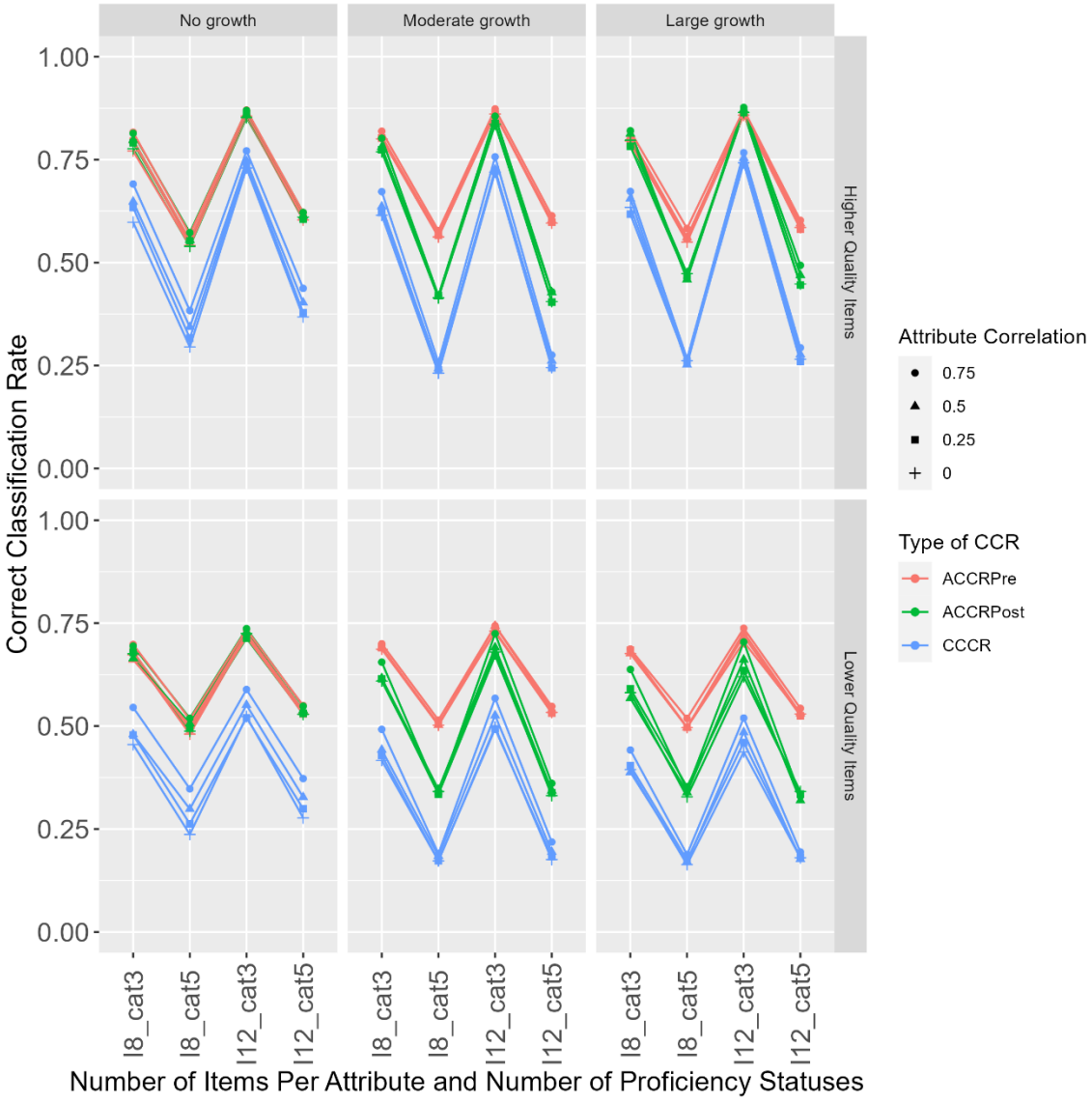


Note. "IRPALL" = average mean absolute difference (MAD) across all IRP types for the condition. "IRP0" = IRP for students with the proficiency status [0]. "IRP1" = IRP for students with the proficiency status [1]. "IRP2" = IRP for students with the proficiency status [2]. "IRP3" = IRP for students with the proficiency status [3]. "IRP4" = IRP for

students with the proficiency status [4]. “I8_cat3” = conditions with eight items and three proficiency statuses per attribute. “I8_cat5” = conditions with eight items and five proficiency statuses per attribute. “I12_cat3” = conditions with 12 items and three proficiency statuses per attribute. “I12_cat5” = conditions with 12 items and five proficiency statuses per attribute. Only the conditions with five proficiency statuses have IRP3 and IRP4.

Figure 5

Correct Classification Rates for the One-Attribute Conditions



Note. "ACCRPre" is the attribute-level correct classification rate for the pre-test only. "ACCRPost" is the attribute-level correct classification rate for the post-test only. For conditions with multiple attributes, the ACCRs are averaged across the attributes within the same testing occasion. "CCCR" is the class-level correct classification rate, which refers to students' entire profiles of proficiency statuses across both testing occasions.

Figure 6

Average Reliability Metrics for the One-Attribute Conditions

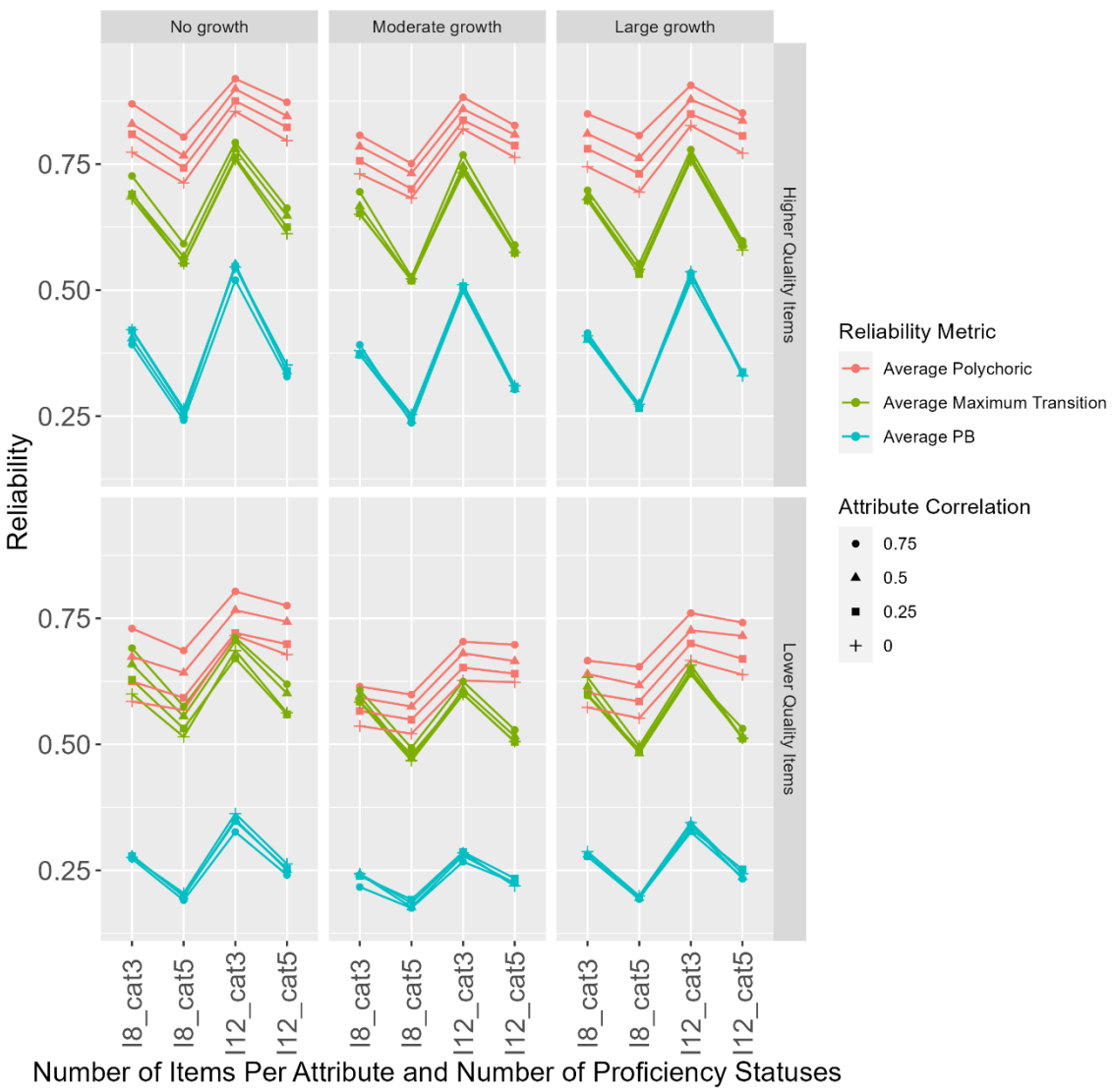


Figure 7

Average Polychoric Reliability Metric for the One-Attribute Conditions

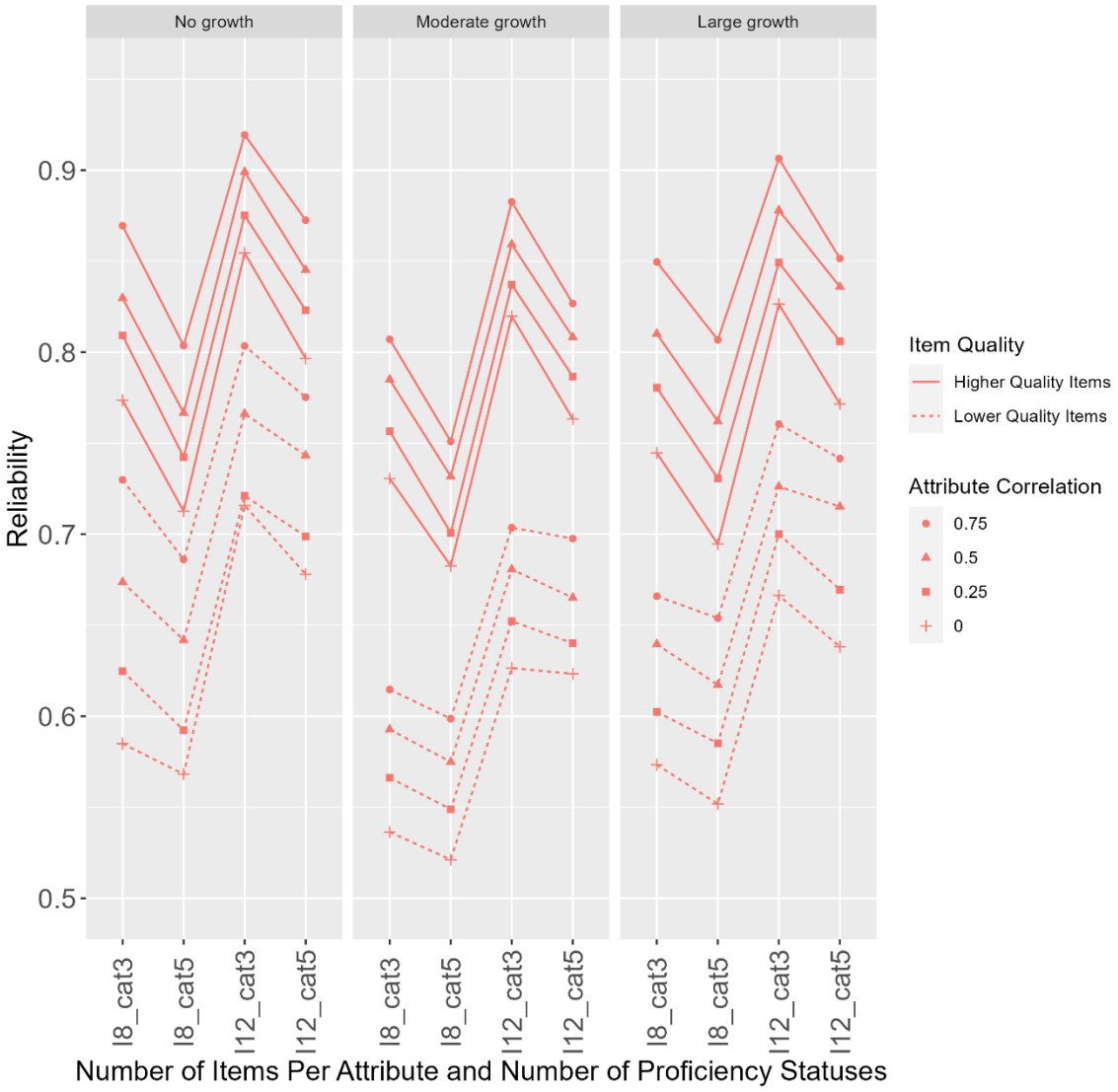


Figure 8

Average Average Maximum Transition Reliability Metric for the One-Attribute Conditions

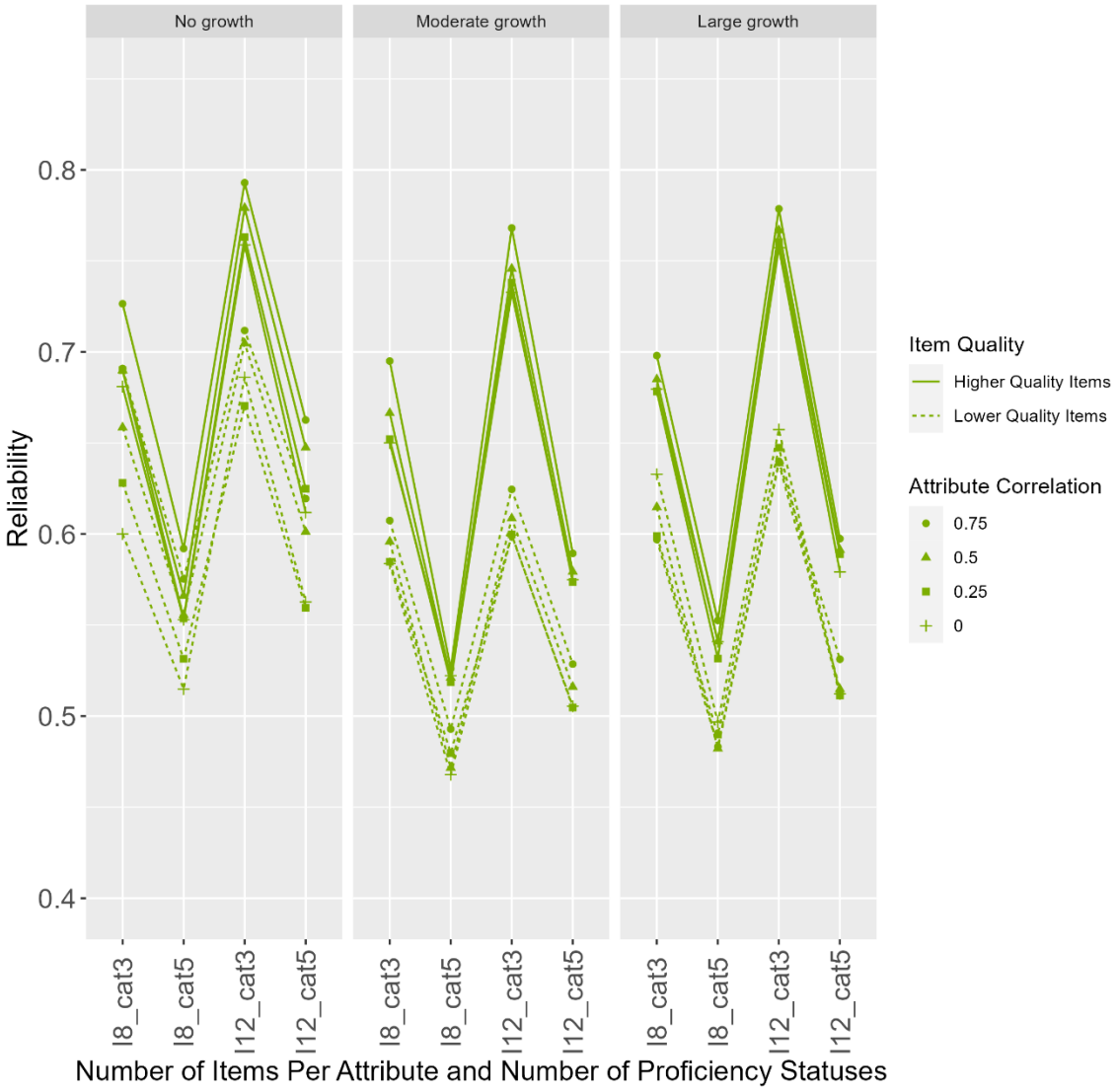


Figure 9

Average Point Biserial Reliability Metric for the One-Attribute Conditions

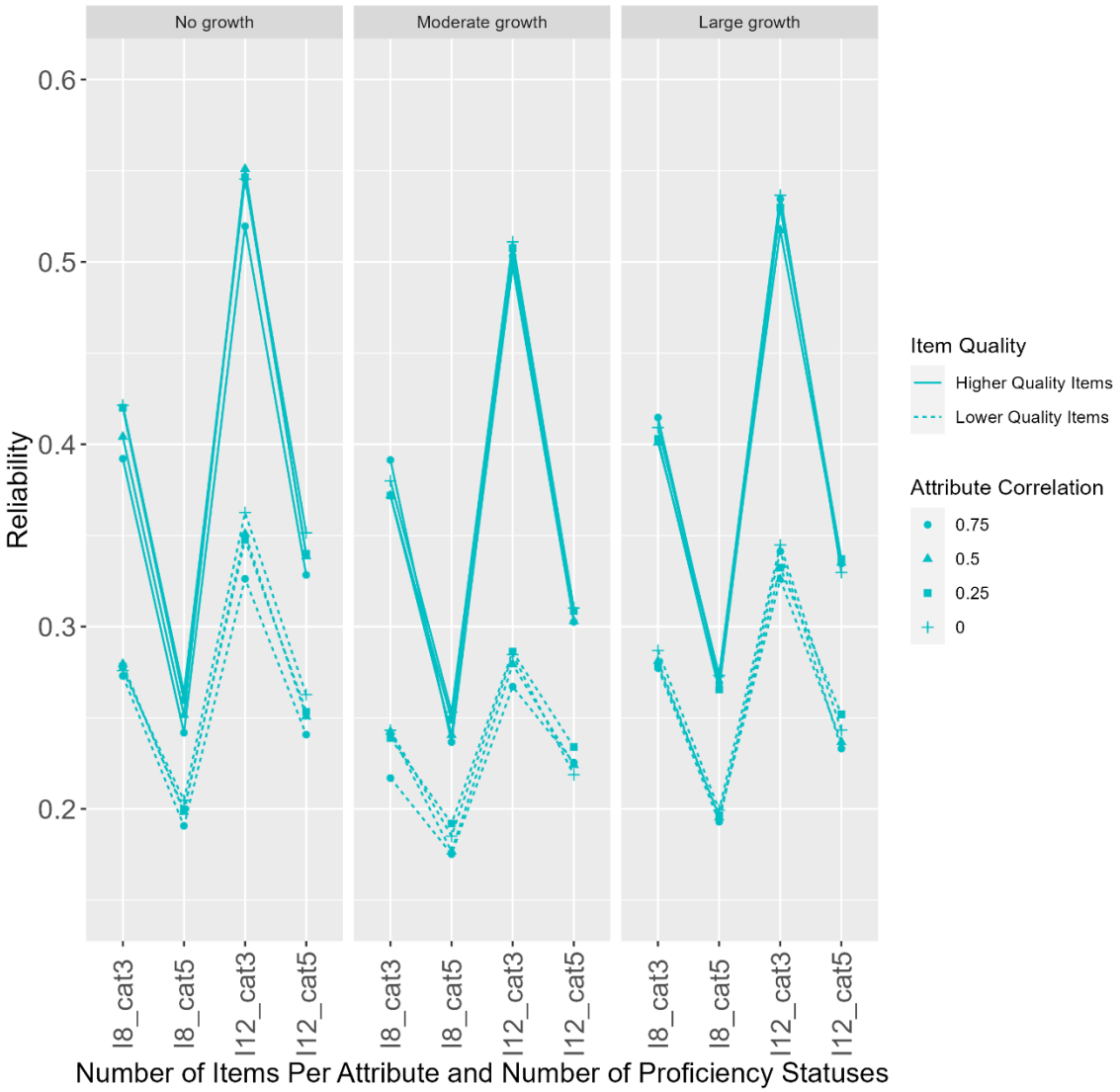


Figure 10

Average Proportion of Maximum Posterior Probabilities for the One-Attribute Conditions

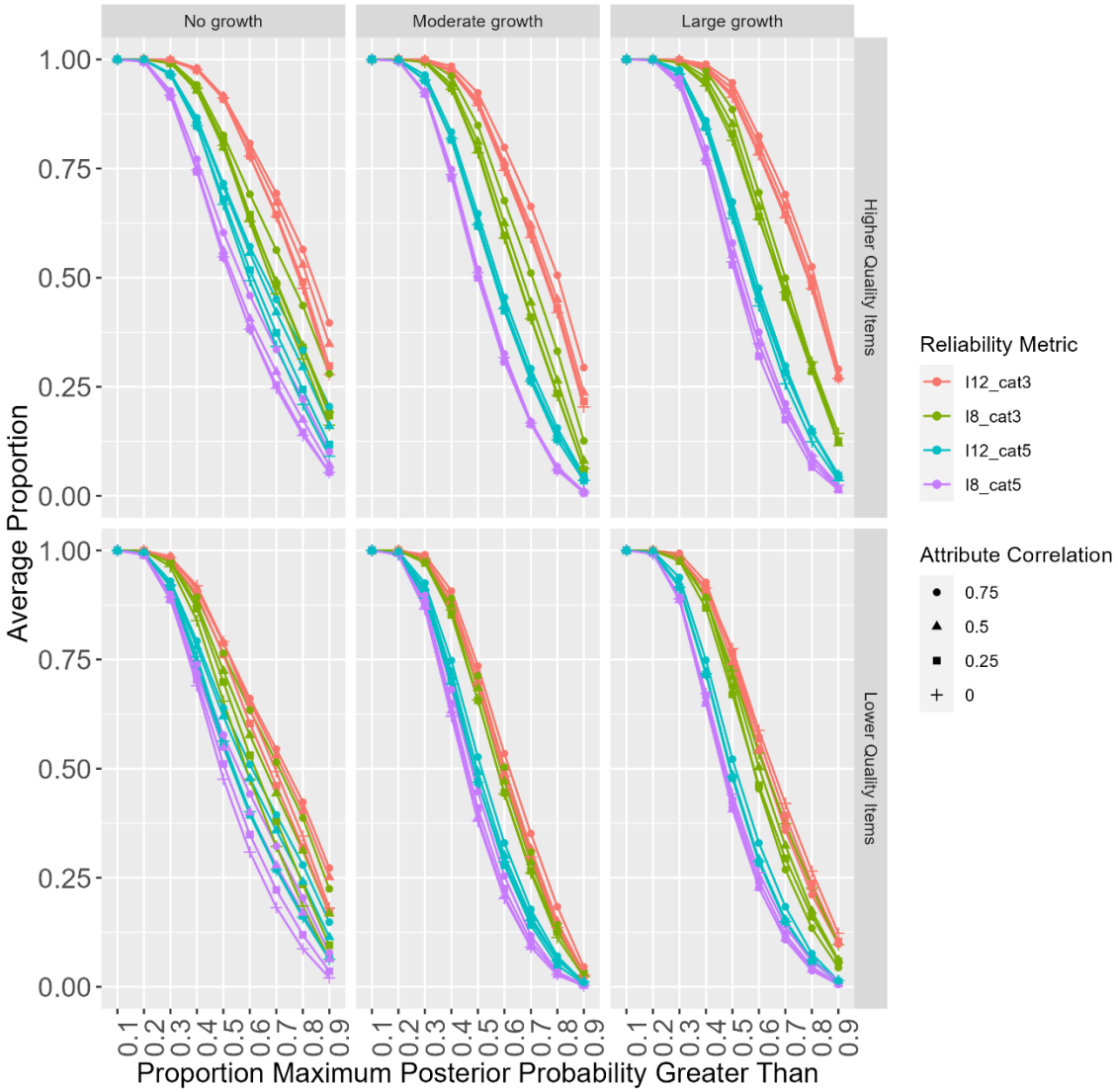
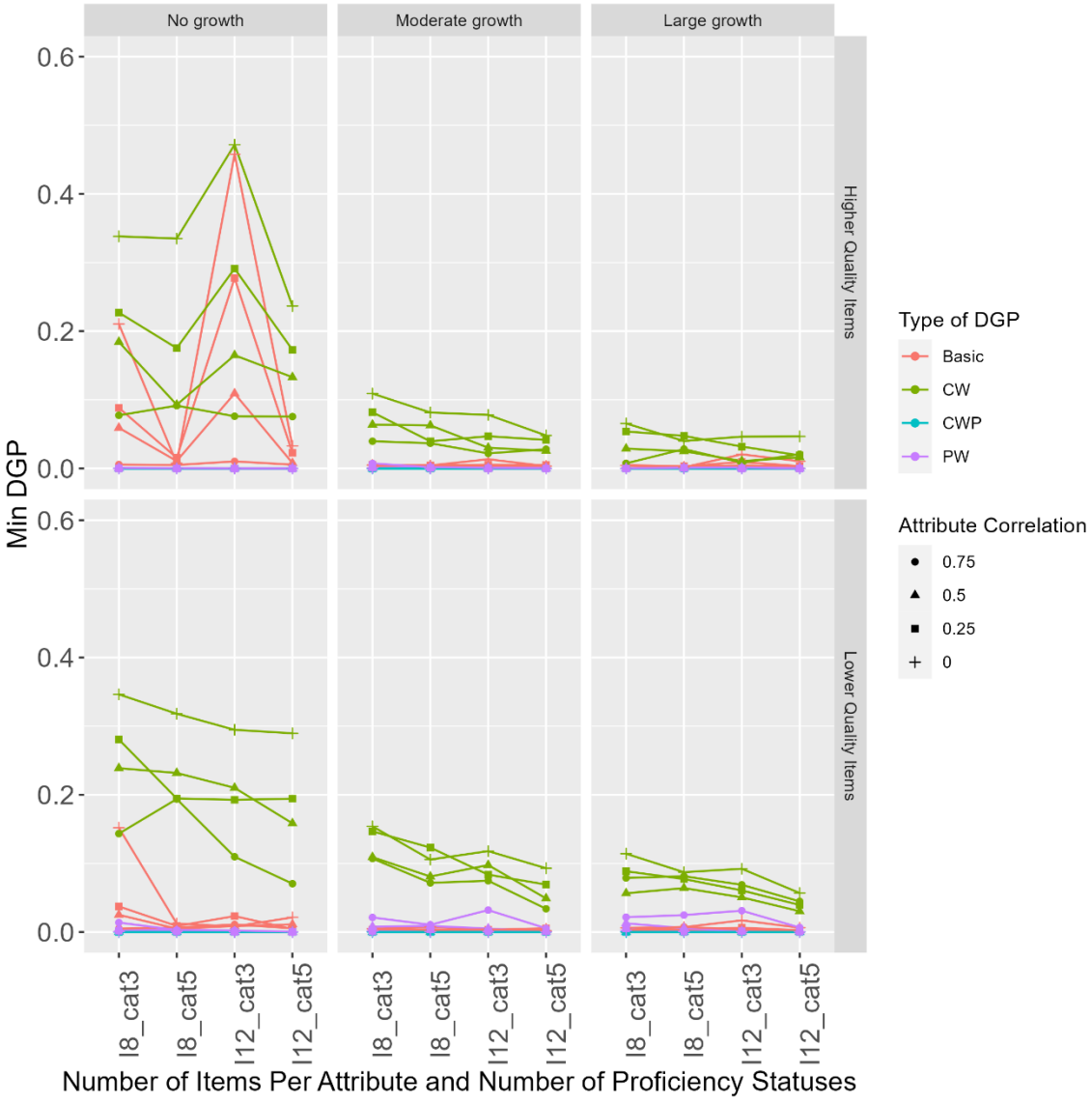


Figure 11

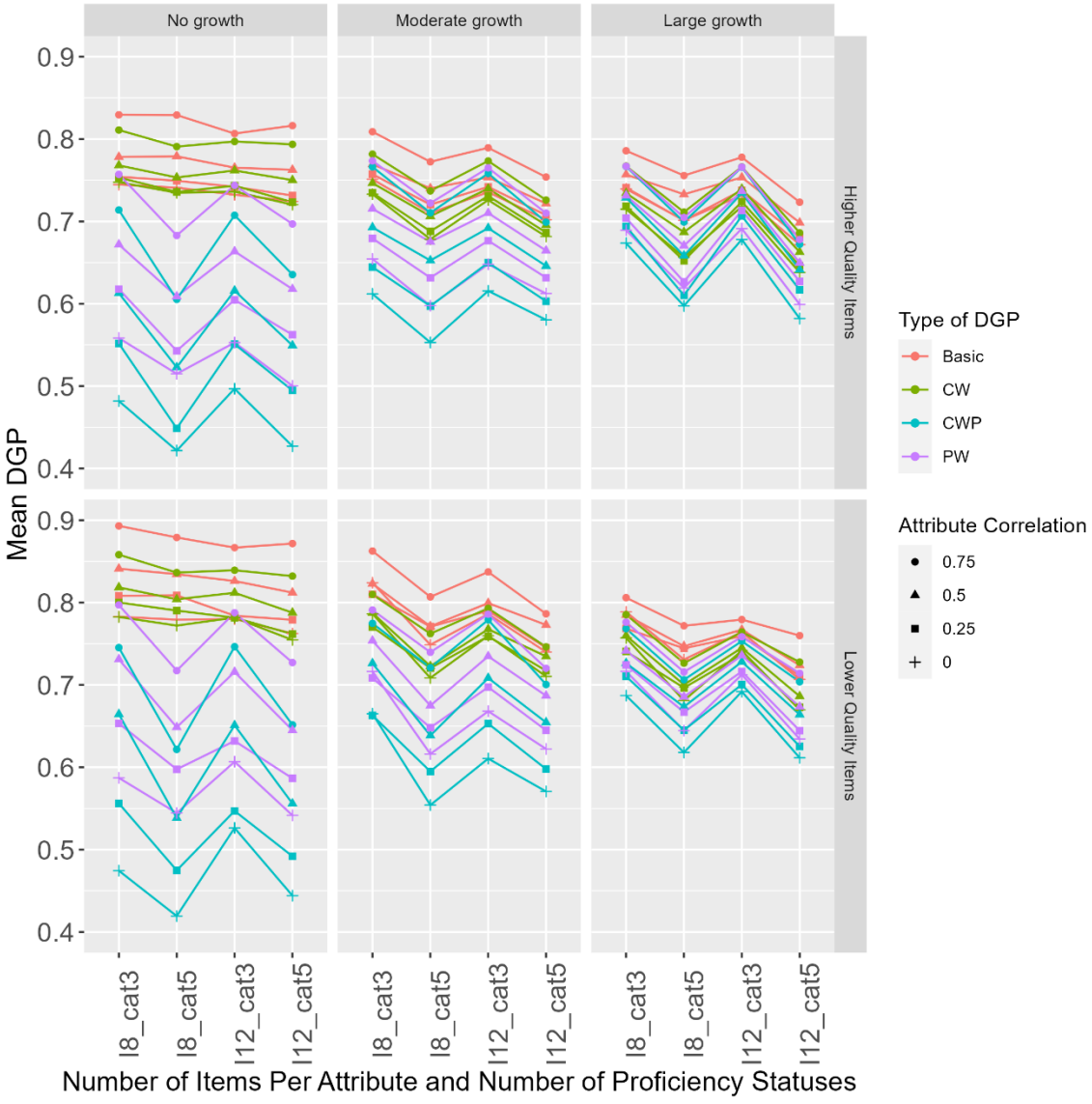
Average Minimum DGPs for the One-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 12

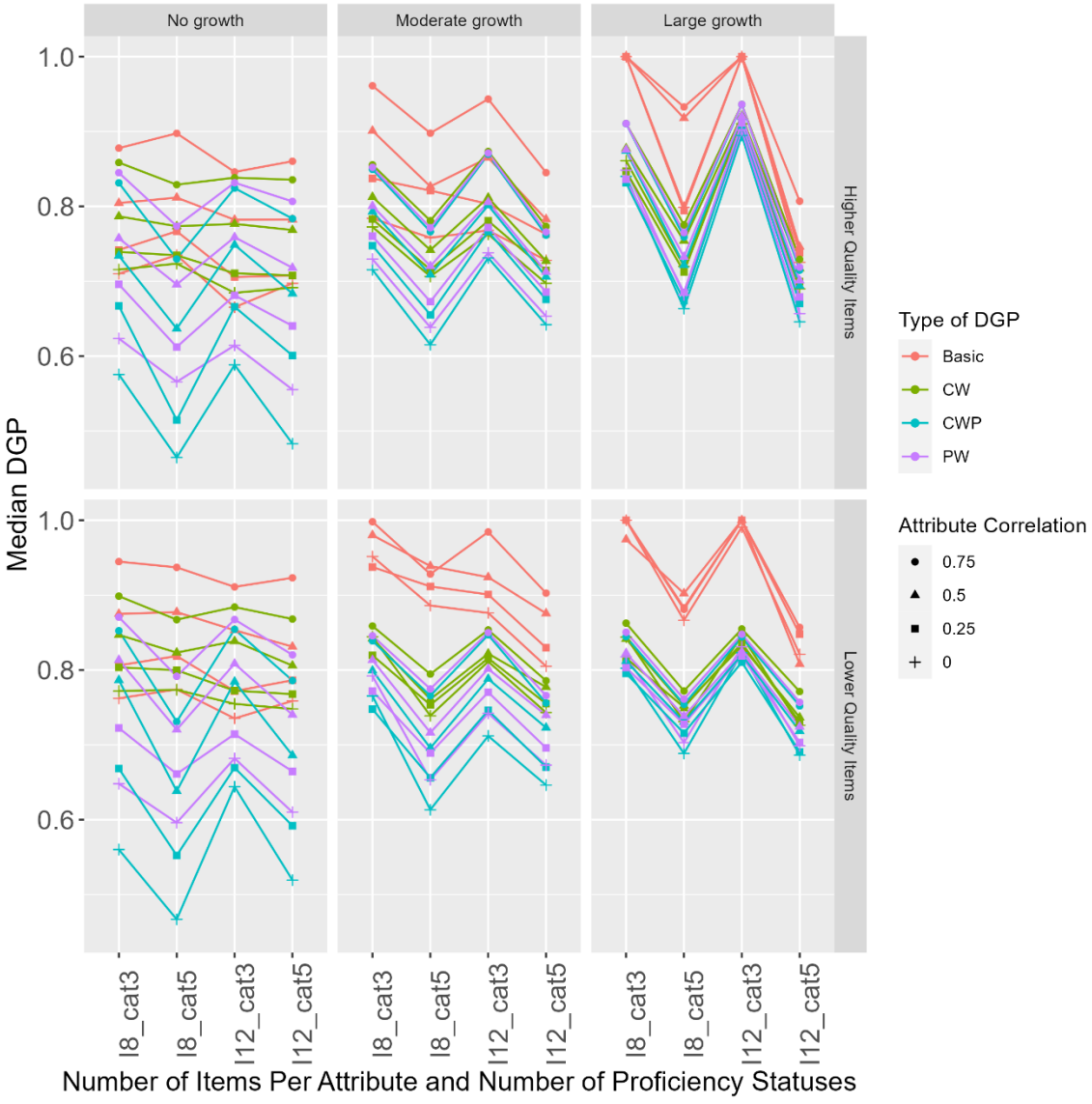
Average Mean DGPs for the One-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 13

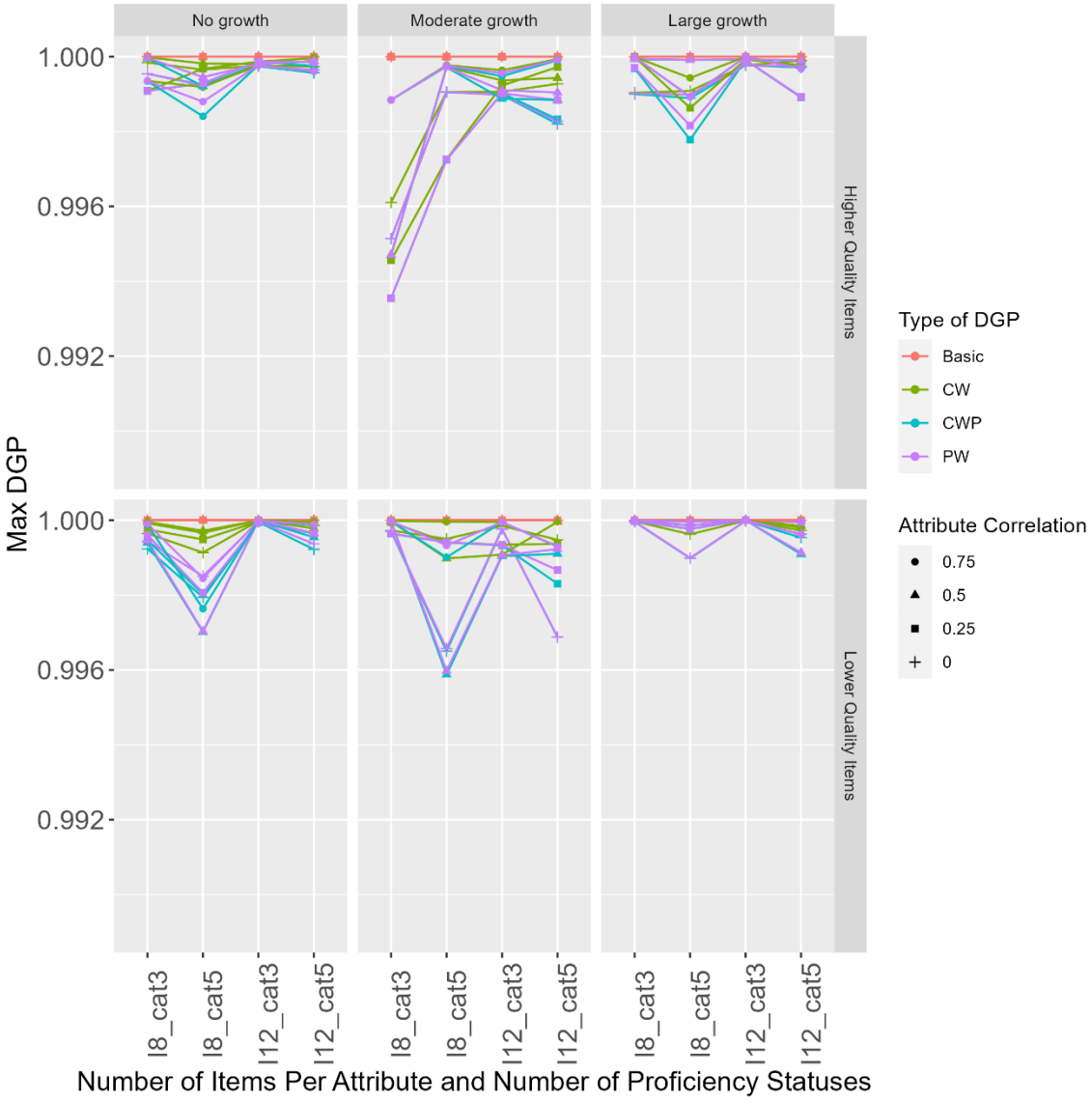
Average Median DGPs for the One-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 14

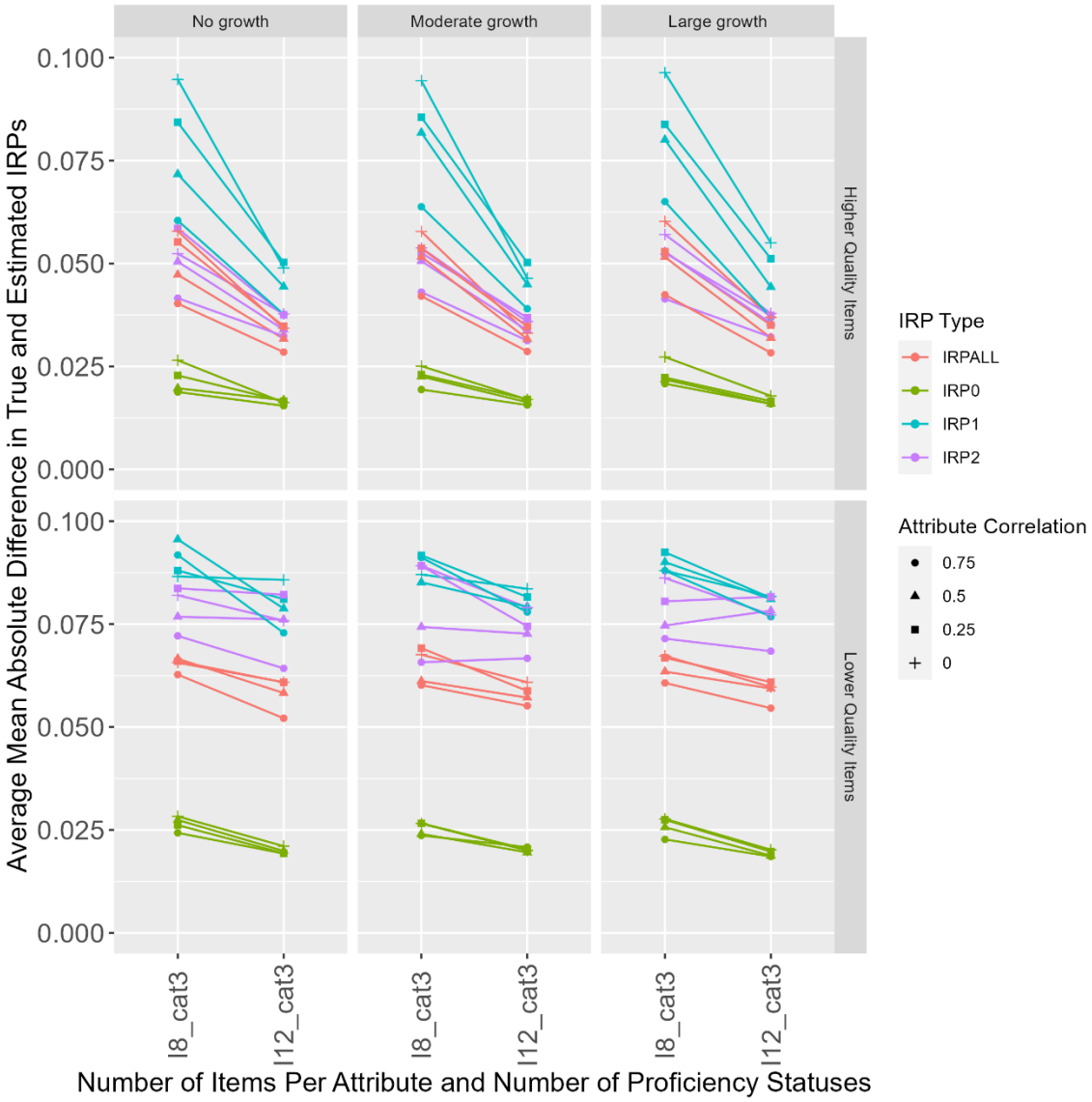
Average Maximum DGPs for the One-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 15

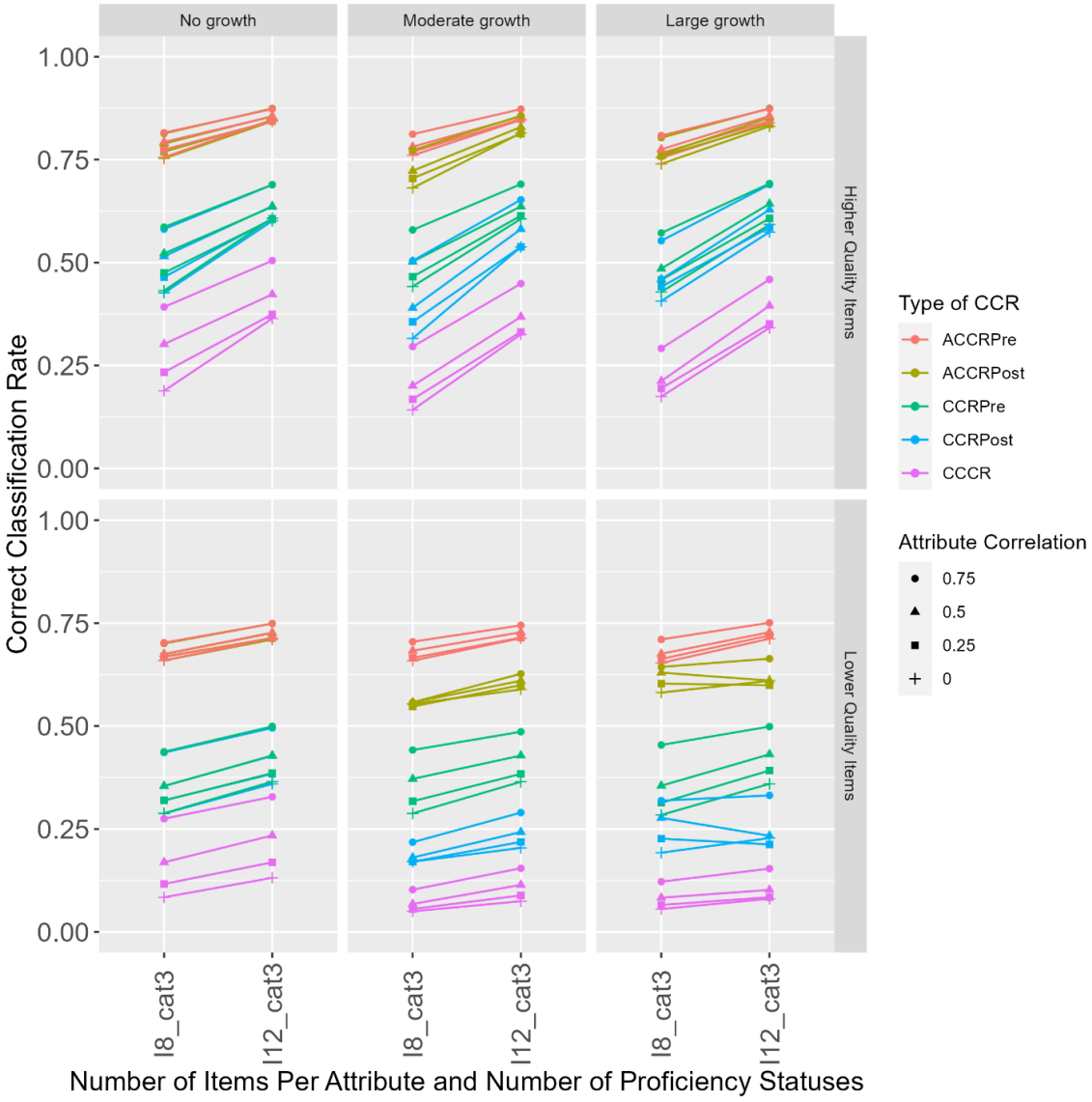
Item Parameter Estimation Accuracy Results: Mean Absolute Difference Between True and Estimated IRPs for the Three-Attribute Conditions



Note. "IRPALL" = average mean absolute difference (MAD) across all IRP types for the condition. "IRP0" = IRP for students with the proficiency status [0]. "IRP1" = IRP for students with the proficiency status [1]. "IRP2" = IRP for students with the proficiency status [2].

Figure 16

Correct Classification Rates for the Three-Attribute Conditions



Note. "ACCRPre" is the attribute-level correct classification rate for the pre-test only. "ACCRPost" is the attribute-level correct classification rate for the post-test only. For conditions with multiple attributes, the ACCRs are averaged across the attributes within the same testing occasion. "CCRPre" is the class-level correct classification rate for the pre-test only. "CCRPost" is the class-level correct classification rate for the post-test

only. “CCCR” is the overall, longitudinal class-level correct classification rate, which refers to students’ entire profiles of proficiency statuses across both testing occasions.

Figure 17

Average Reliability Metrics for the Three-Attribute Conditions

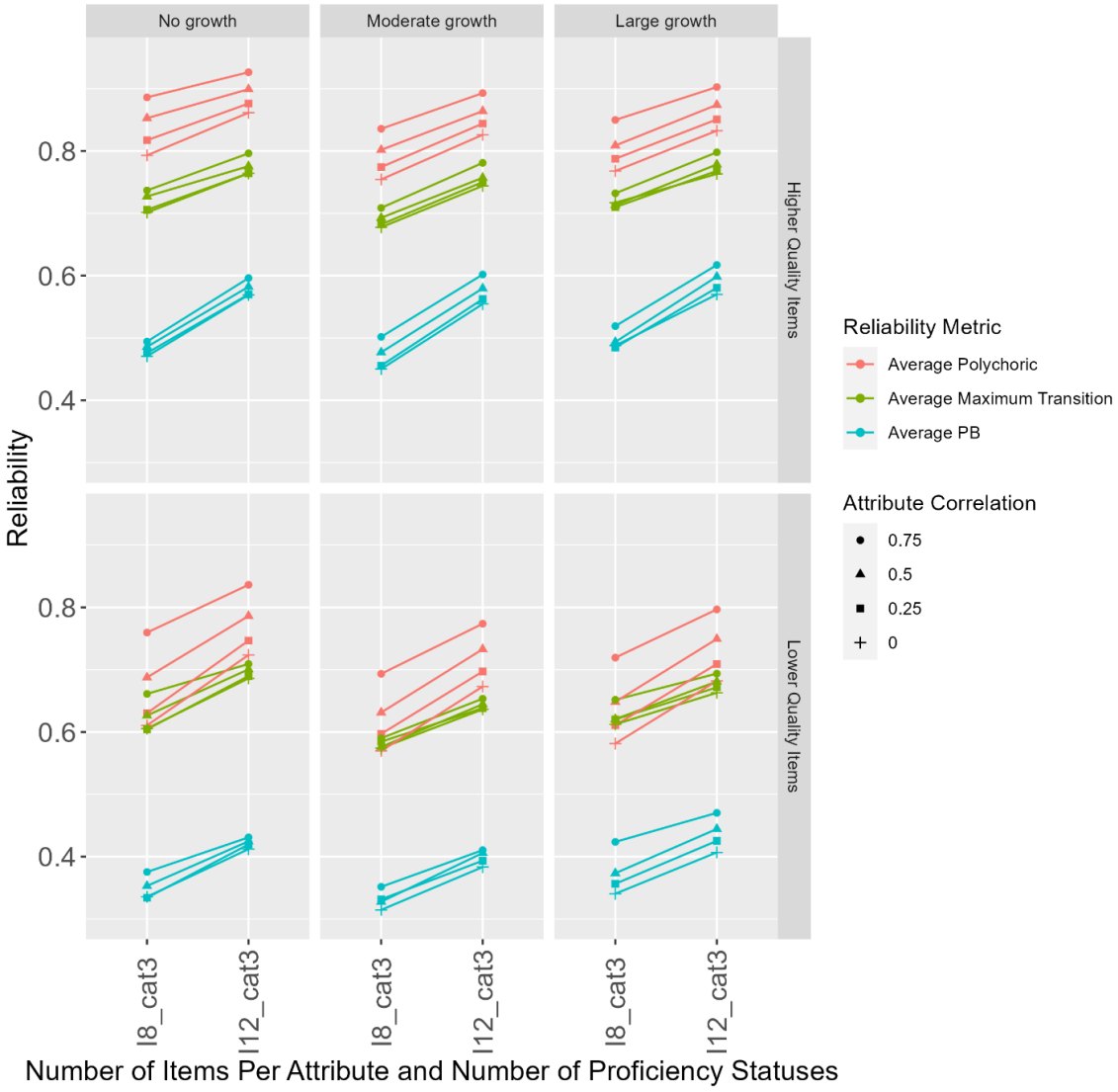


Figure 18

Average Polychoric Reliability Metric for the Three-Attribute Conditions

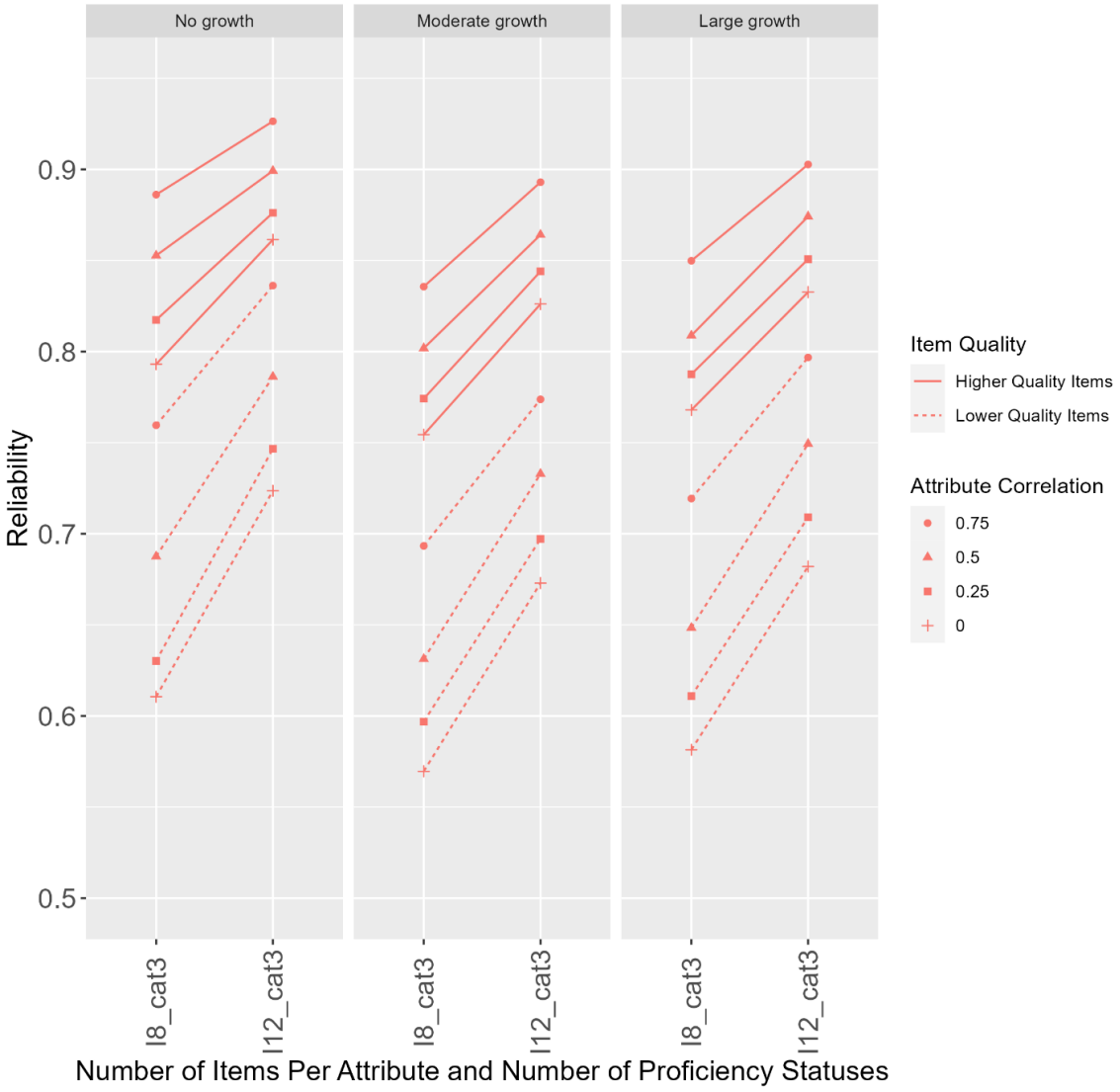


Figure 19

Average Average Maximum Transition Reliability Metric for the Three-Attribute Conditions

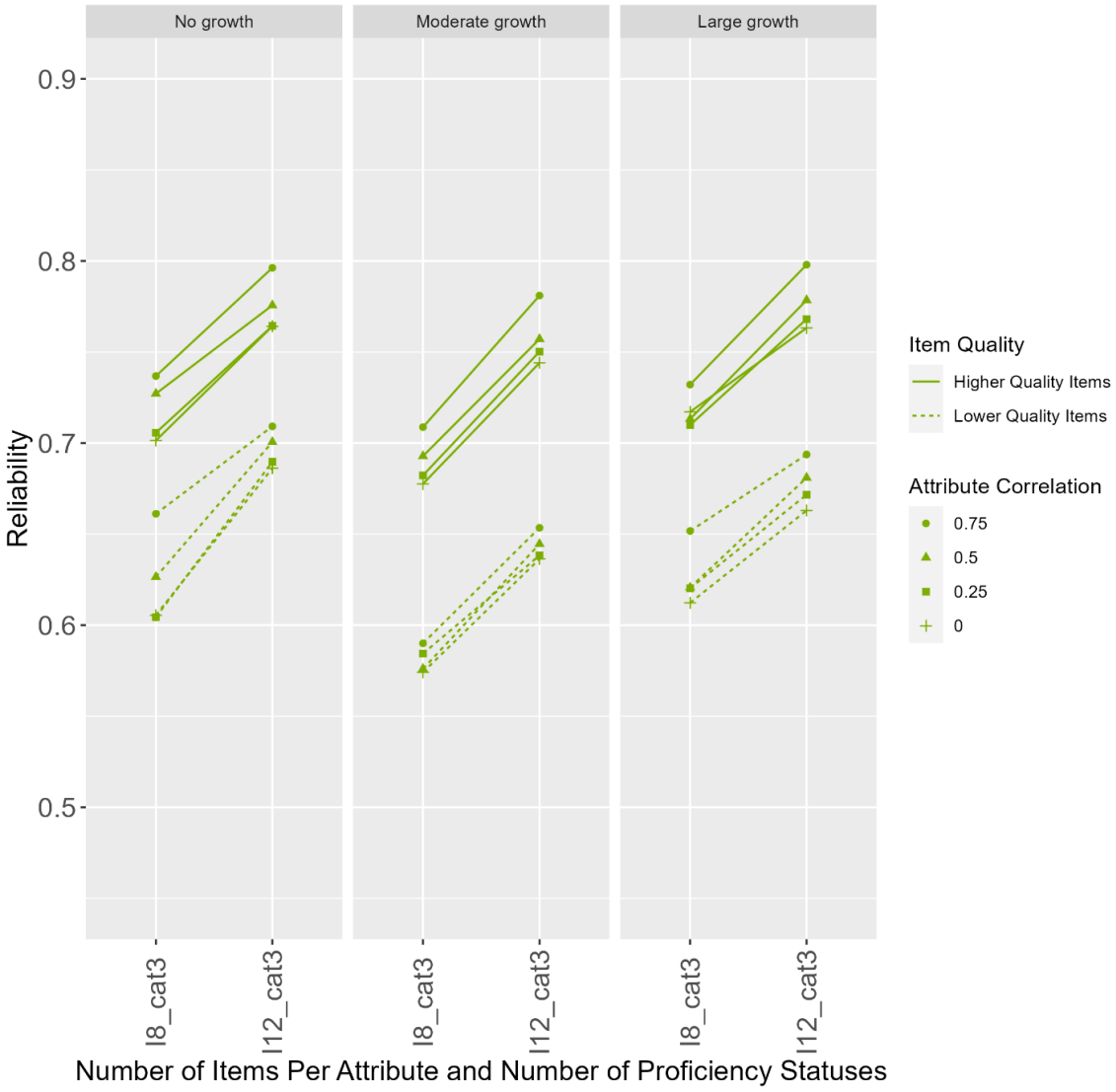


Figure 20

Average Point Biserial Reliability Metric for the Three-Attribute Conditions

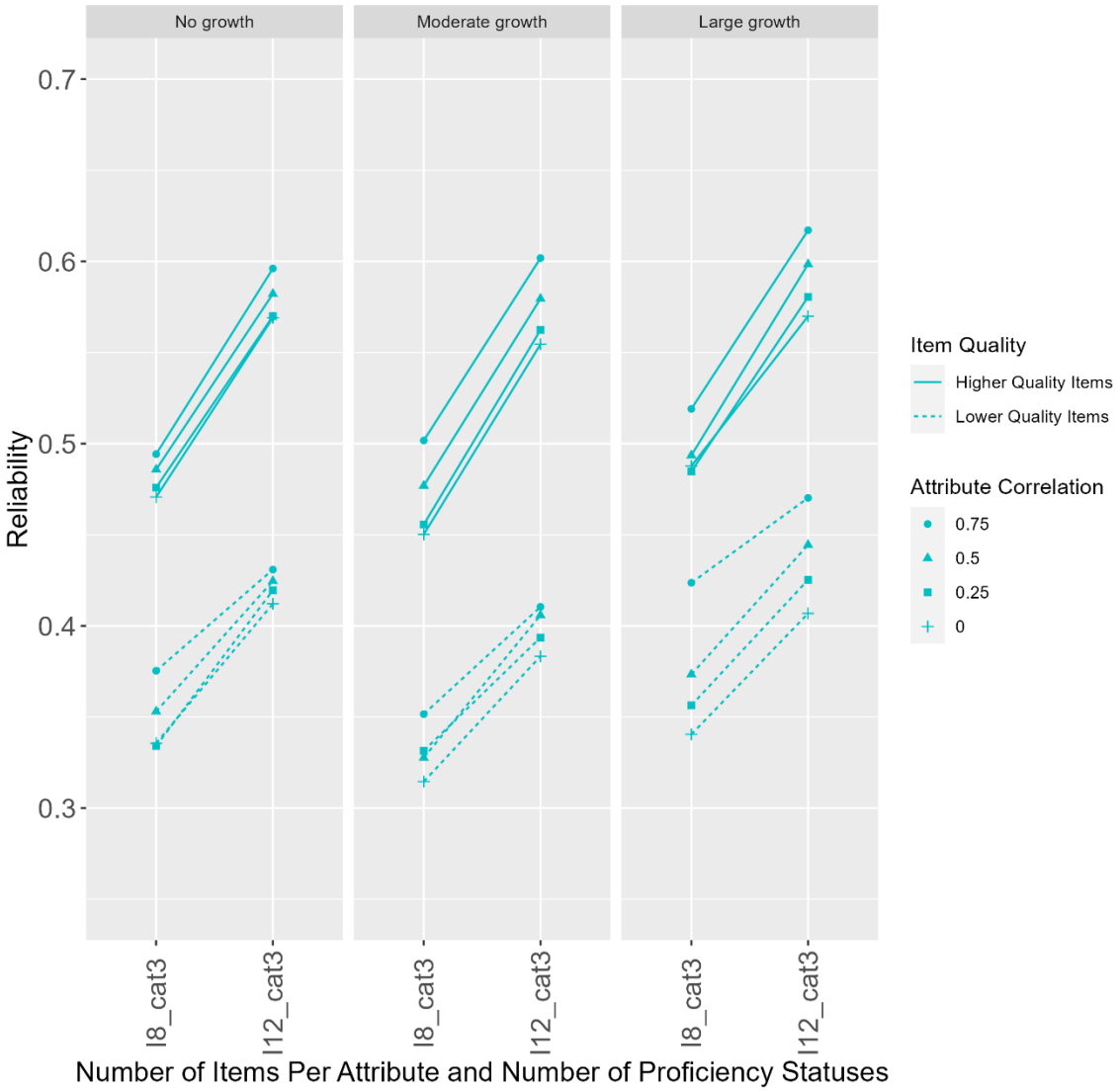


Figure 21

Average Proportion of Maximum Posterior Probabilities for the Three-Attribute Conditions

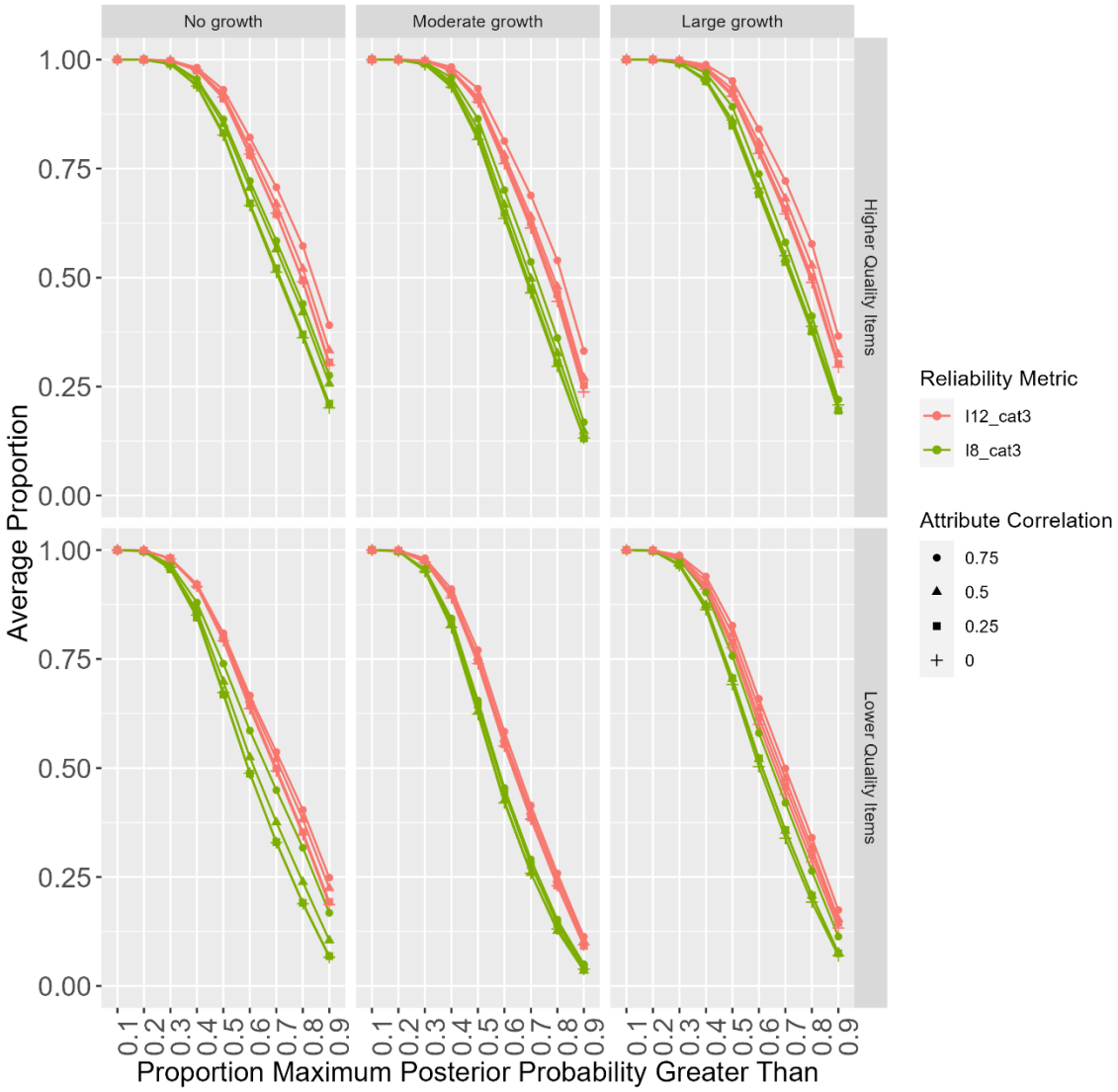
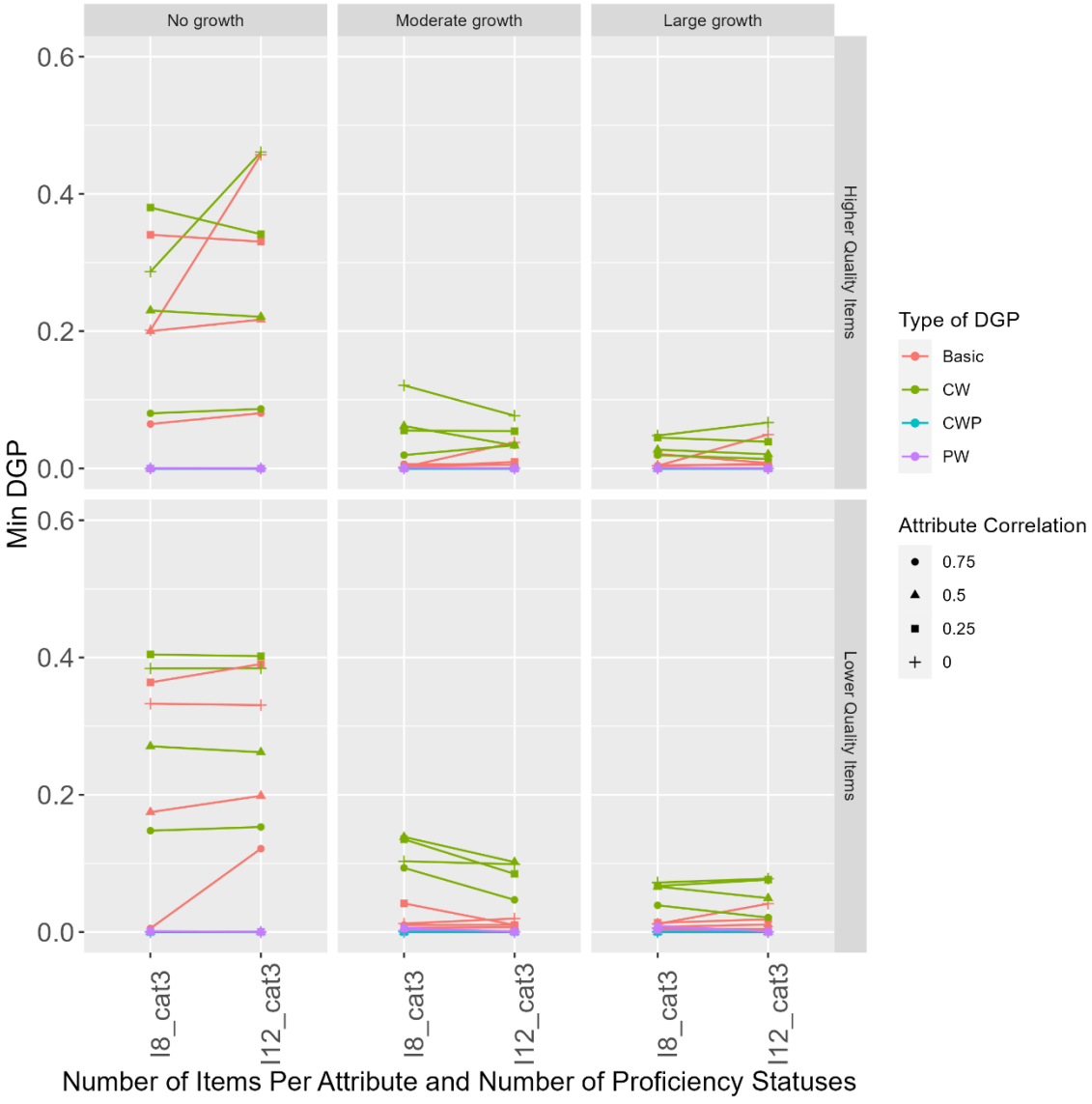


Figure 22

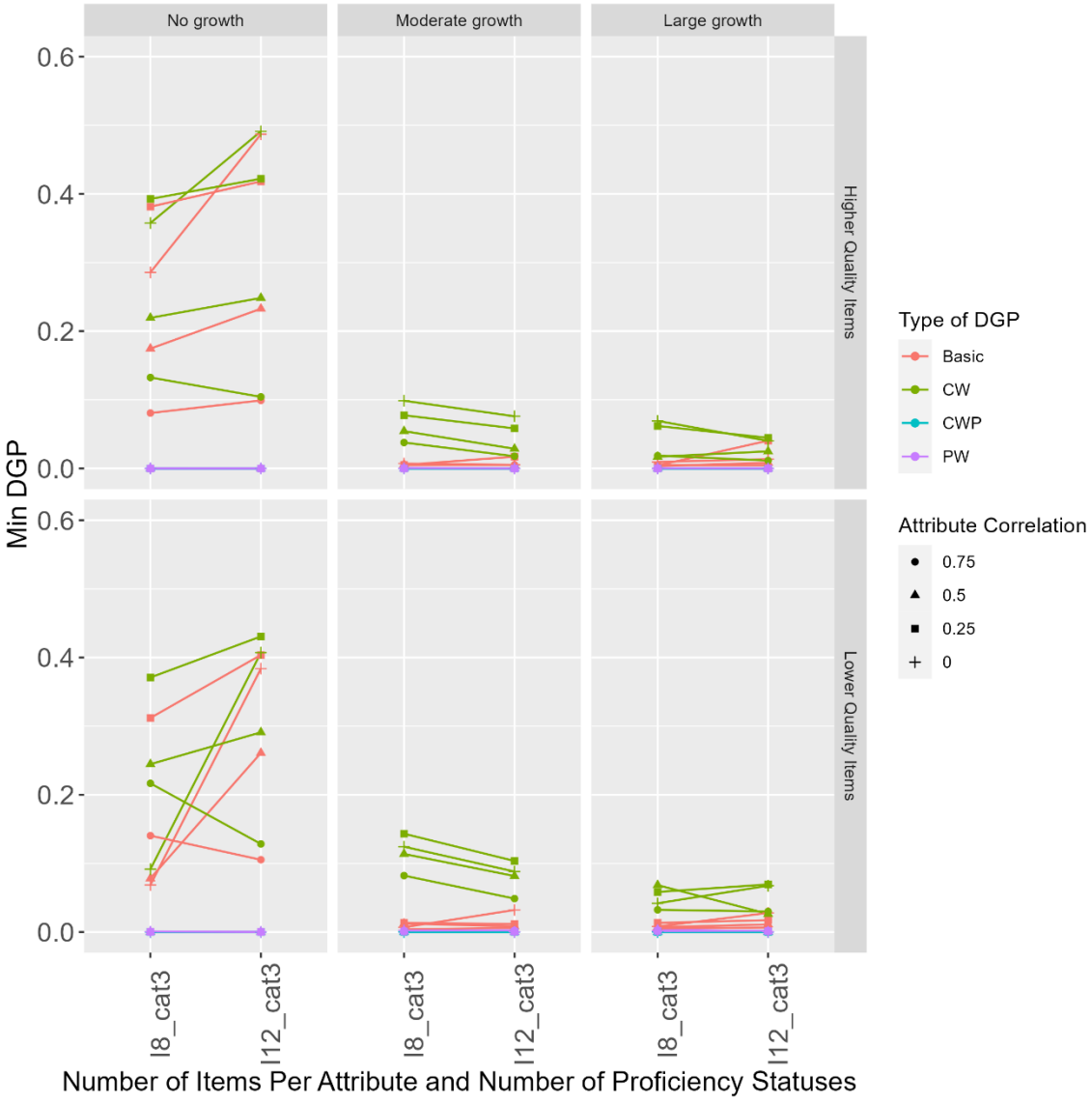
Average Minimum DGPs for Attribute 1 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 23

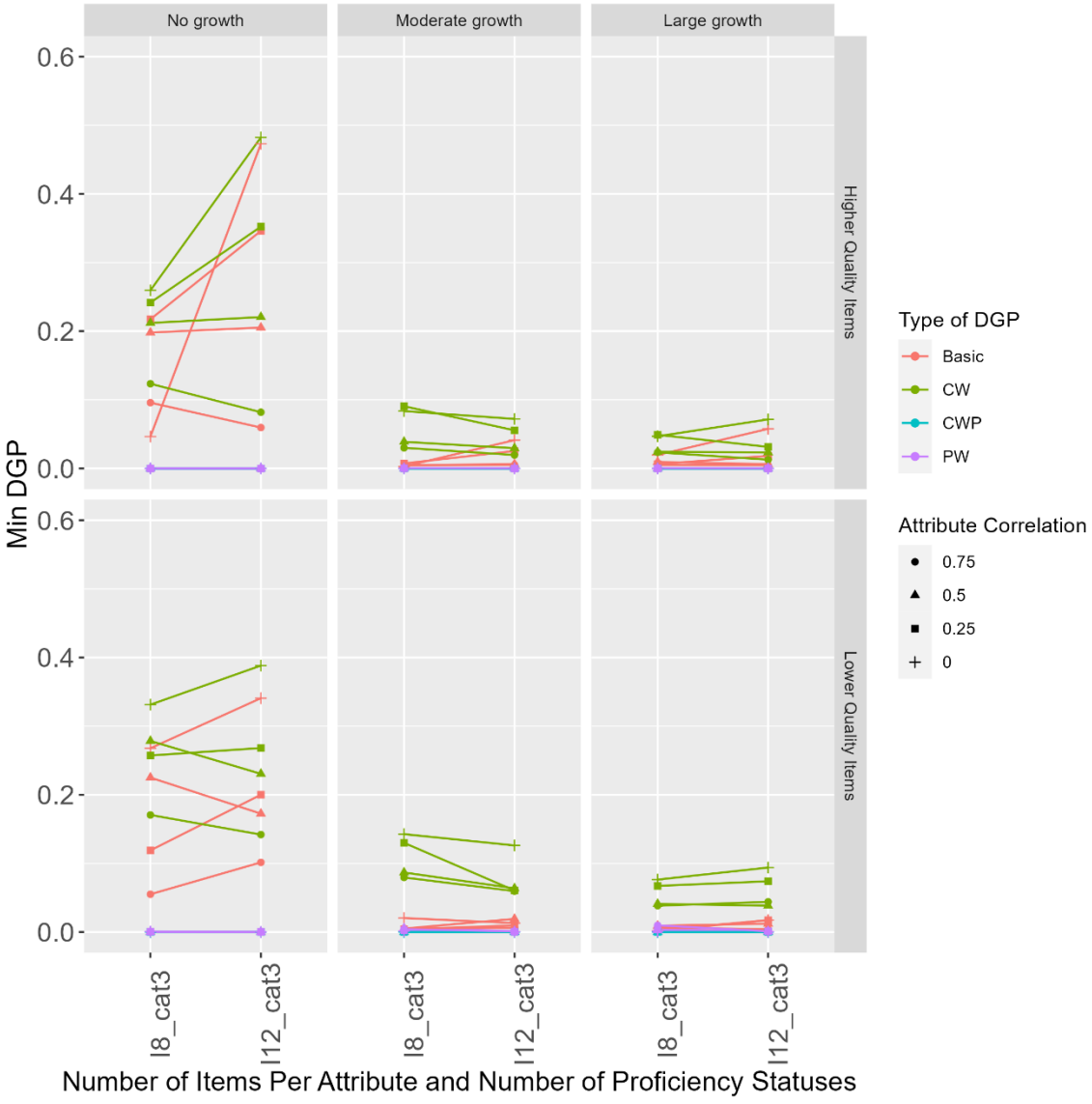
Average Minimum DGPs for Attribute 2 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 24

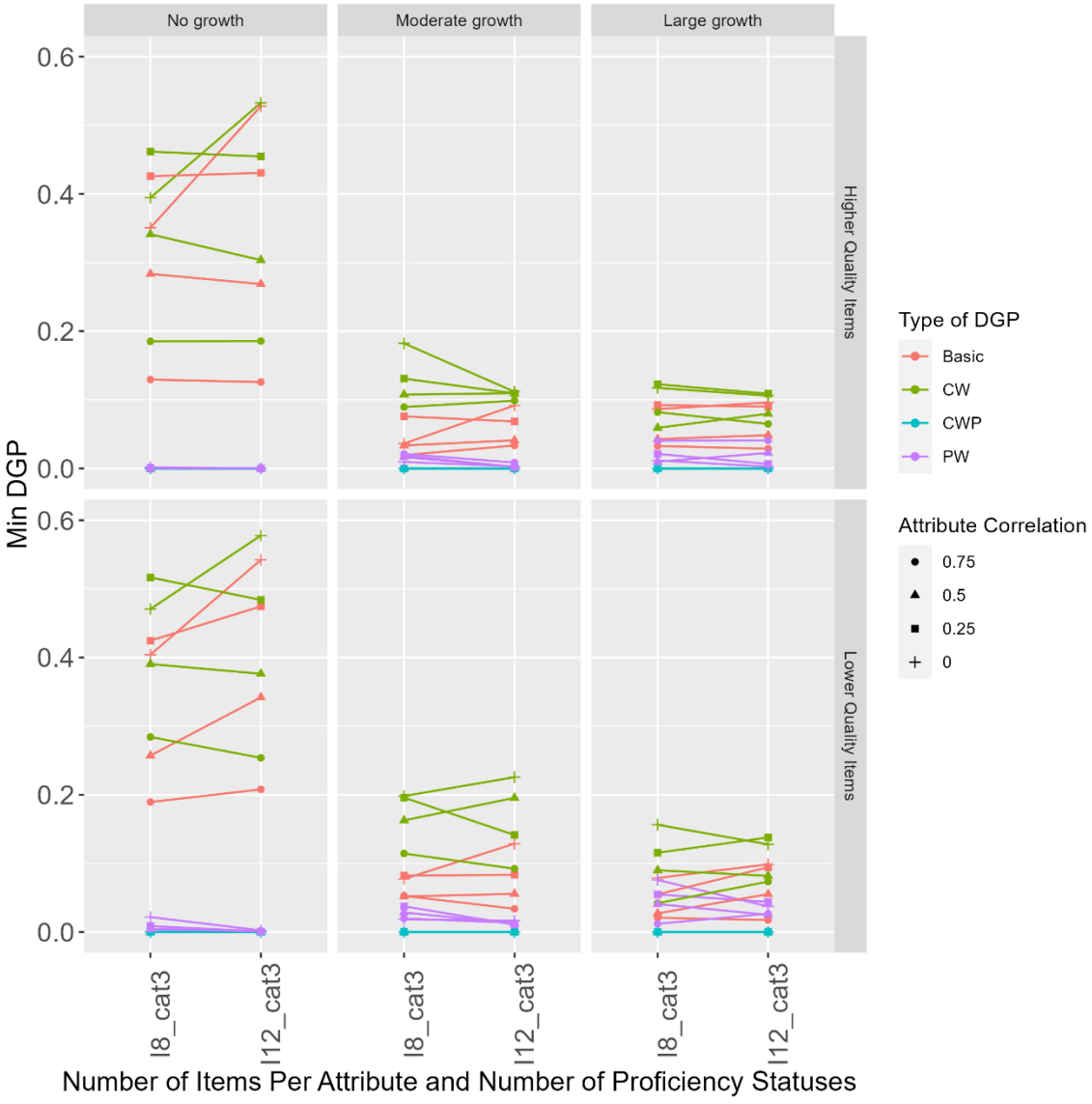
Average Minimum DGPs for Attribute 3 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 25

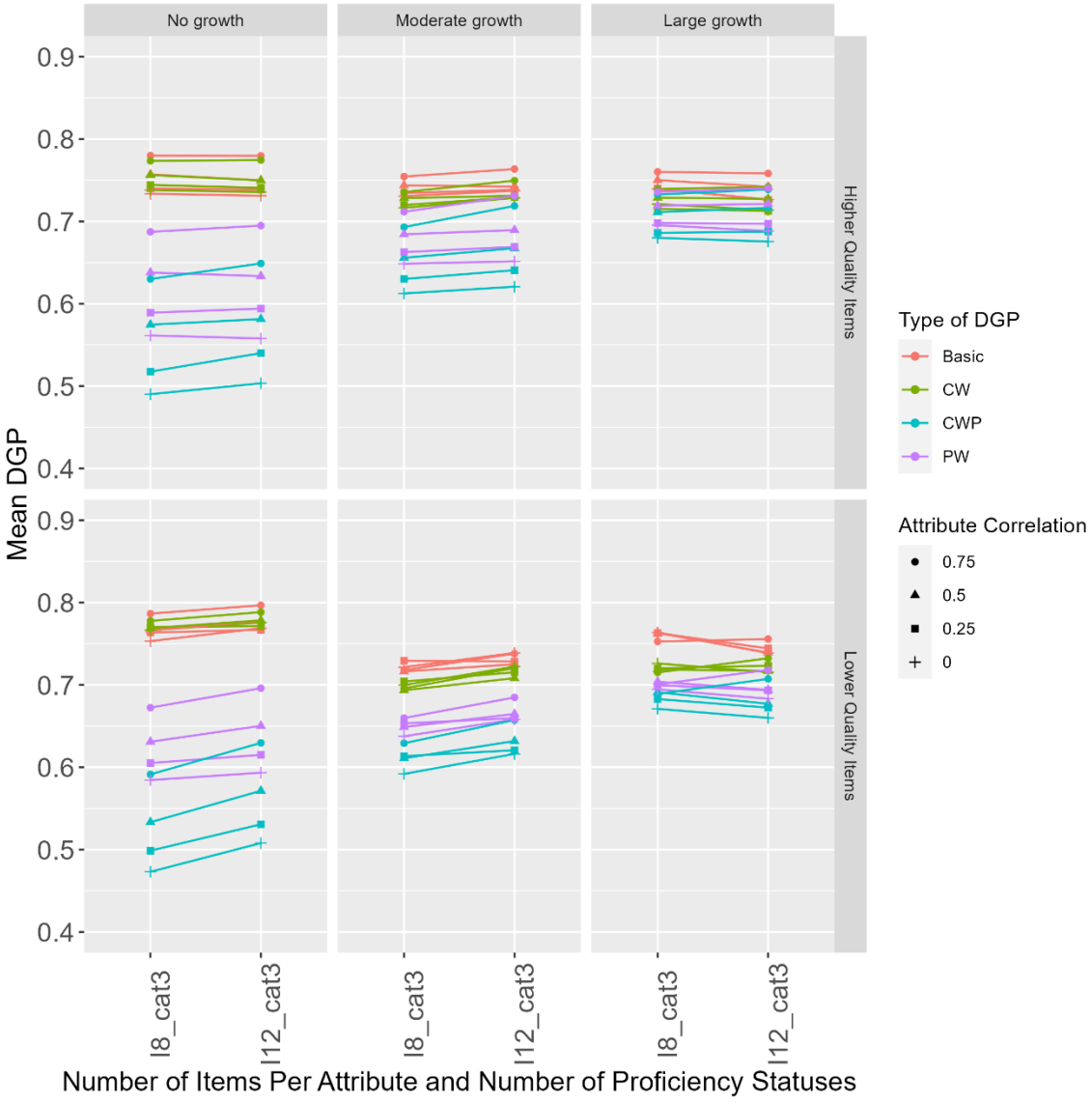
Average Minimum DGPs for the Profile-Level DGPs in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 26

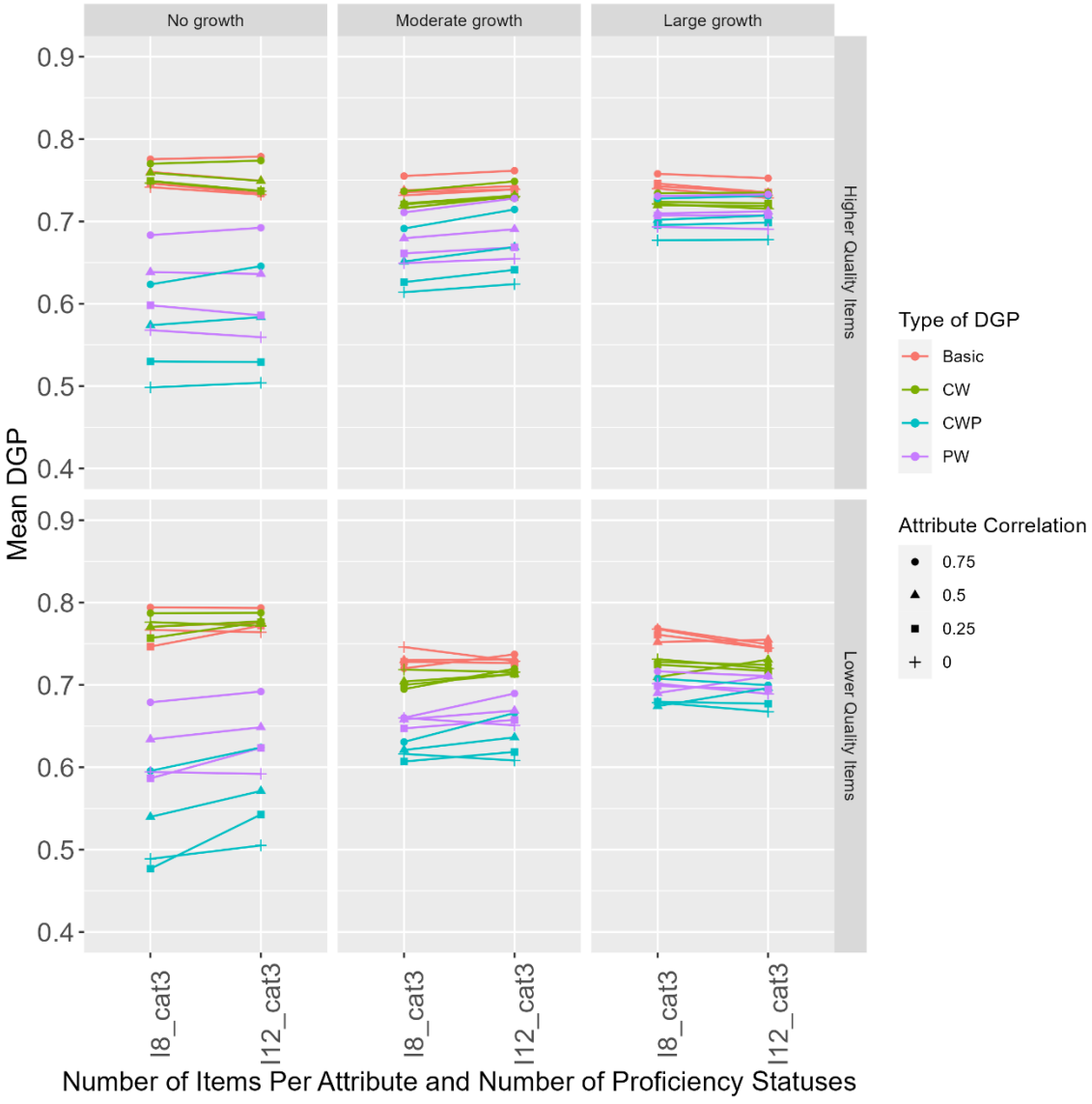
Average Mean DGPs for Attribute 1 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 27

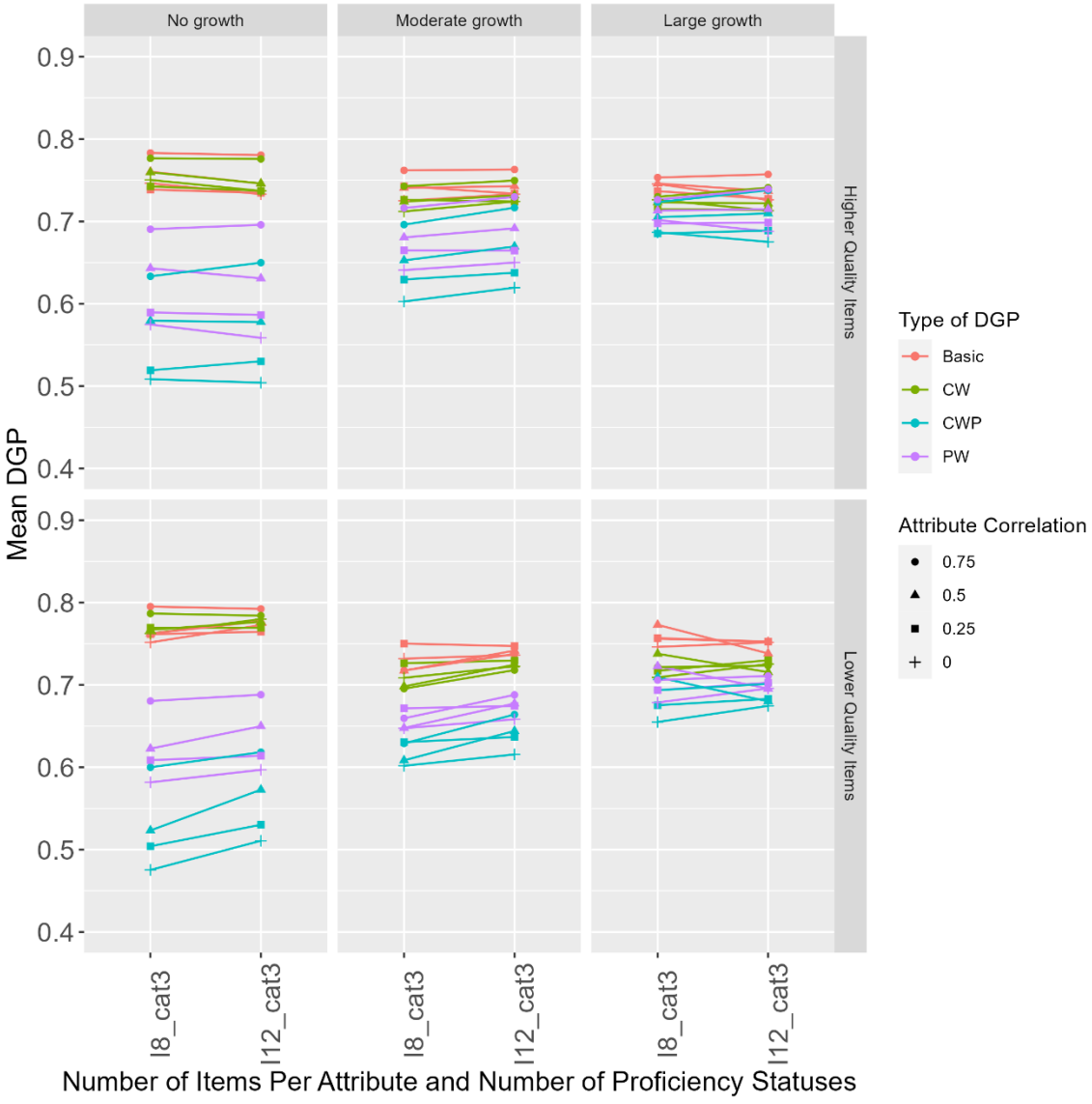
Average Mean DGPs for Attribute 2 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 28

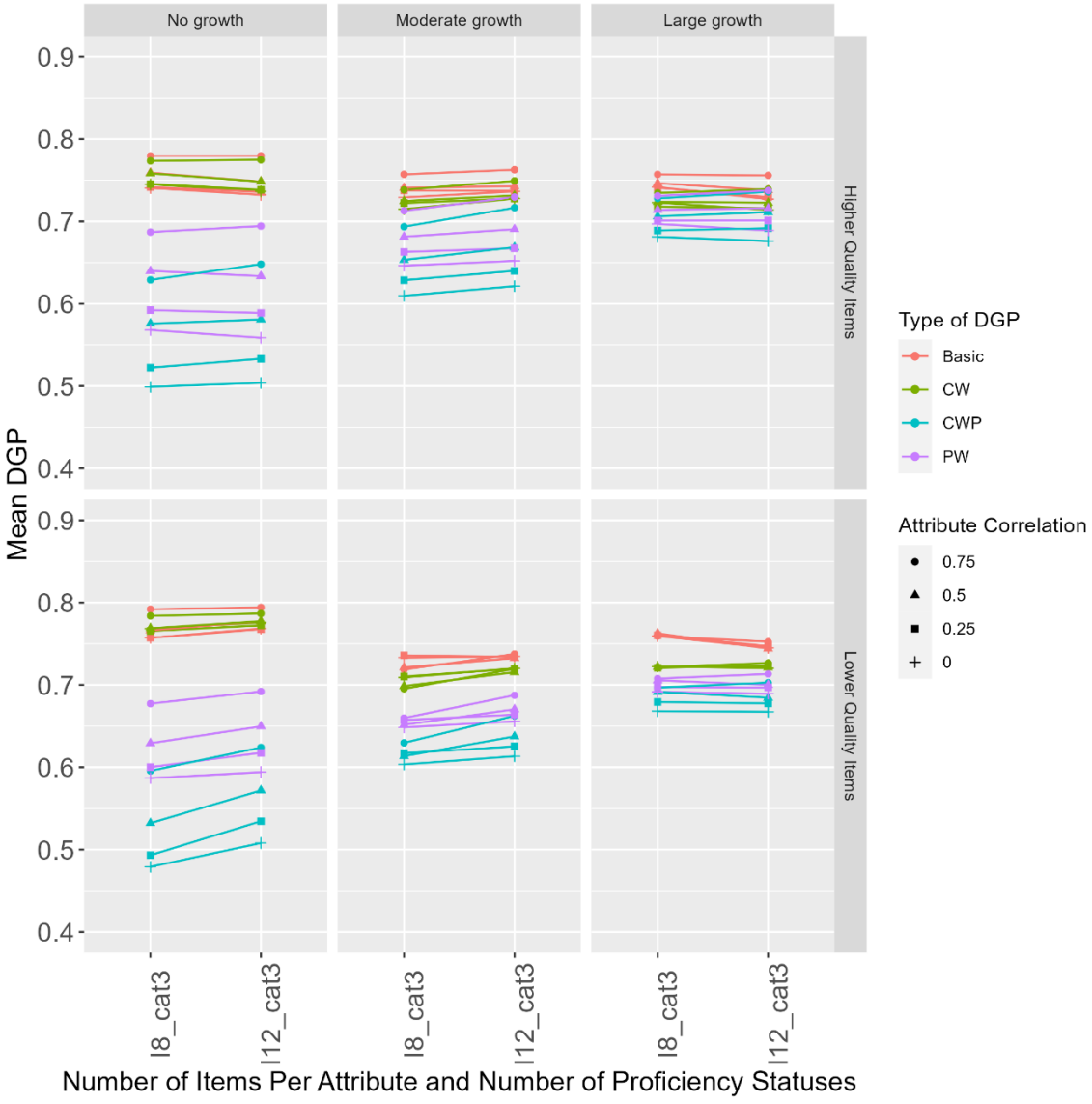
Average Mean DGPs for Attribute 3 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 29

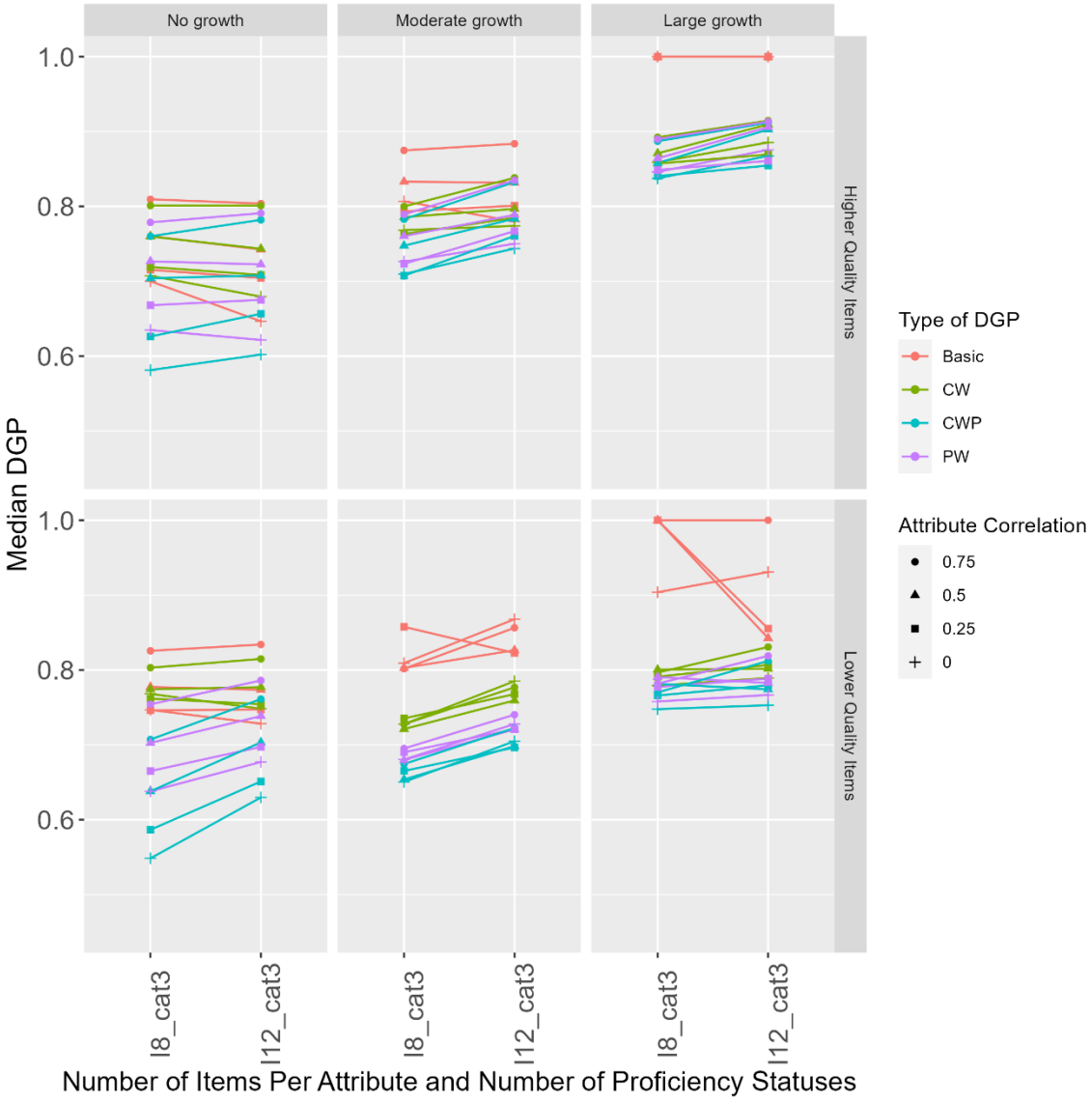
Average Mean DGPs for Profile-Level DGPs in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 30

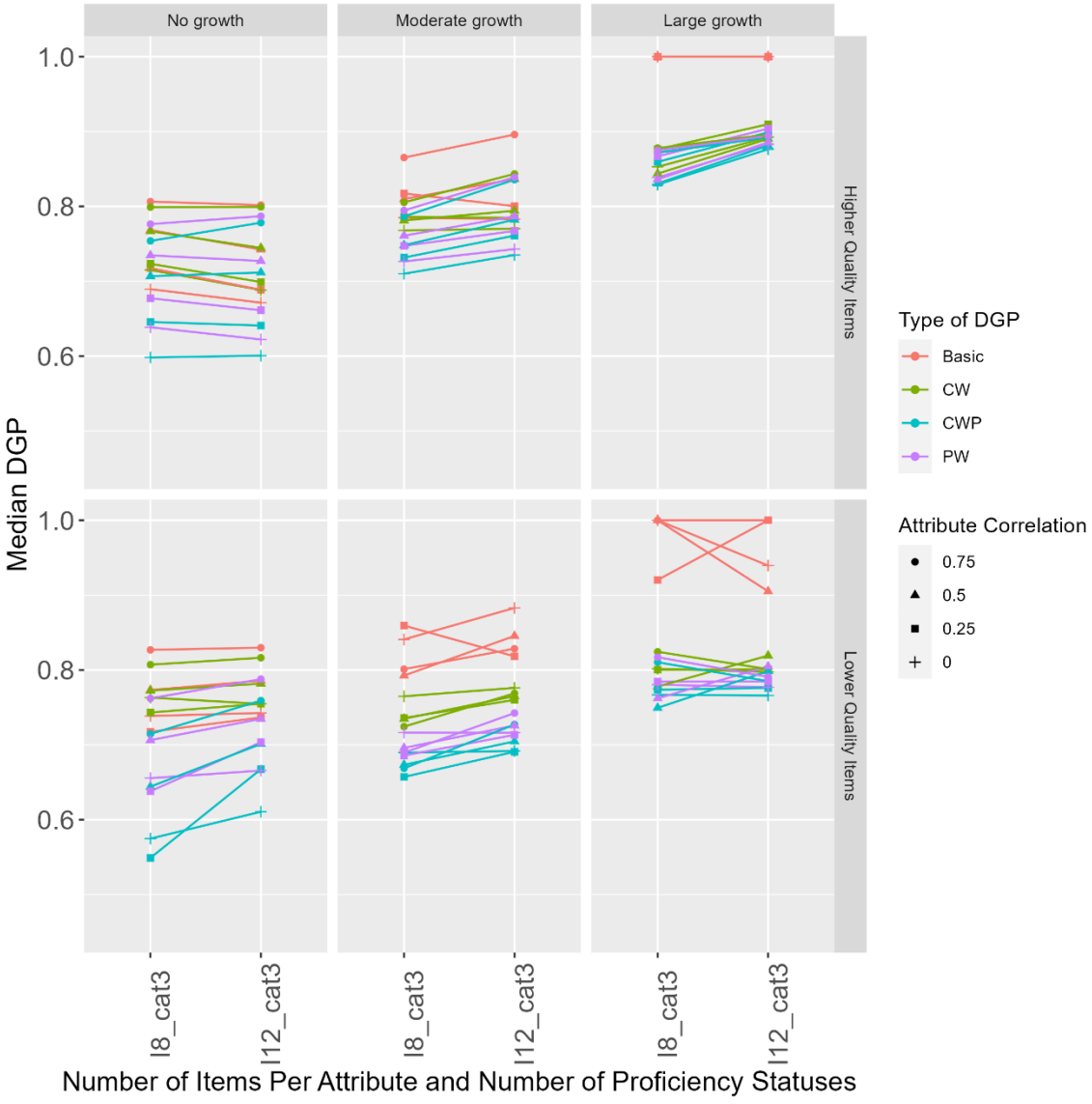
Average Median DGPs for Attribute 1 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 31

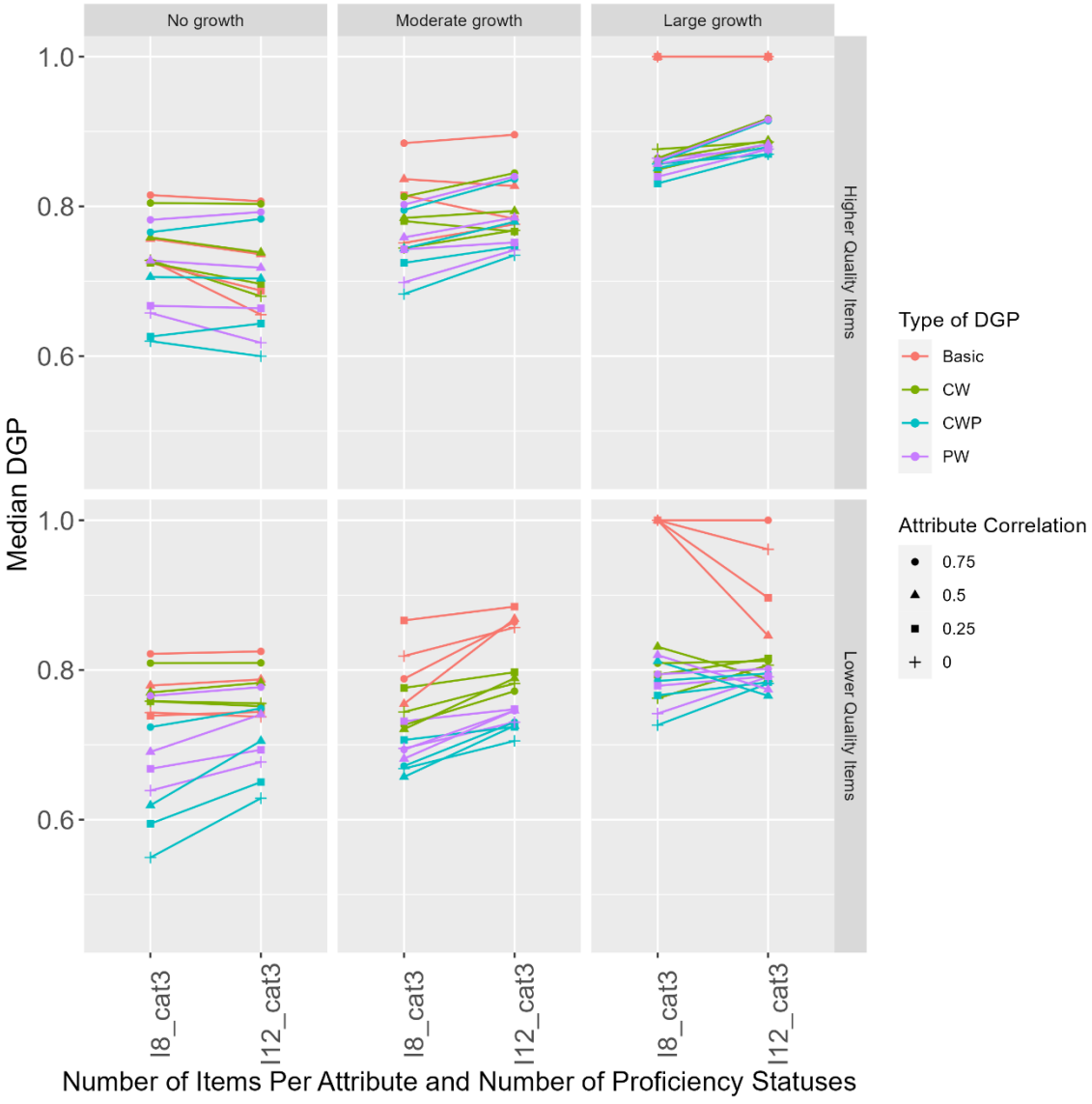
Average Median DGPs for Attribute 2 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 32

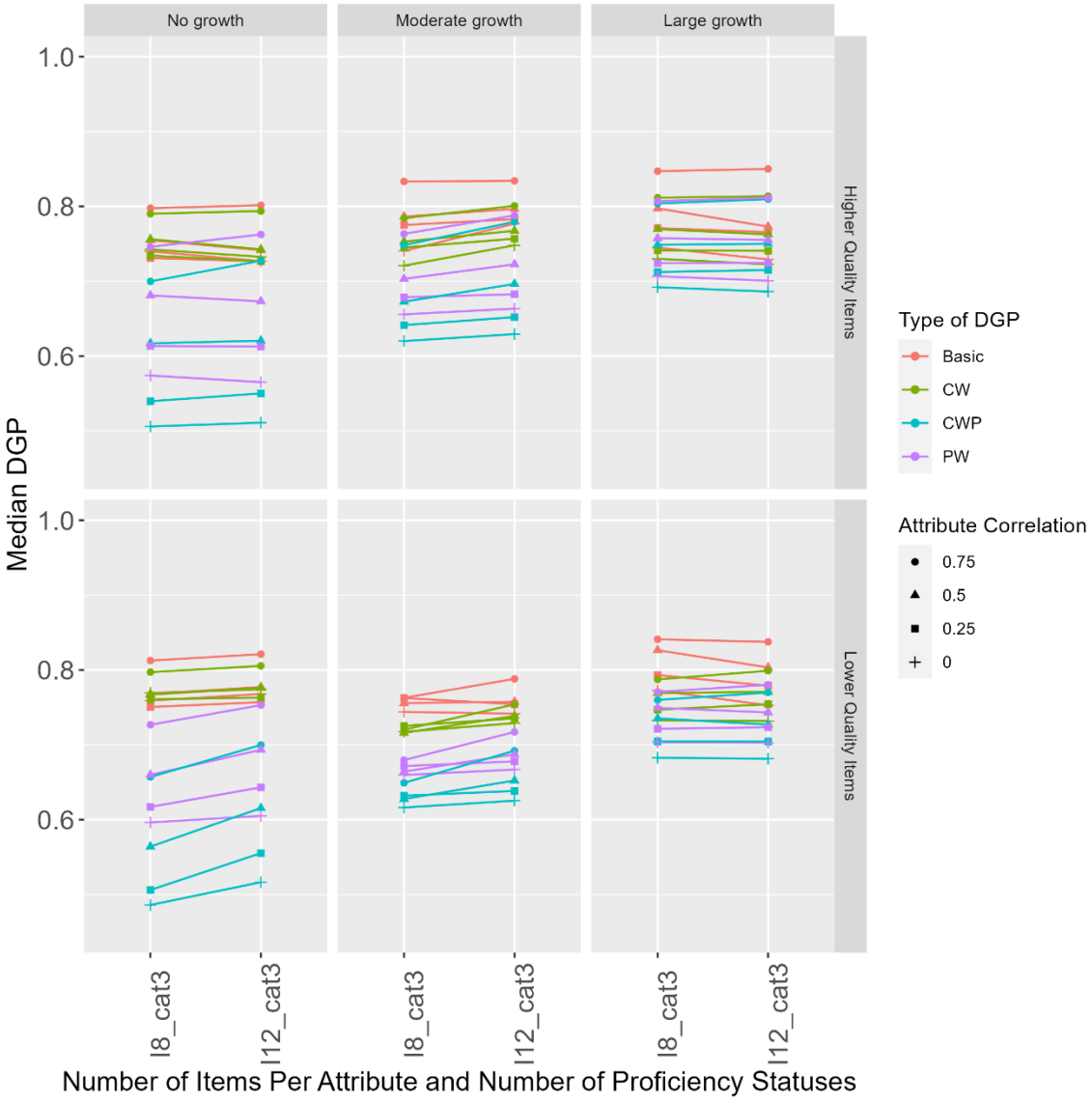
Average Median DGPs for Attribute 3 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 33

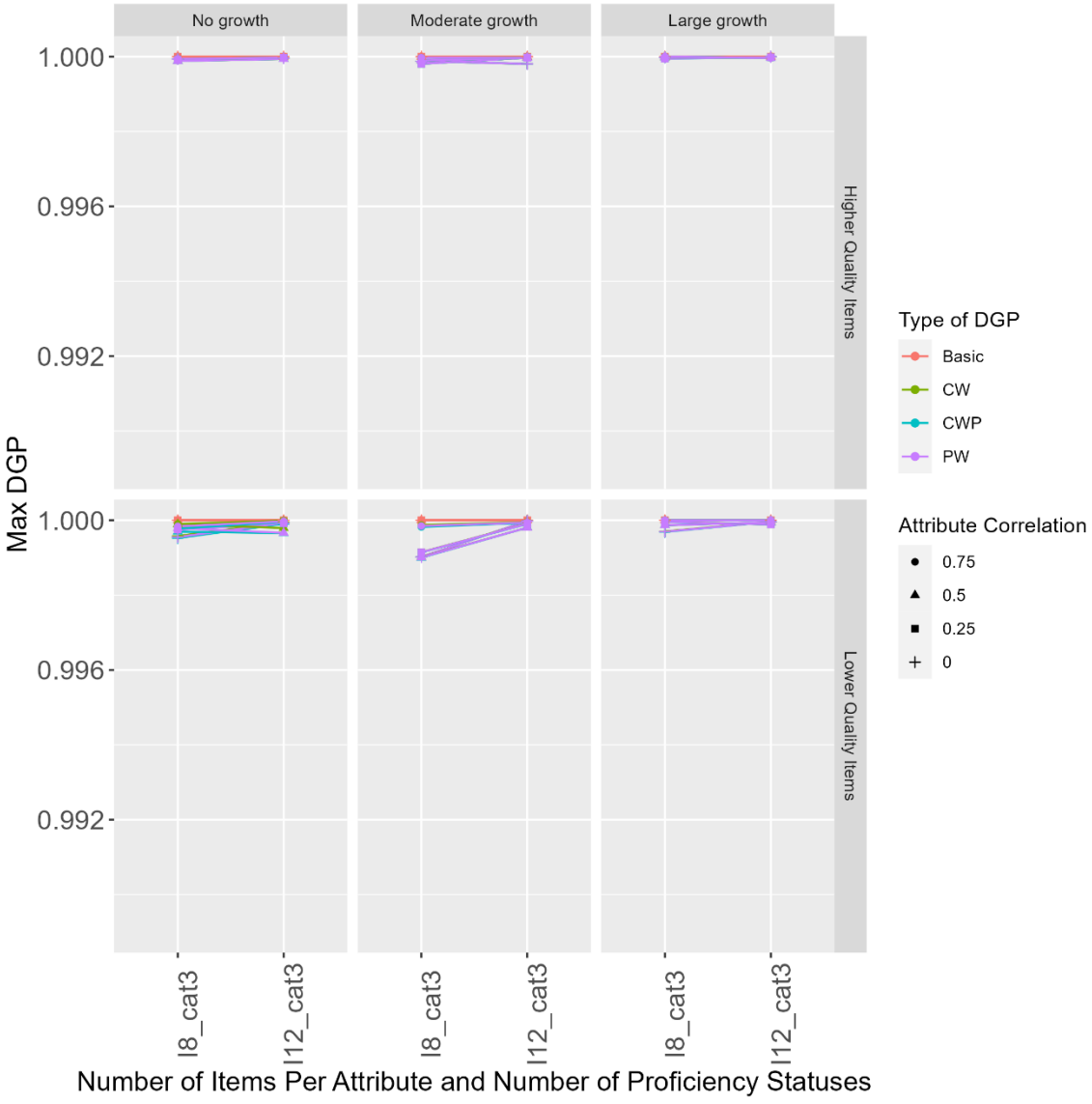
Average Median DGPs for the Profile-Level DGPs in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 34

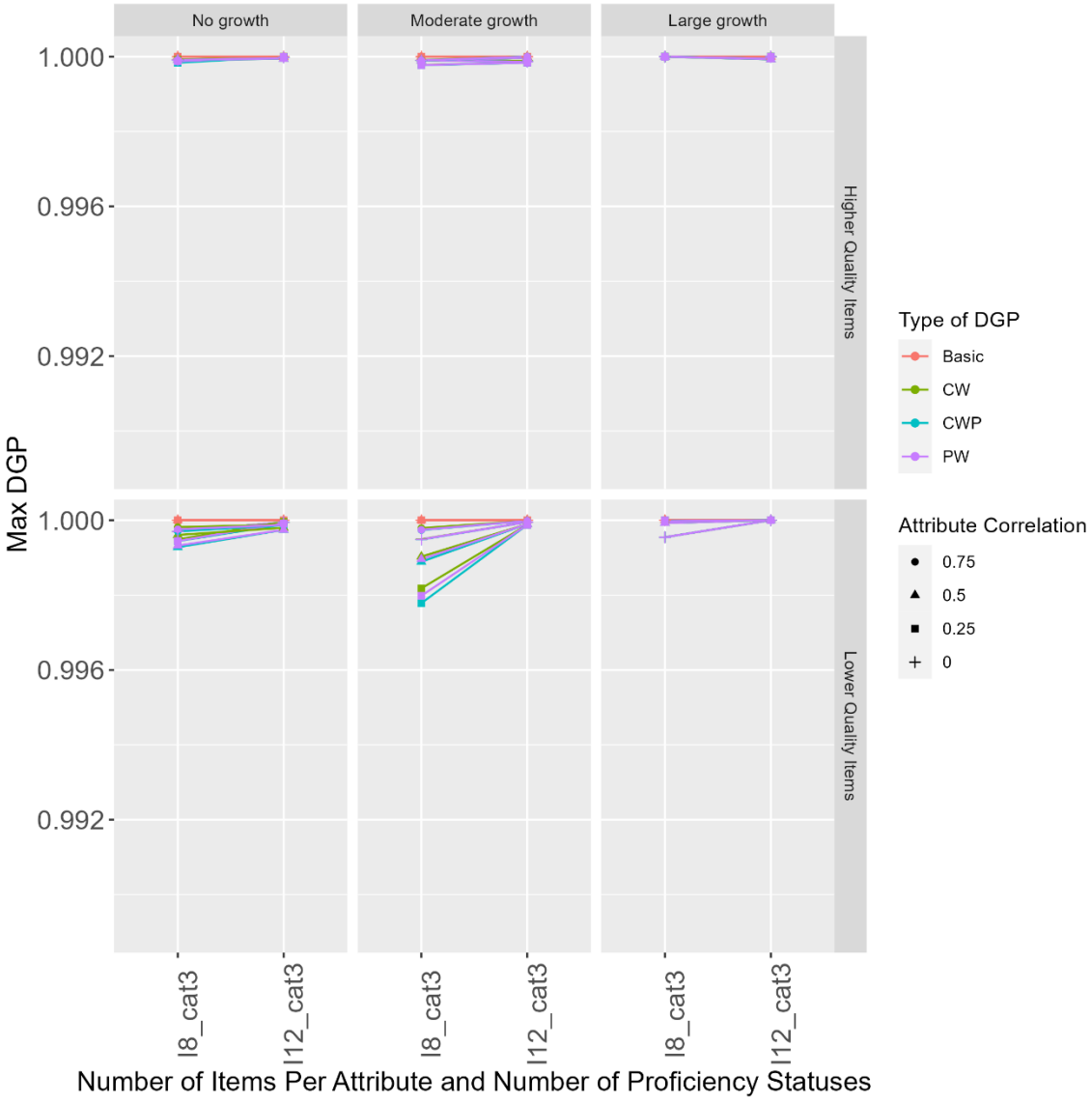
Average Maximum DGPs for Attribute 1 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 35

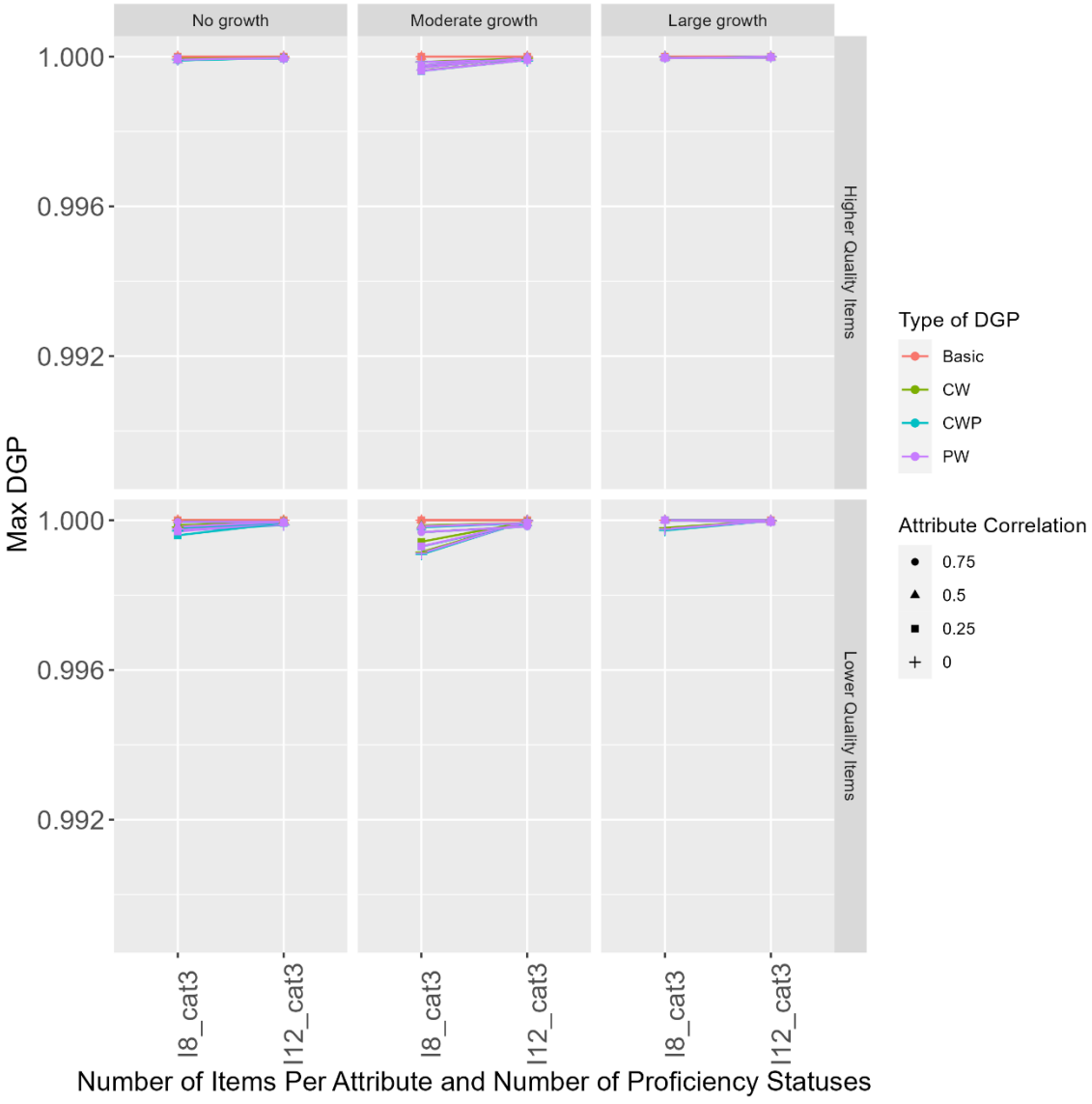
Average Maximum DGPs for Attribute 2 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 36

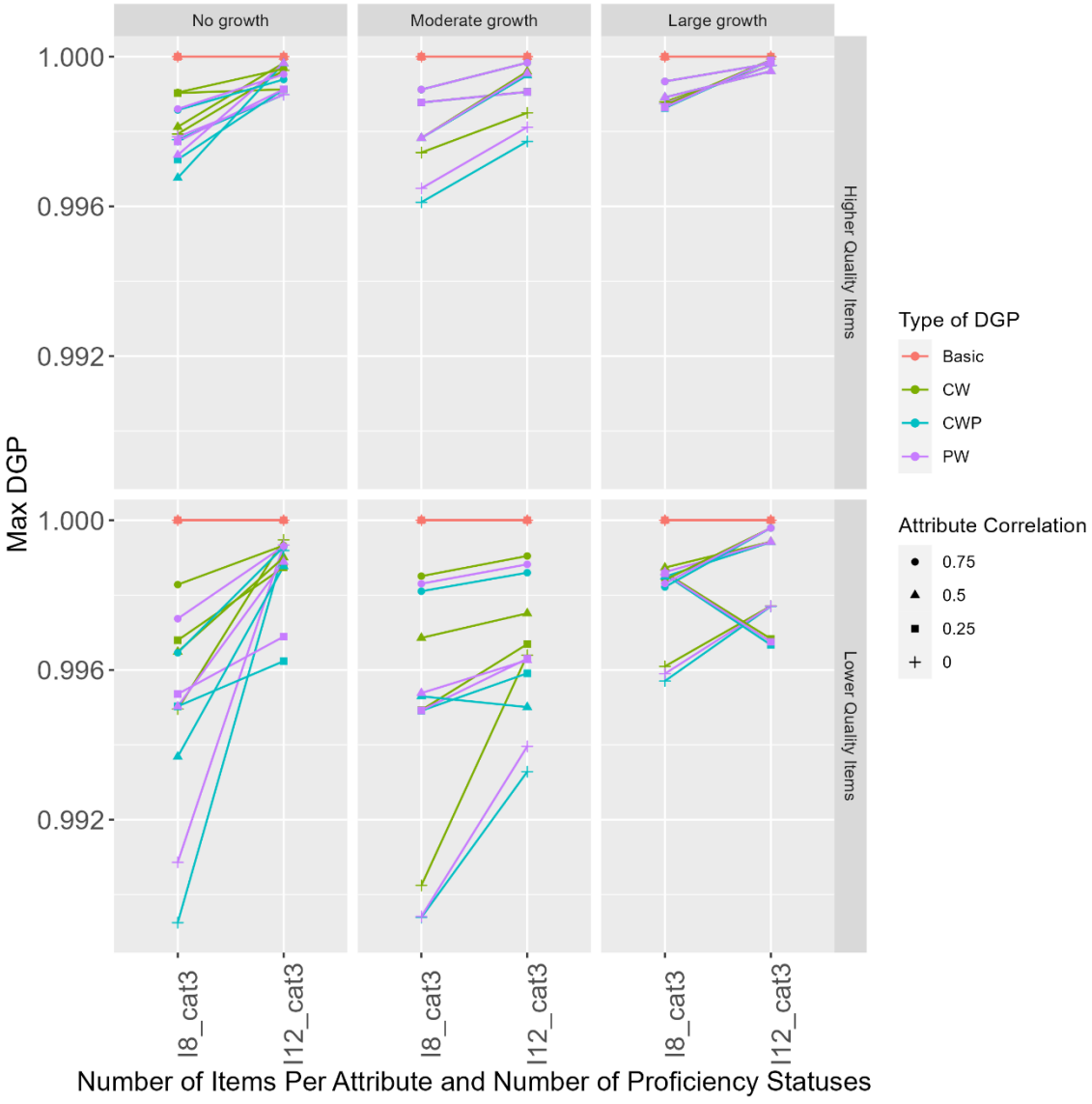
Average Maximum DGPs for Attribute 3 in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 37

Average Maximum DGPs for the Profile-Level DGPs in the Three-Attribute Conditions



Note. "CW" = adjusted DGP with complete weighting. "CWP" = adjusted DGP with complete weighting and a penalty for forgetting. "PW" = adjusted DGP with partial weighting.

Figure 38

Item Response Probability Estimates for the Empirical Data Analysis with Four Two-Category Attributes

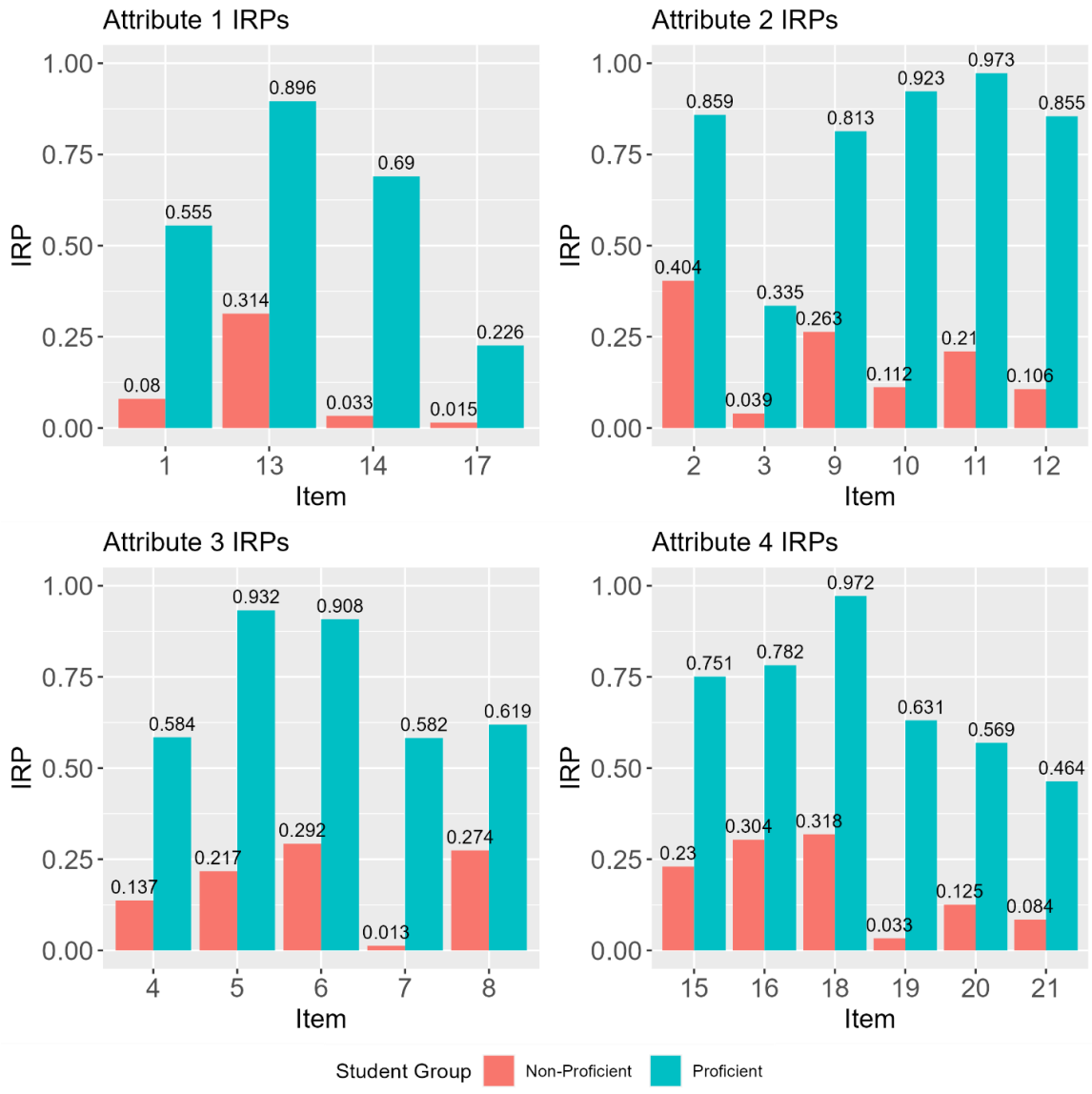
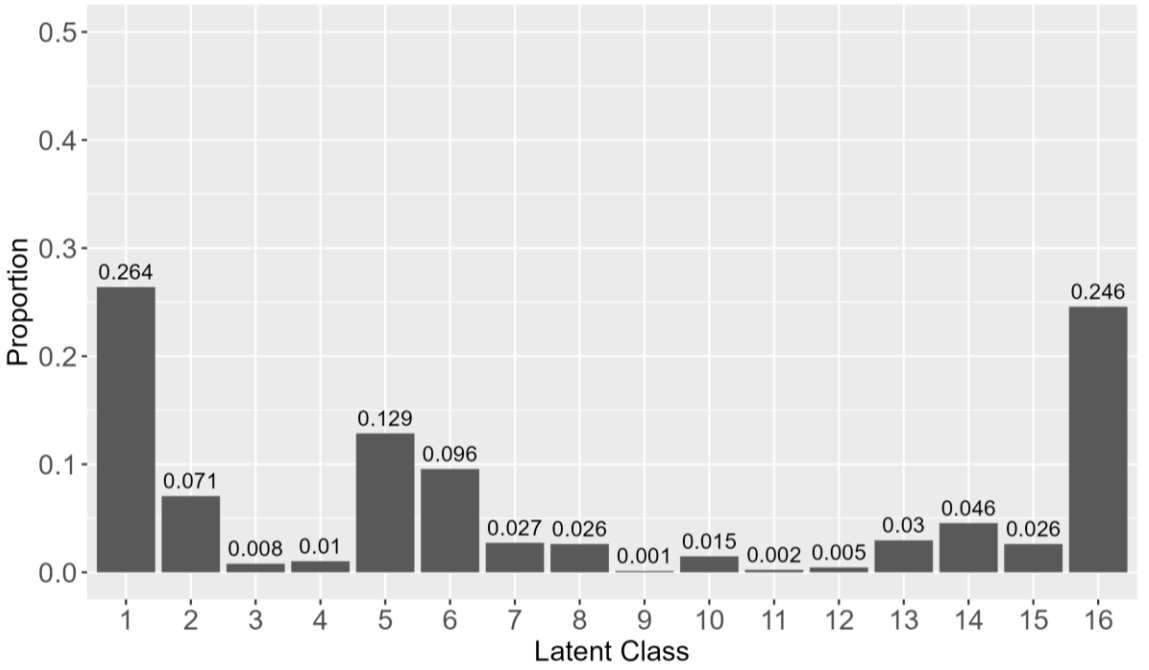


Figure 39

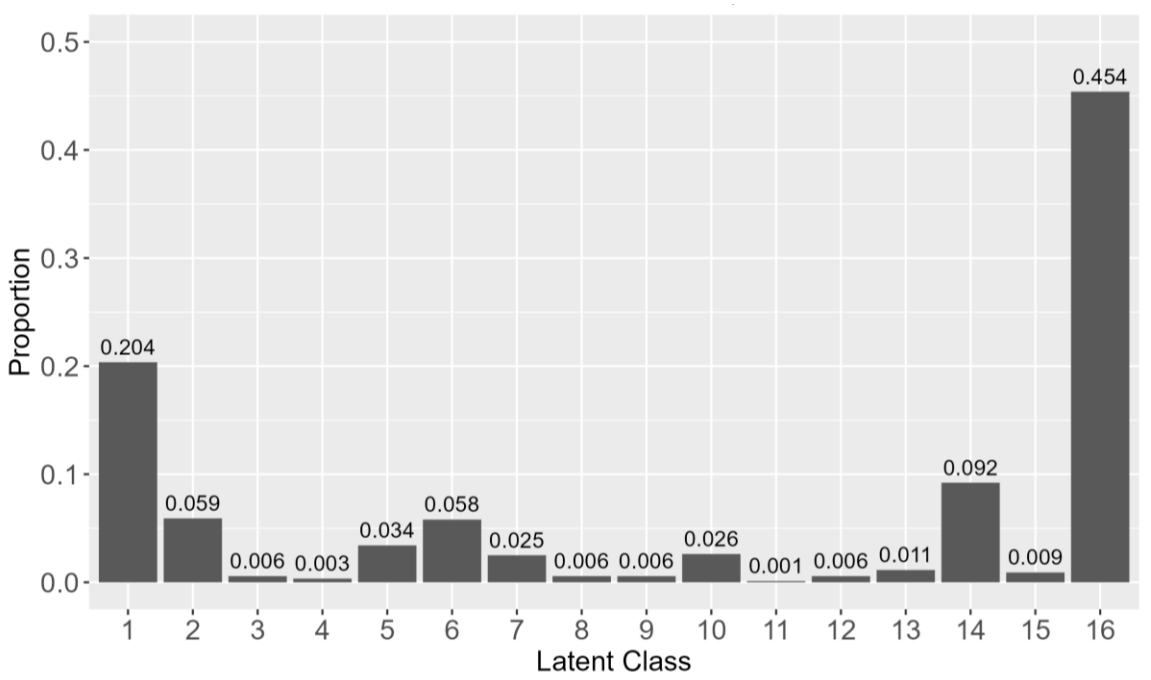
Pre-Test Latent Class Proportions for the Empirical Data Analysis with Four Two-Category Attributes



Note. Latent Class 1 has the attribute profile [0000]. Latent Class 2 = [0001]. Latent Class 3 = [0010]. Latent Class 4 = [0011]. Latent Class 5 = [0100]. Latent Class 6 = [0101]. Latent Class 7 = [0110]. Latent Class 8 = [0111]. Latent Class 9 = [1000]. Latent Class 10 = [1001]. Latent Class 11 = [1010]. Latent Class 12 = [1011]. Latent Class 13 [1100]. Latent Class 14 = [1101]. Latent Class 15 = [1110]. Latent Class 16 = [1111].

Figure 40

Post-Test Latent Class Proportions for the Empirical Data Analysis with Four Two-Category Attributes



Note. Latent Class 1 has the attribute profile [0000]. Latent Class 2 = [0001]. Latent Class 3 = [0010]. Latent Class 4 = [0011]. Latent Class 5 = [0100]. Latent Class 6 = [0101]. Latent Class 7 = [0110]. Latent Class 8 = [0111]. Latent Class 9 = [1000]. Latent Class 10 = [1001]. Latent Class 11 = [1010]. Latent Class 12 = [1011]. Latent Class 13 [1100]. Latent Class 14 = [1101]. Latent Class 15 = [1110]. Latent Class 16 = [1111].

Figure 41

Base Rates for the Empirical Data Analysis with Four Two-Category Attributes

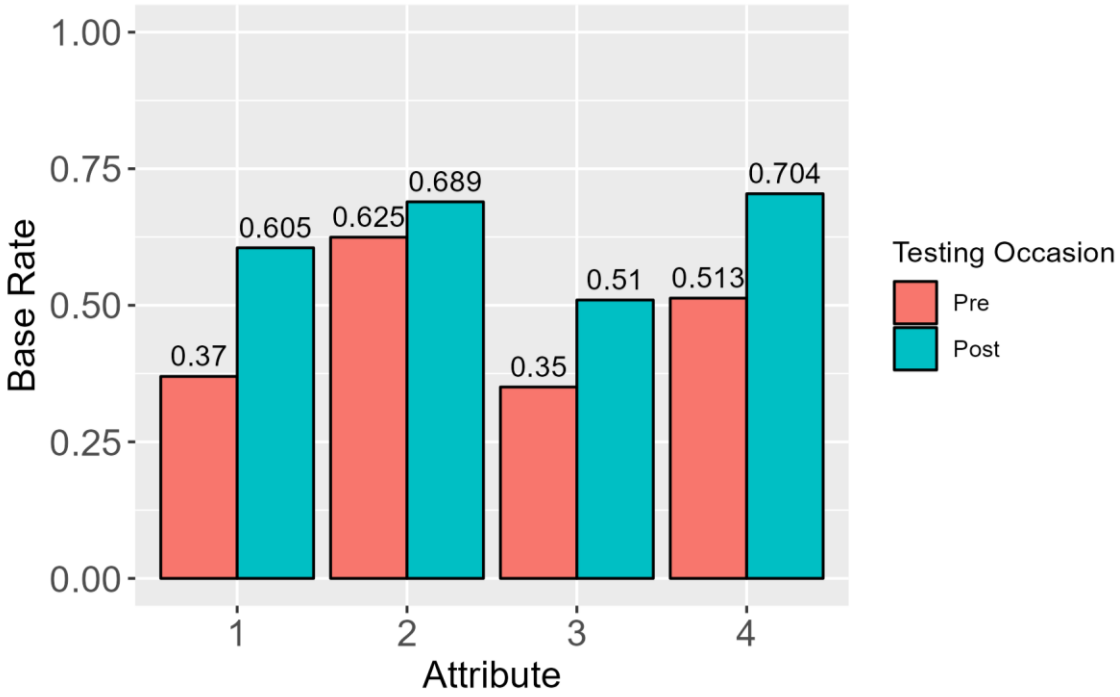
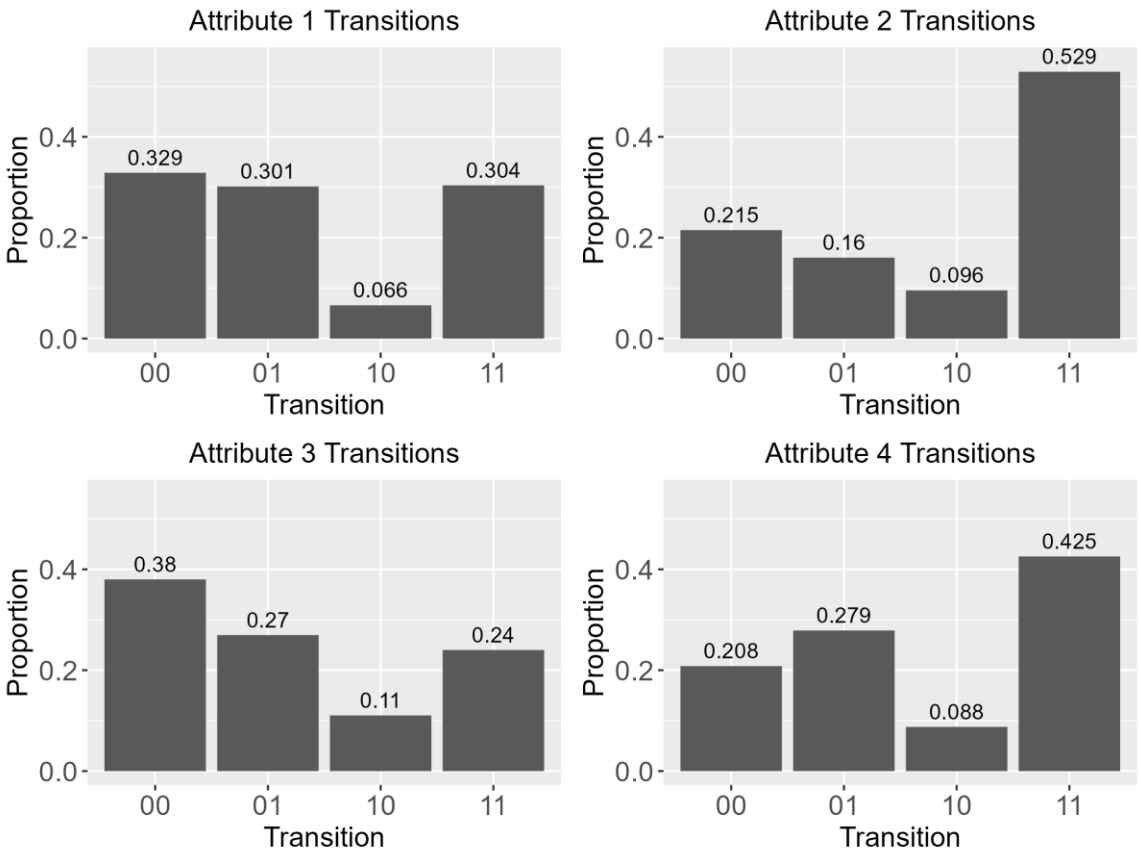


Figure 42

Attribute Transitions for the Empirical Data Analysis with Four Two-Category Attributes



Note. In the transitions on the horizontal axis, "0" indicates the "Non-Proficient" proficiency status and "1" indicates the "Proficient" proficiency status.

Figure 43

PTDCM Reliability for the Empirical Data Analysis with Four Two-Category Attributes

Note. PB = point biserial reliability metric; PBW = weighted point biserial reliability metric; PF = parallel forms reliability metric; PFW = weighted parallel forms reliability metric; IG = information gain reliability metric; IGW = weighted information gain reliability metric; Polychoric = polychoric reliability metric; AvgMax = average maximum transition reliability metric

Figure 44

Analysis of Average Maximum Posterior Probability for the Empirical Data Analysis with Four Two-Category Attributes

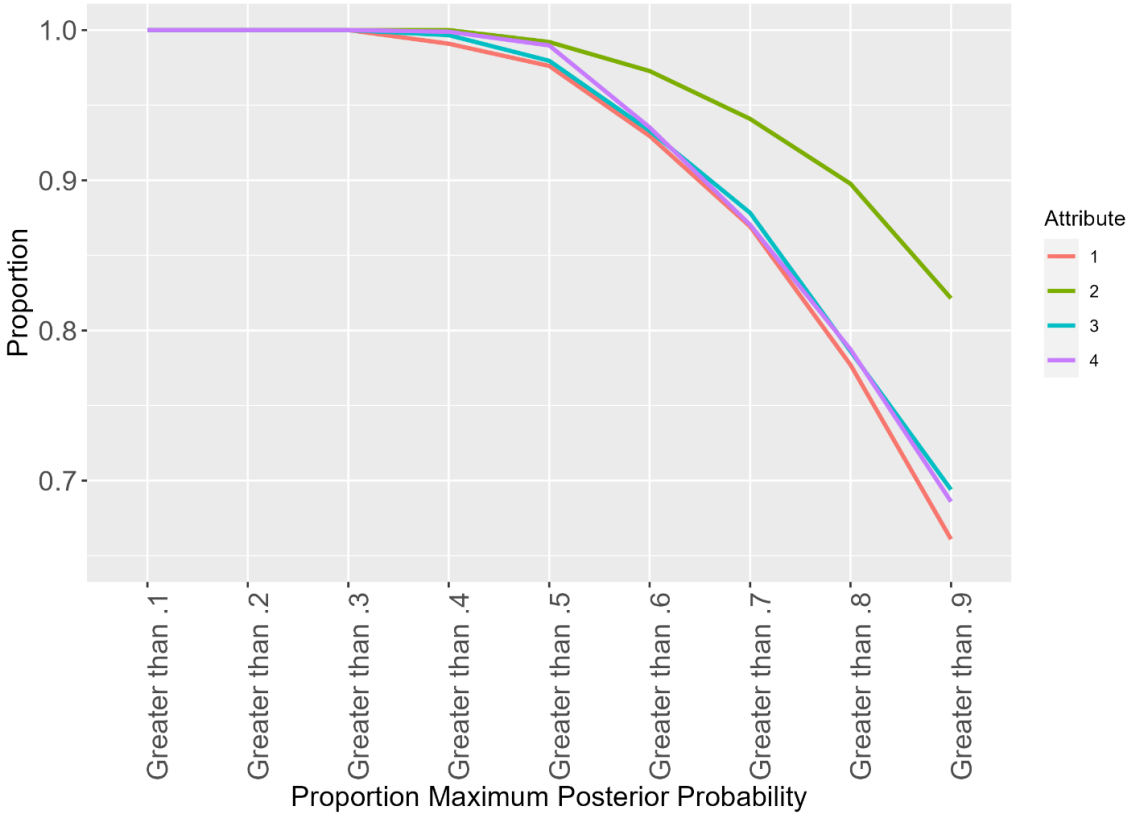


Figure 45

Basic DGPs for the Empirical Data Analysis with Four Two-Category Attributes

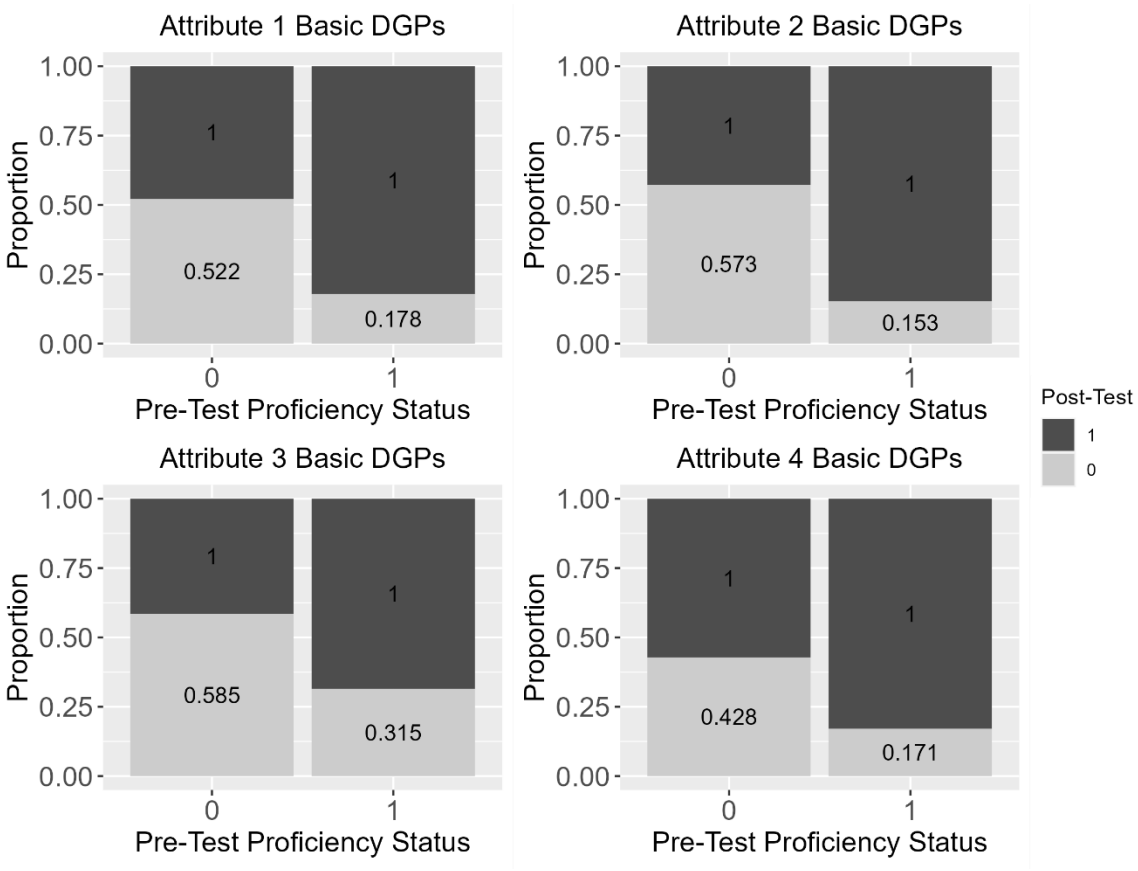
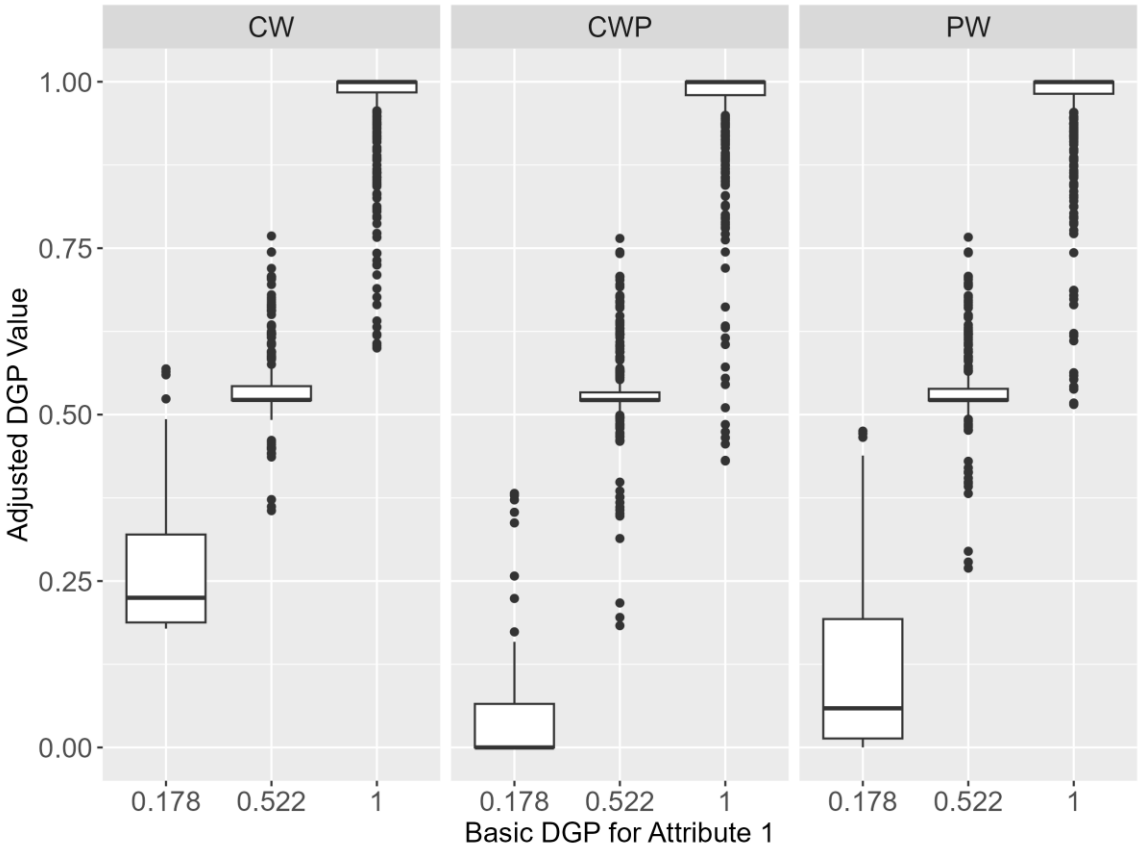


Figure 46

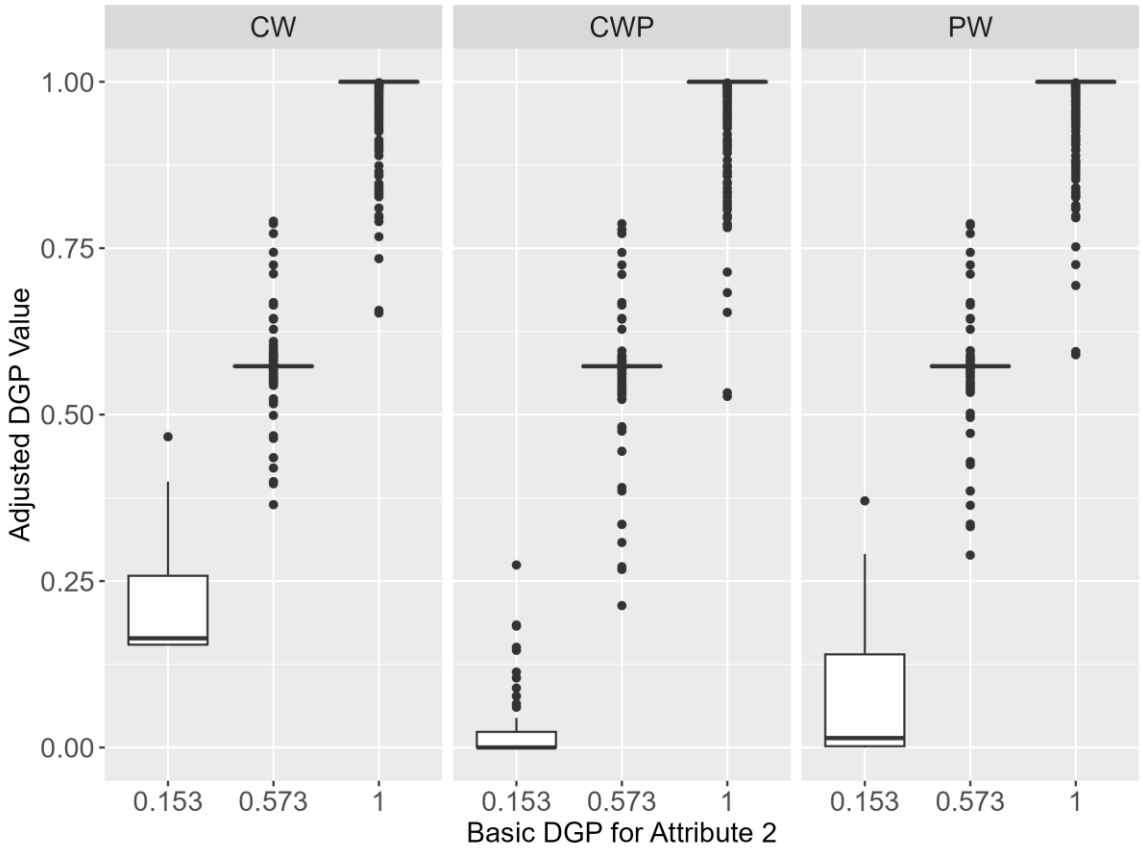
Adjusted DGP Boxplot for Attribute 1 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 47

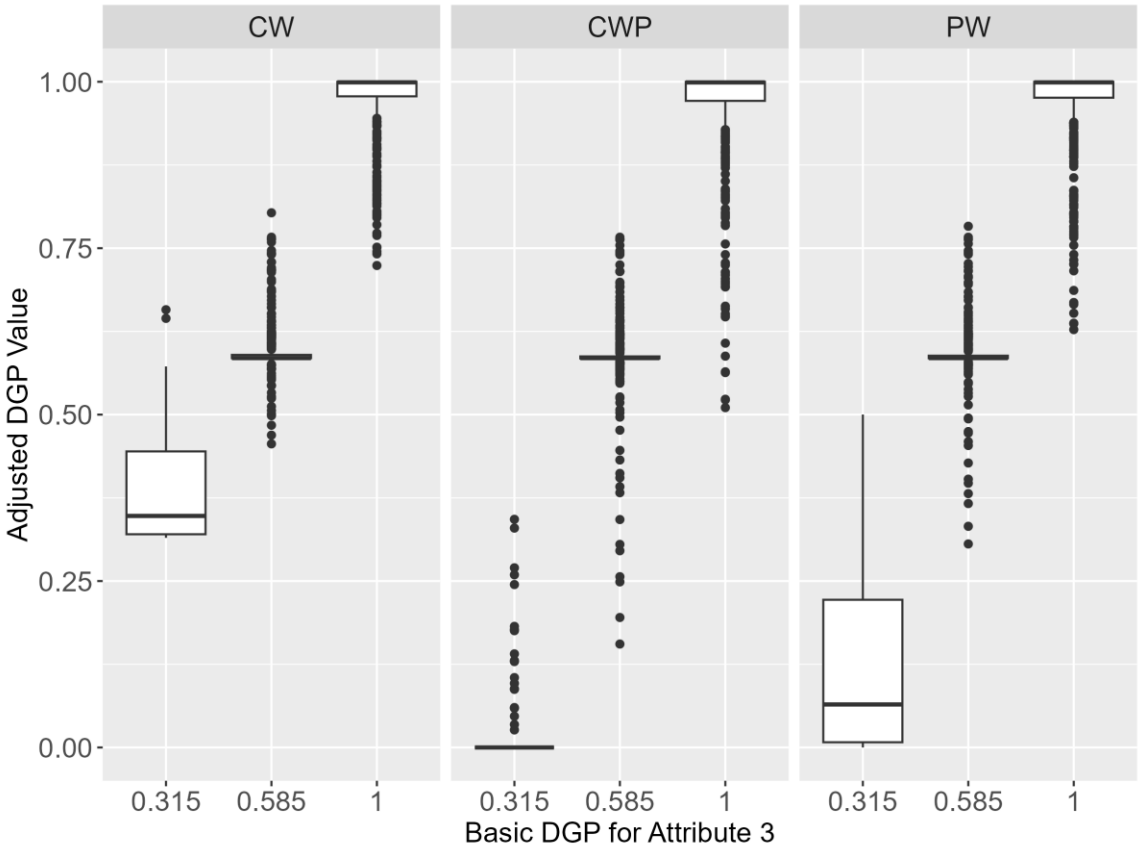
Adjusted DGP Boxplot for Attribute 2 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 48

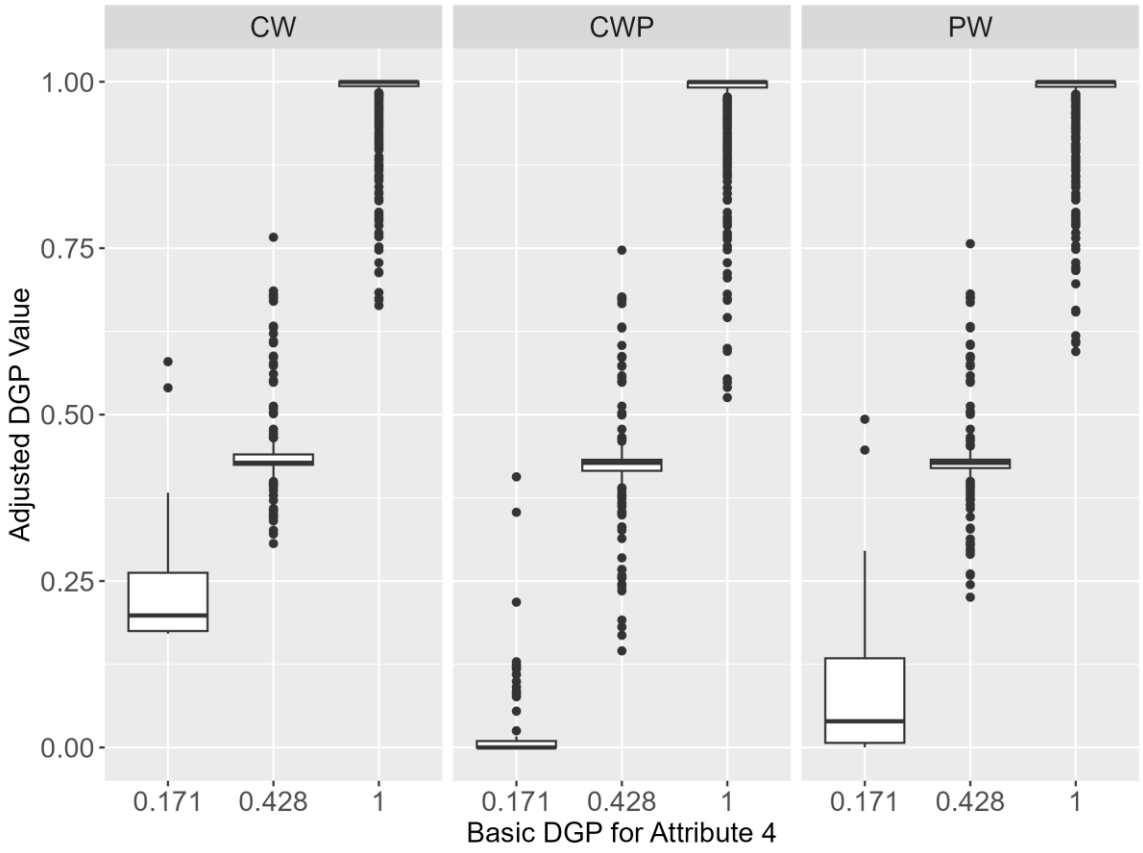
Adjusted DGP Boxplot for Attribute 3 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 49

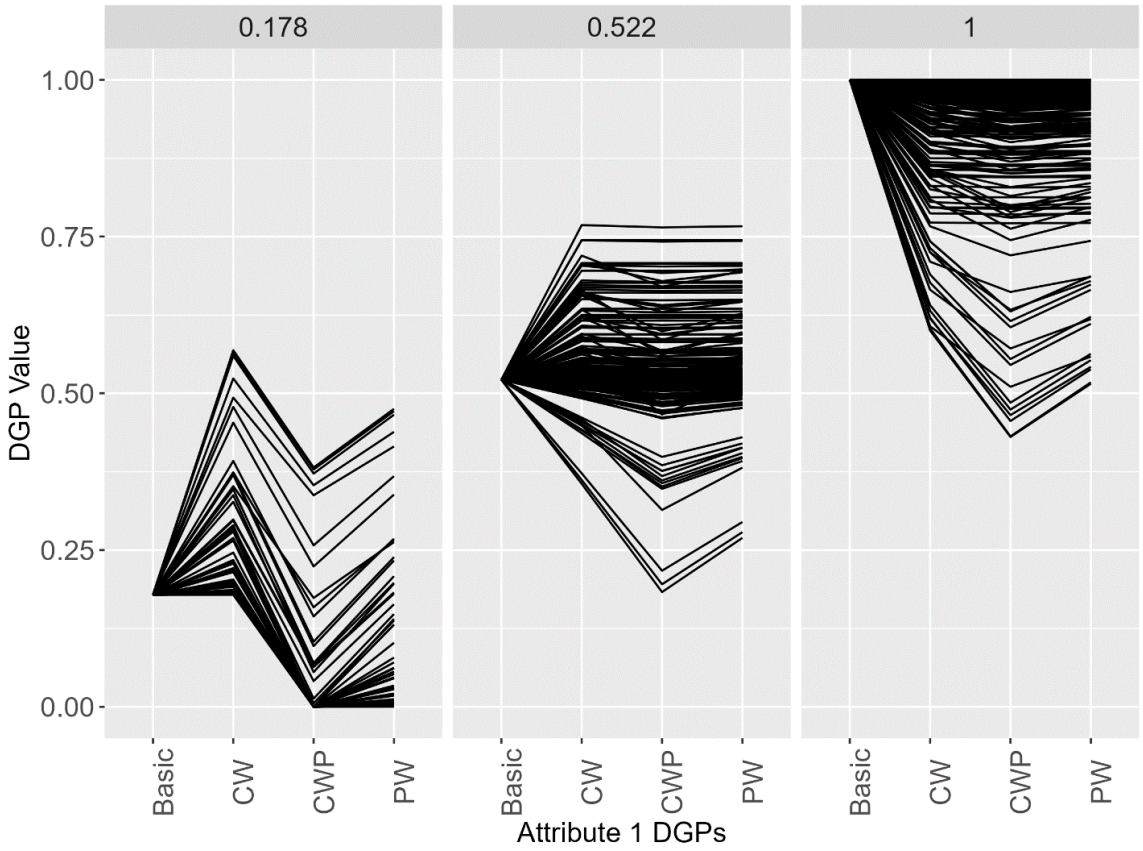
Adjusted DGP Boxplot for Attribute 4 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 50

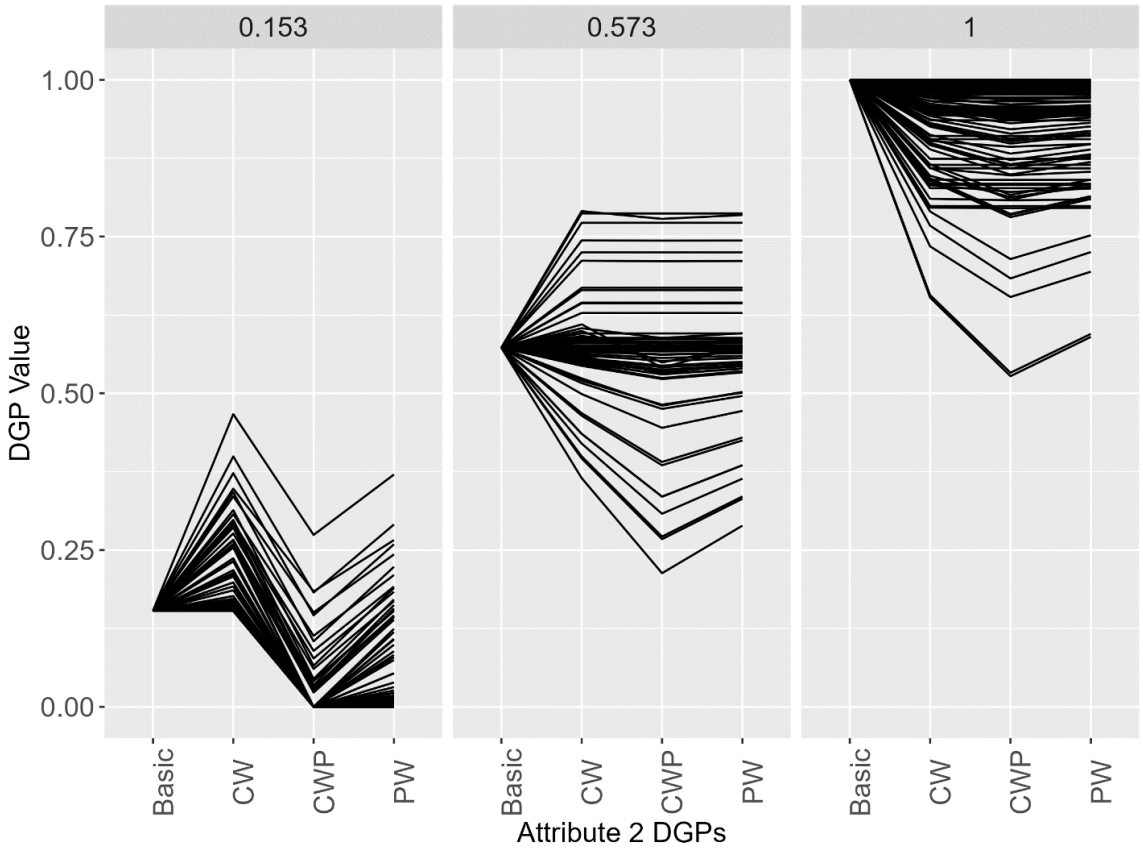
Student's DGP Plots Split by Basic DGP for Attribute 1 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 51

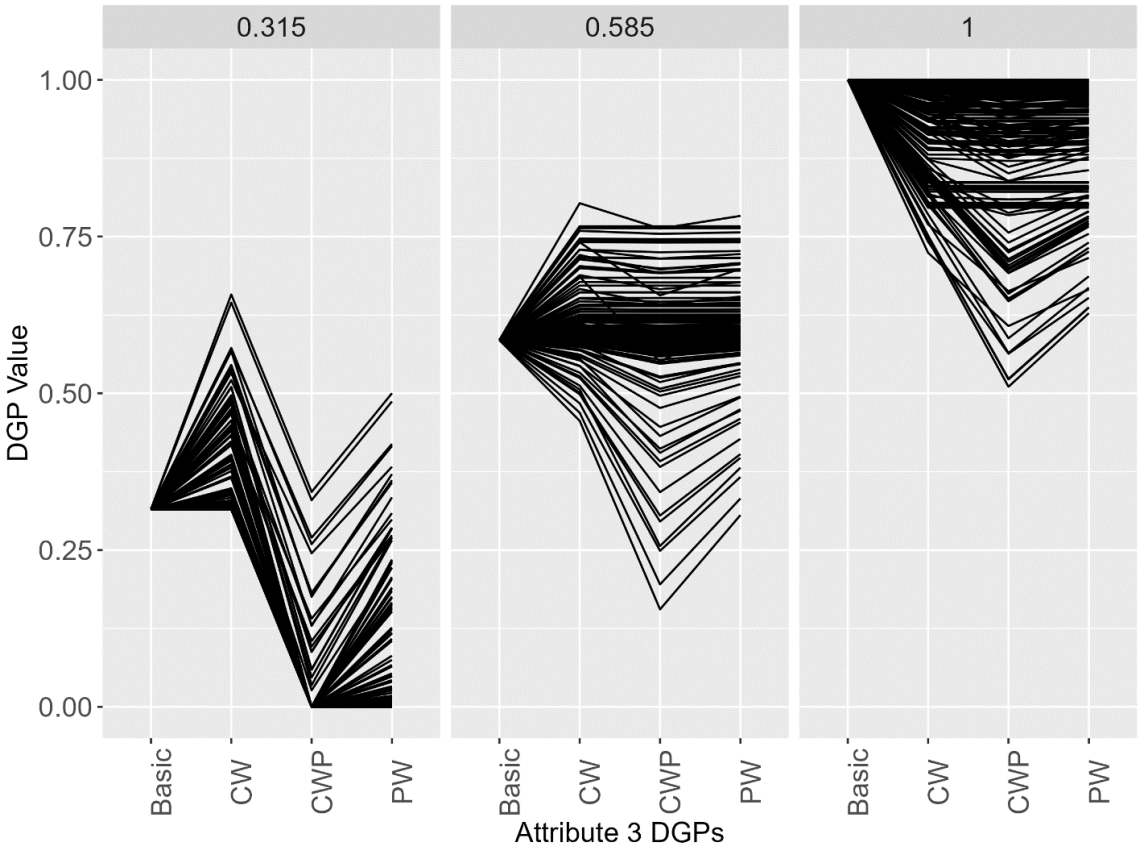
Student's DGP Plots Split by Basic DGP for Attribute 2 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 52

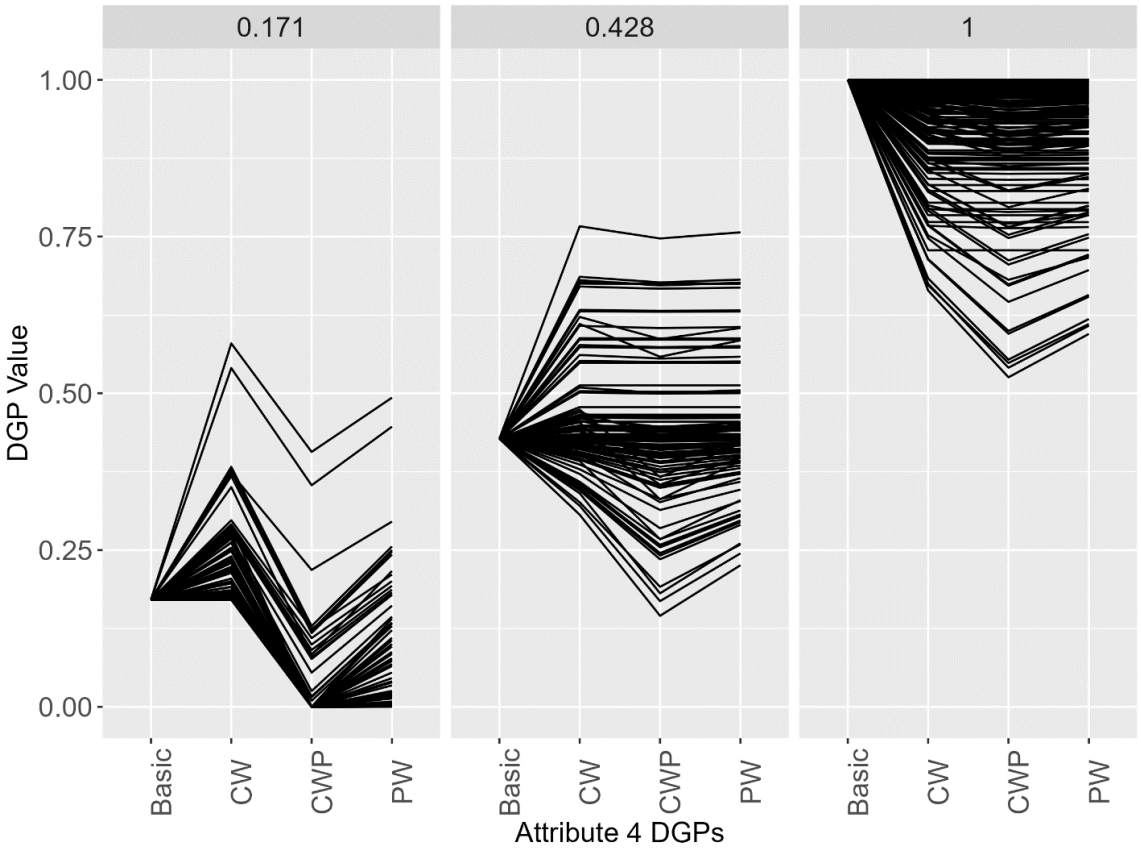
Student's DGP Plots Split by Basic DGP for Attribute 3 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 53

Student's DGP Plots Split by Basic DGP for Attribute 4 in the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 54

Attribute 1 DGPs for Eight Example Students in the Empirical Data Analysis with Four Two-Category Attributes

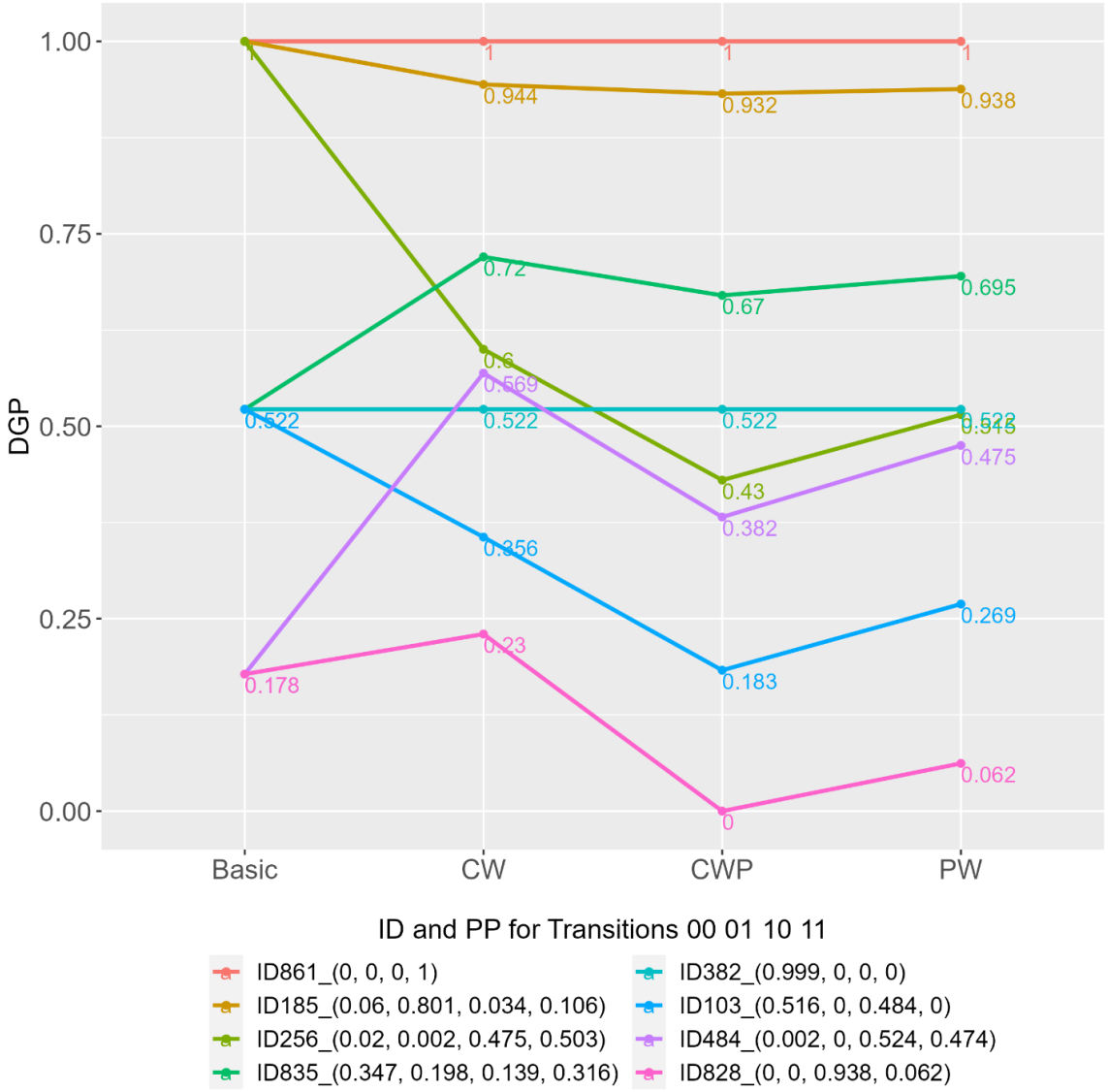
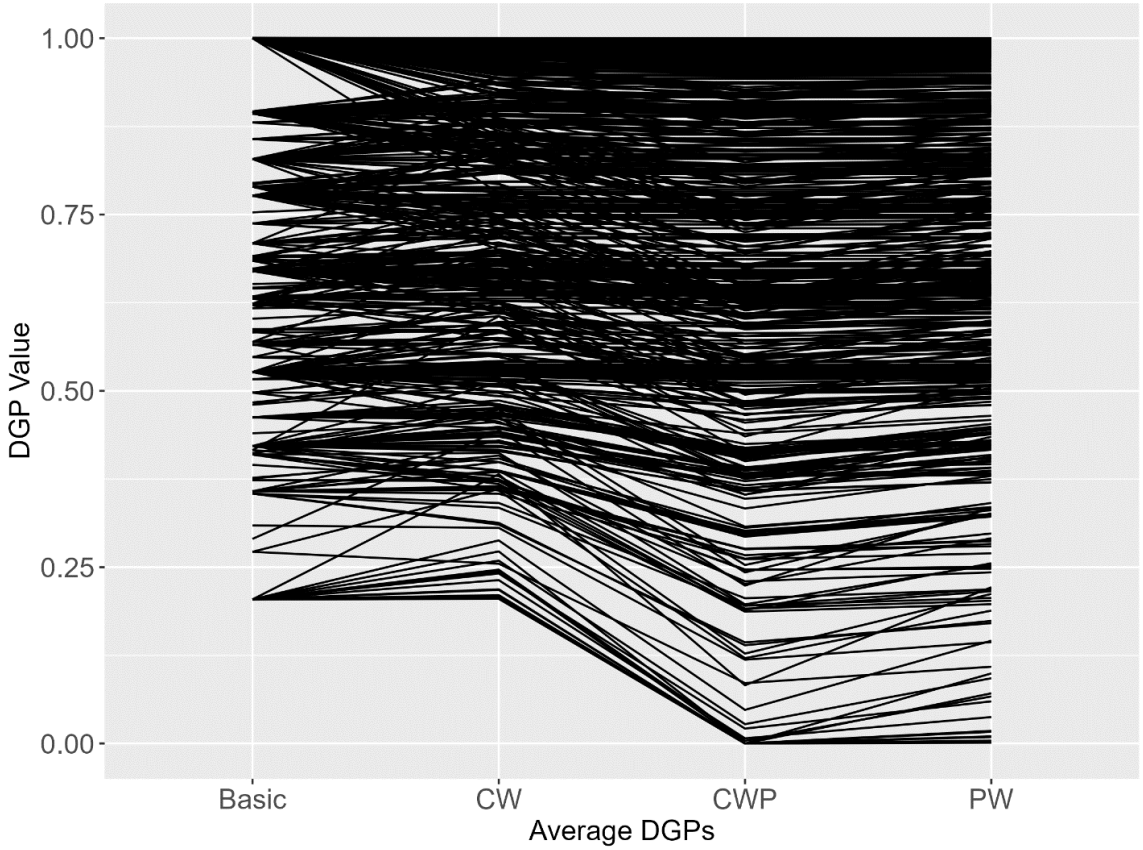


Figure 55

Students' Average DGPs for the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting

Figure 56

DGPs Aggregated Across Students for the Empirical Data Analysis with Four Two-Category Attributes

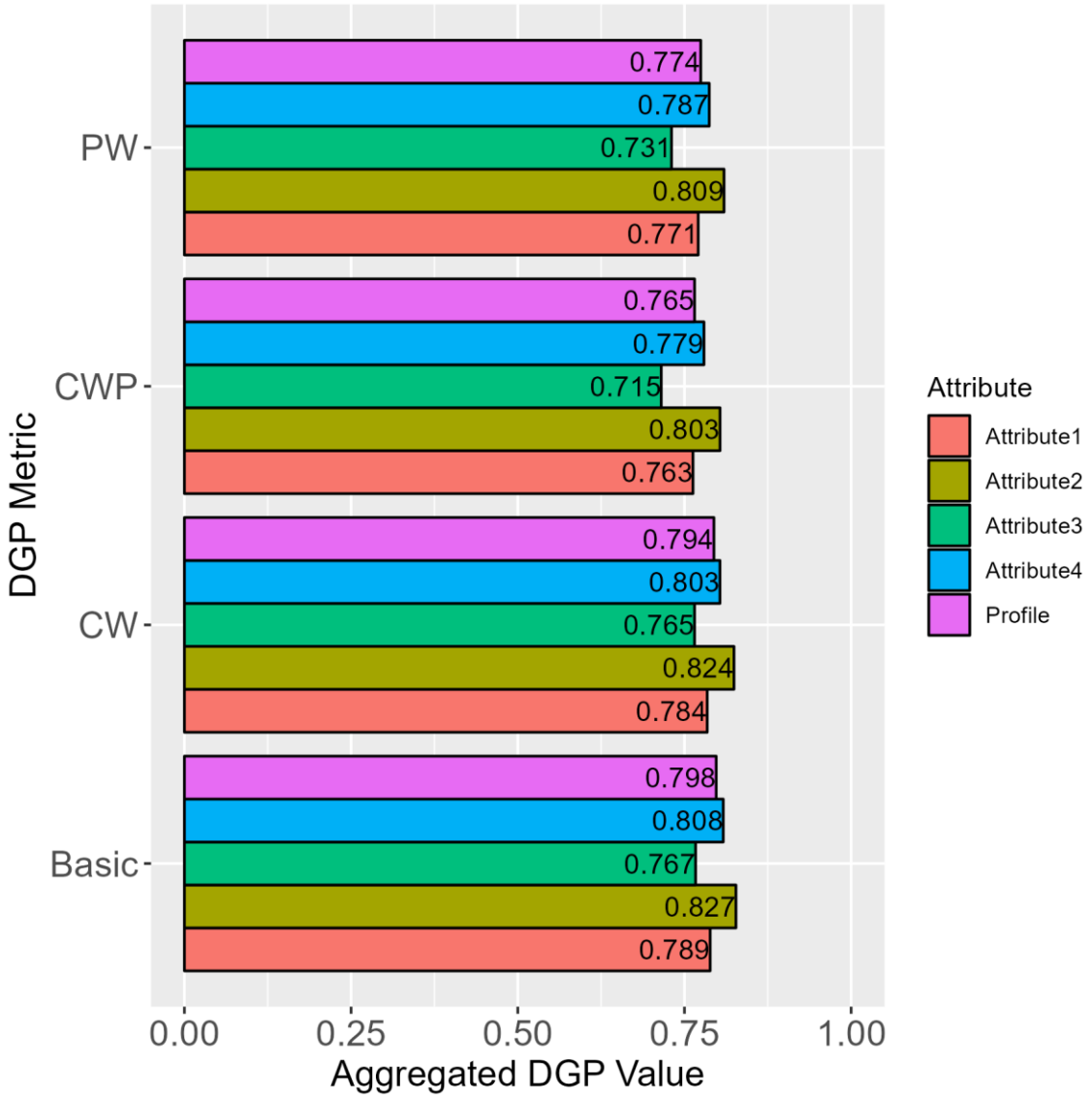
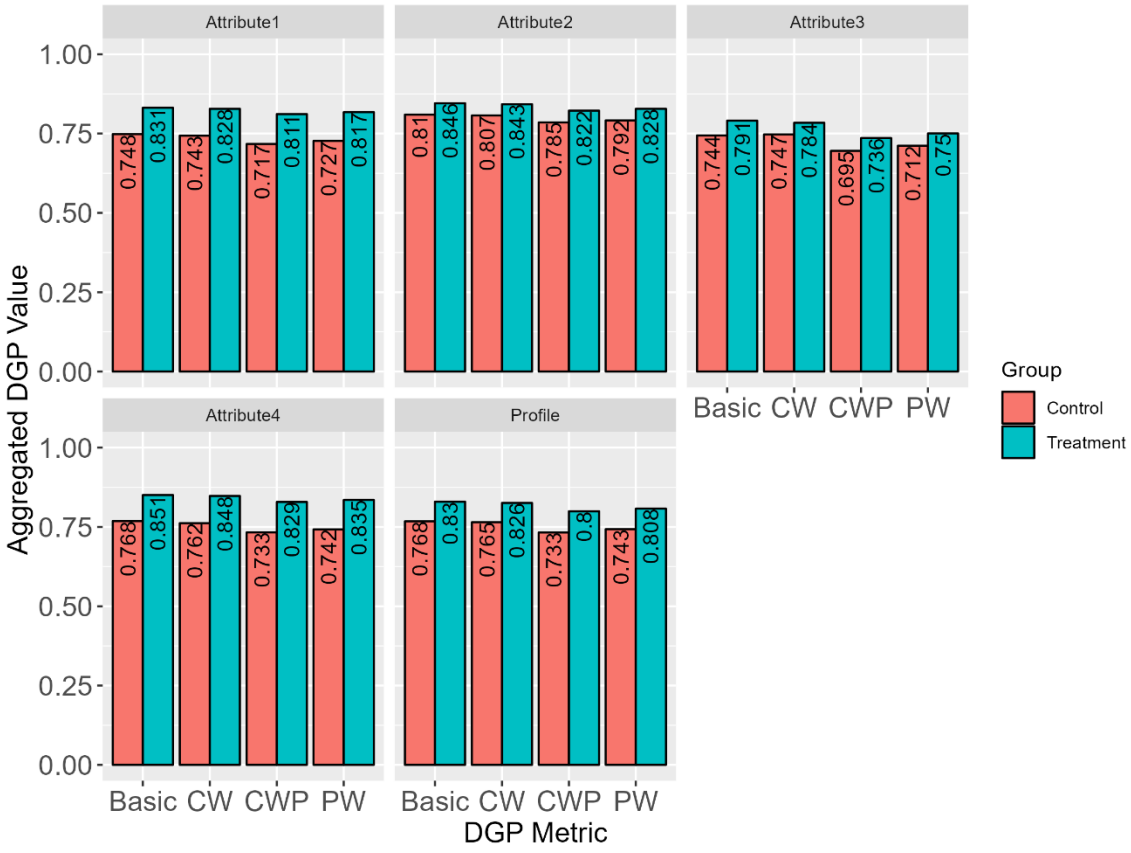


Figure 57

DGPs Aggregated Across Treatment Groups for the Empirical Data Analysis with Four Two-Category Attributes



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting.

Figure 58

Item Response Probability Estimates for the Empirical Data Analysis with One Three-Category Attribute

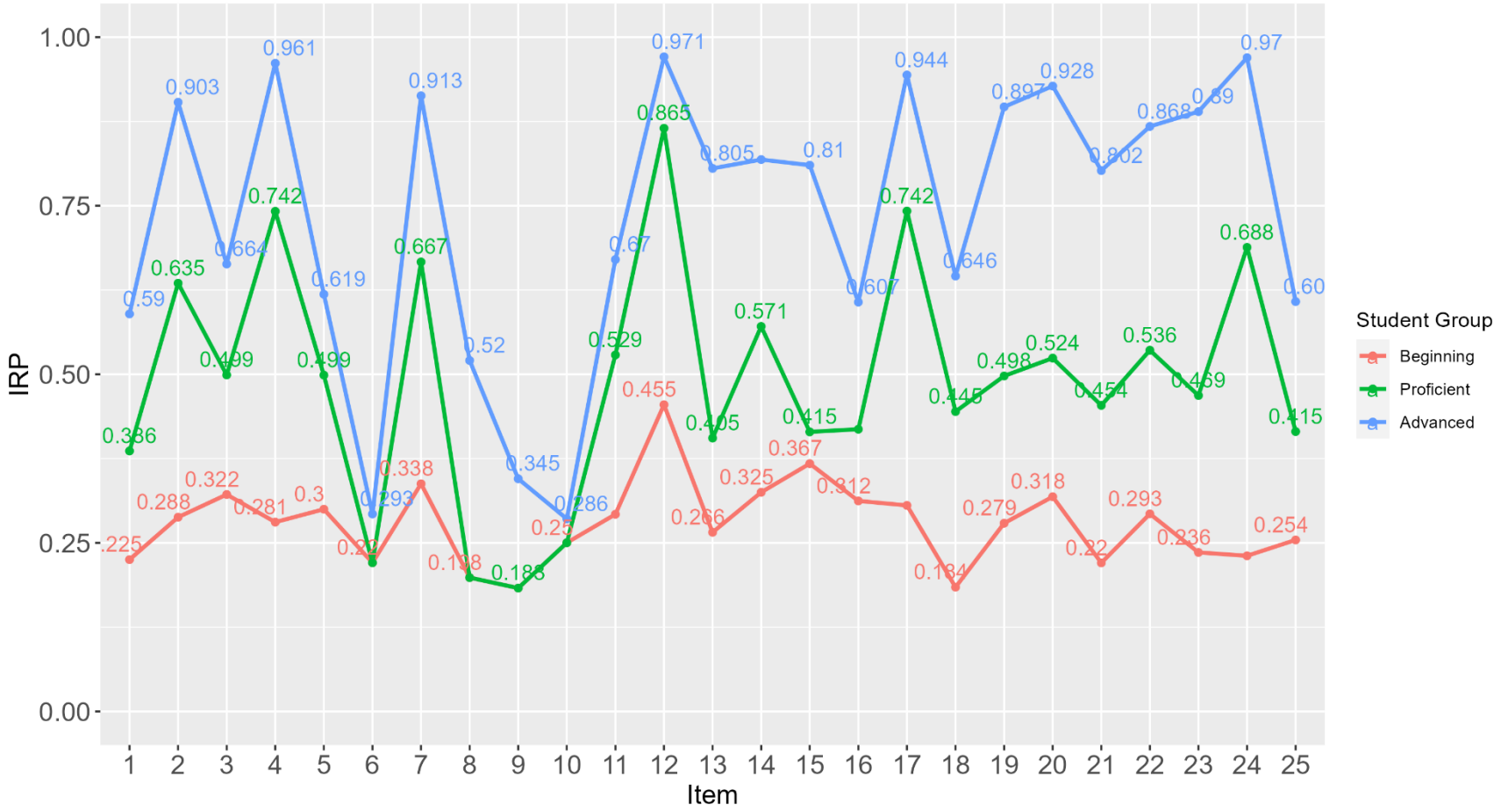
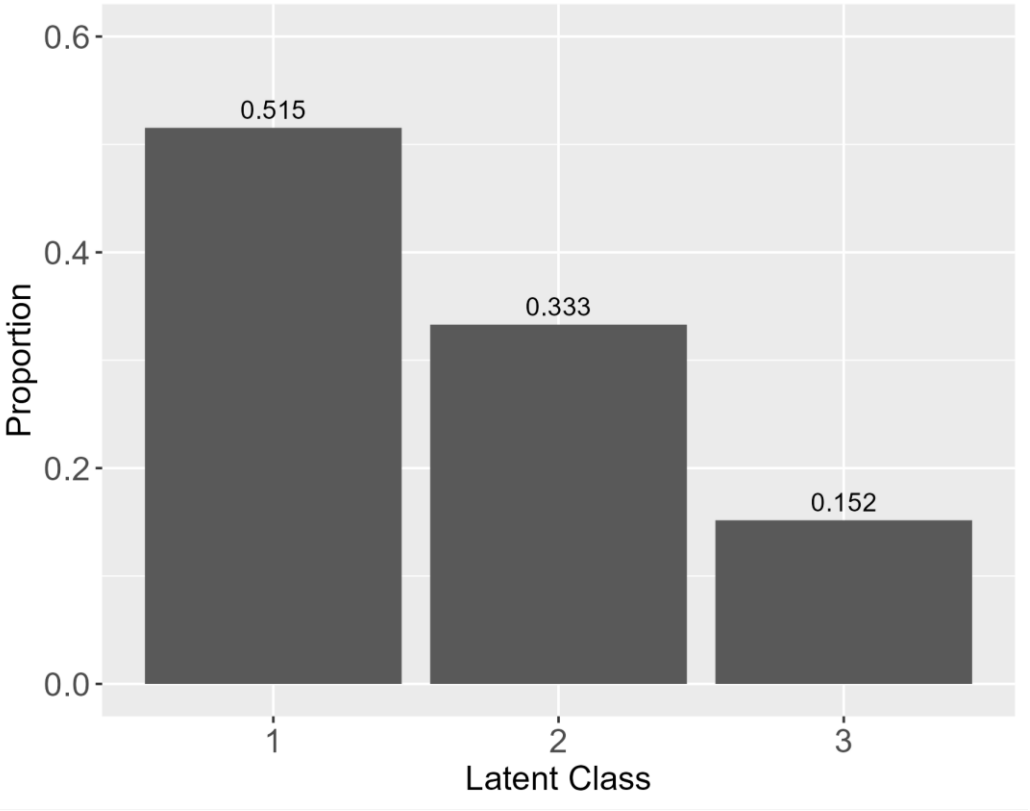


Figure 59

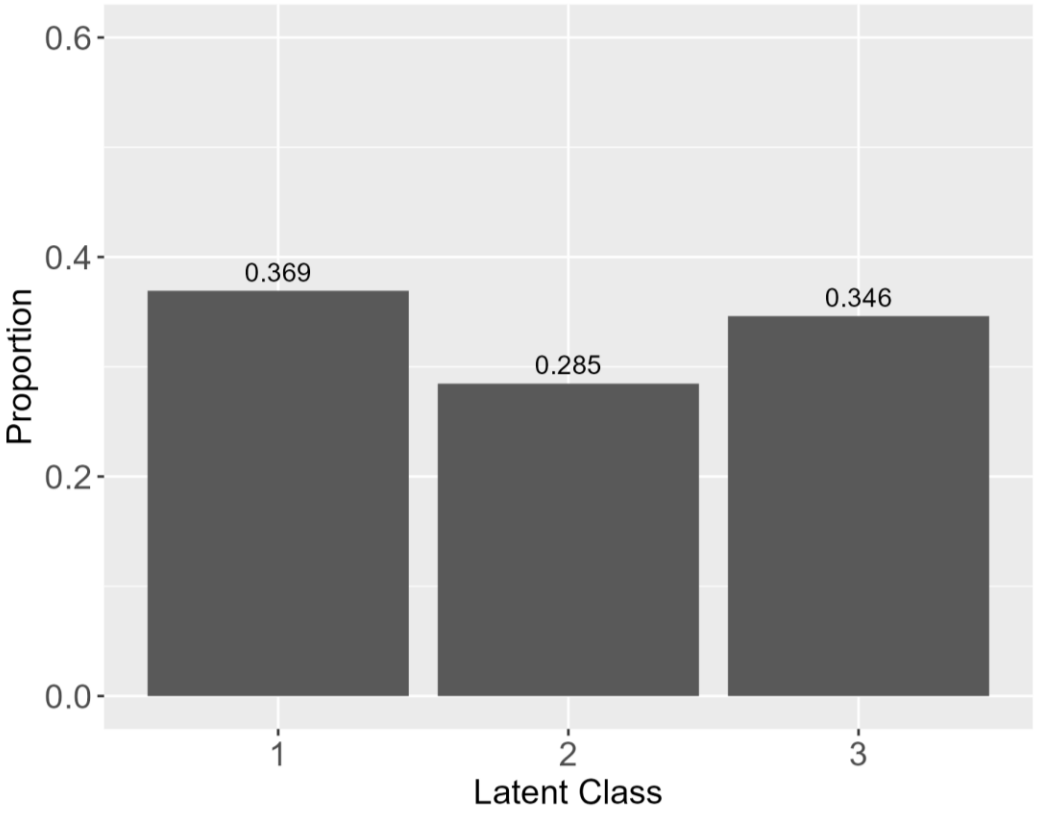
Pre-Test Latent Class Proportions for the Empirical Data Analysis with One Three-Category Attribute



Note. Latent Class 1 has the attribute profile [0] and the label “Beginning”. Latent Class 2 has the attribute profile [1] and the label “Proficient”. Latent Class 3 has the attribute profile [2] and the label “Advanced”.

Figure 60

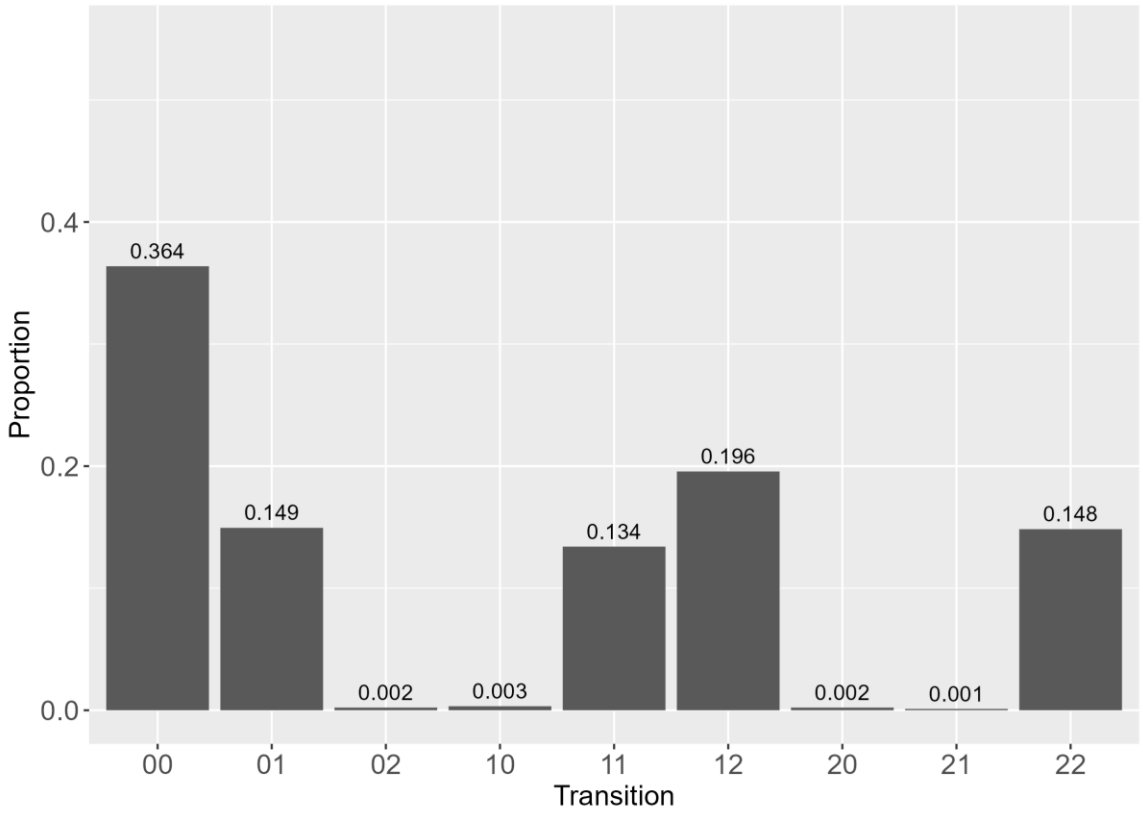
Post-Test Latent Class Proportions for the Empirical Data Analysis with One Three-Category Attribute



Note. Latent Class 1 has the attribute profile [0] and the label “Beginning”. Latent Class 2 has the attribute profile [1] and the label “Proficient”. Latent Class 3 has the attribute profile [2] and the label “Advanced”.

Figure 61

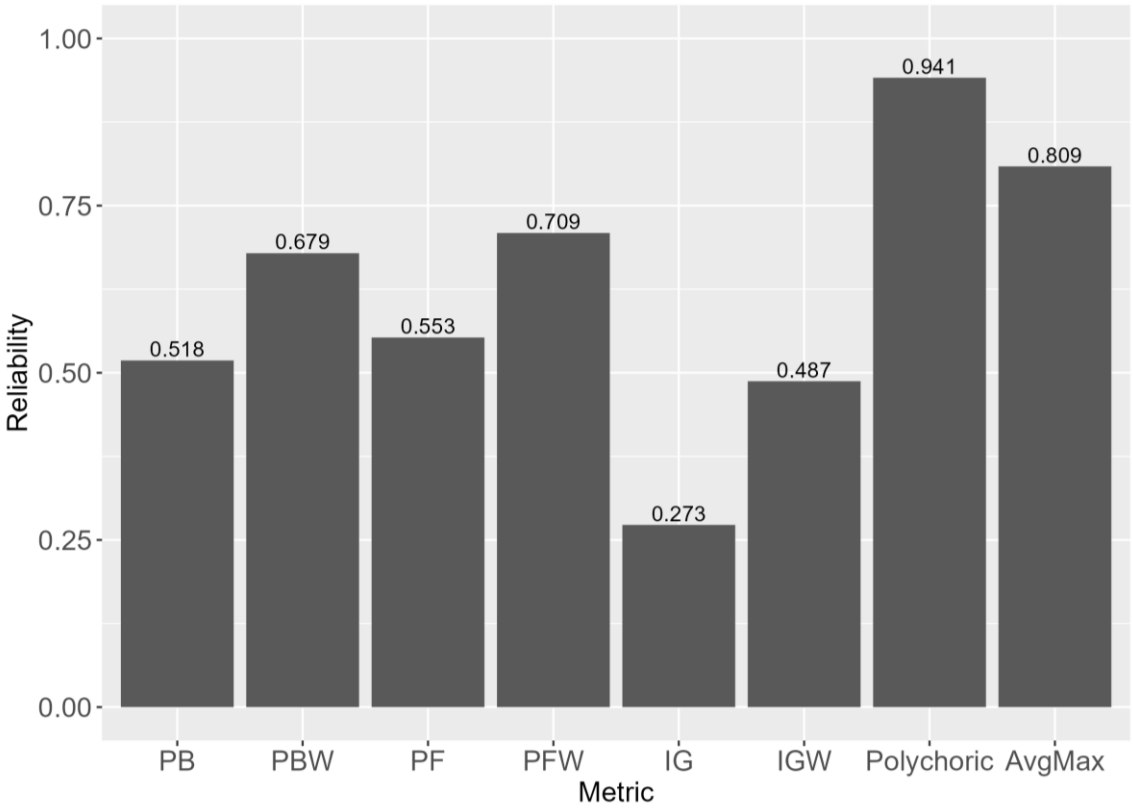
Attribute Transitions for the Empirical Data Analysis with One Three-Category Attribute



Note. In the transitions on the horizontal axis, "0" indicates the "Beginning" proficiency status, "1" indicates the "Proficient" proficiency status, and "2" indicates the "Advanced" proficiency status.

Figure 62

PTDCM Reliability for Each Attribute for the Empirical Data Analysis with One Three-Category Attribute



Note. PB = point biserial reliability metric; PBW = weighted point biserial reliability metric; PF = parallel forms reliability metric; PFW = weighted parallel forms reliability metric; IG = information gain reliability metric; IGW = weighted information gain reliability metric; Polychoric = polychoric reliability metric; AvgMax = average maximum transition reliability metric

Figure 63

Analysis of Average Maximum Posterior Probability for the Empirical Data Analysis with One Three-Category Attribute

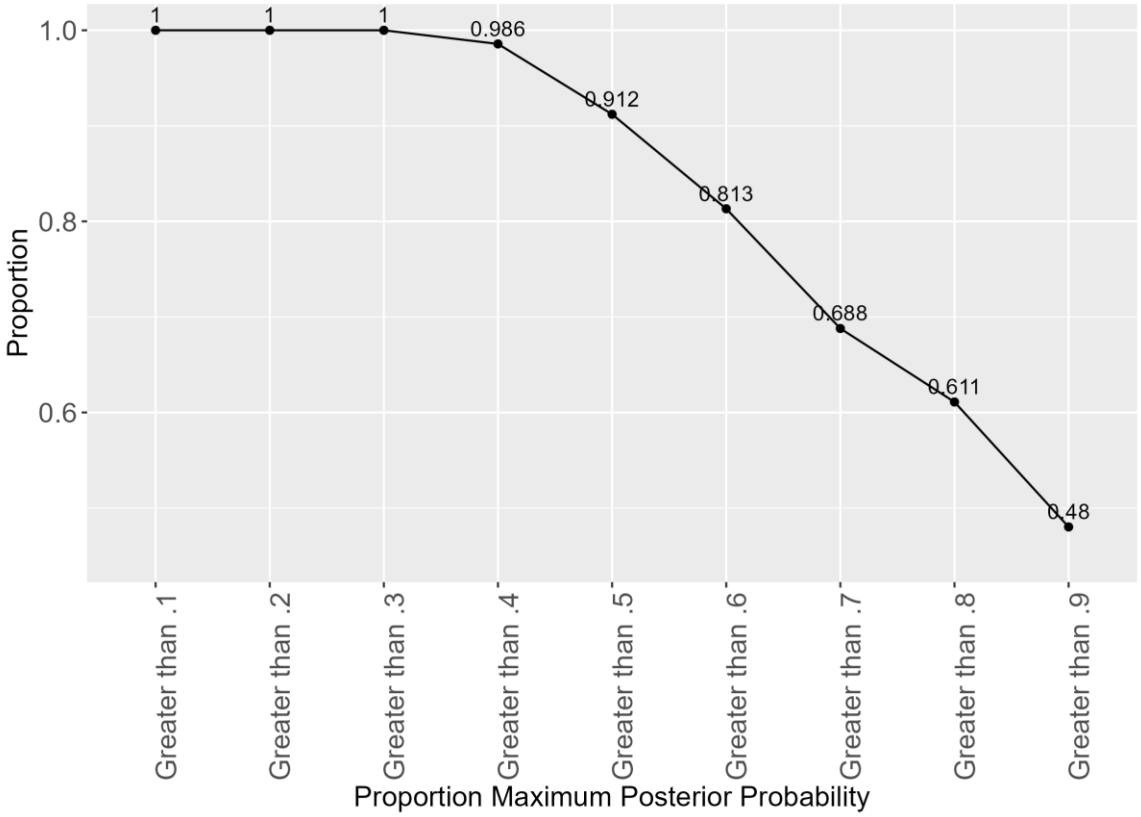


Figure 64

Basic DGPs for the Empirical Data Analysis with One Three-Category Attribute

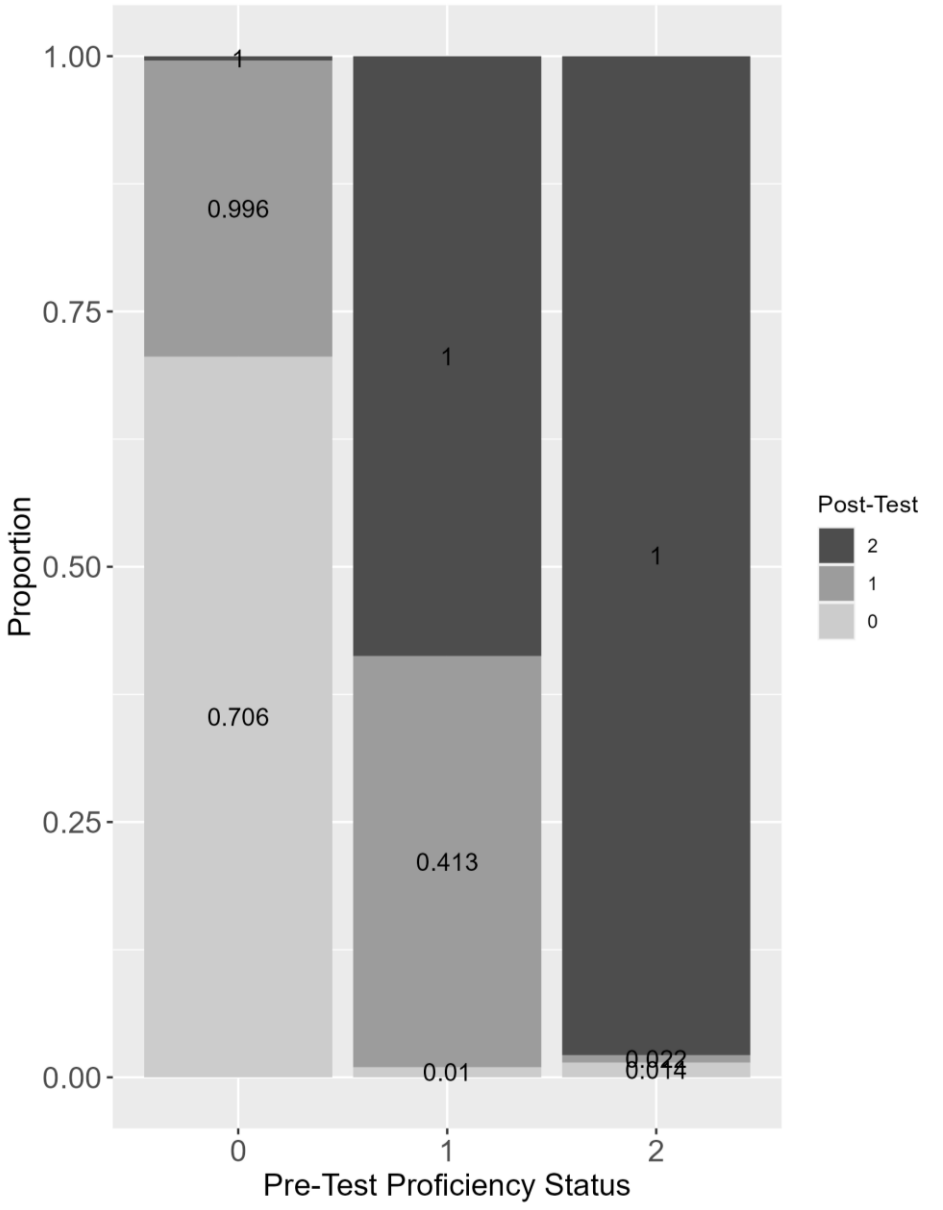
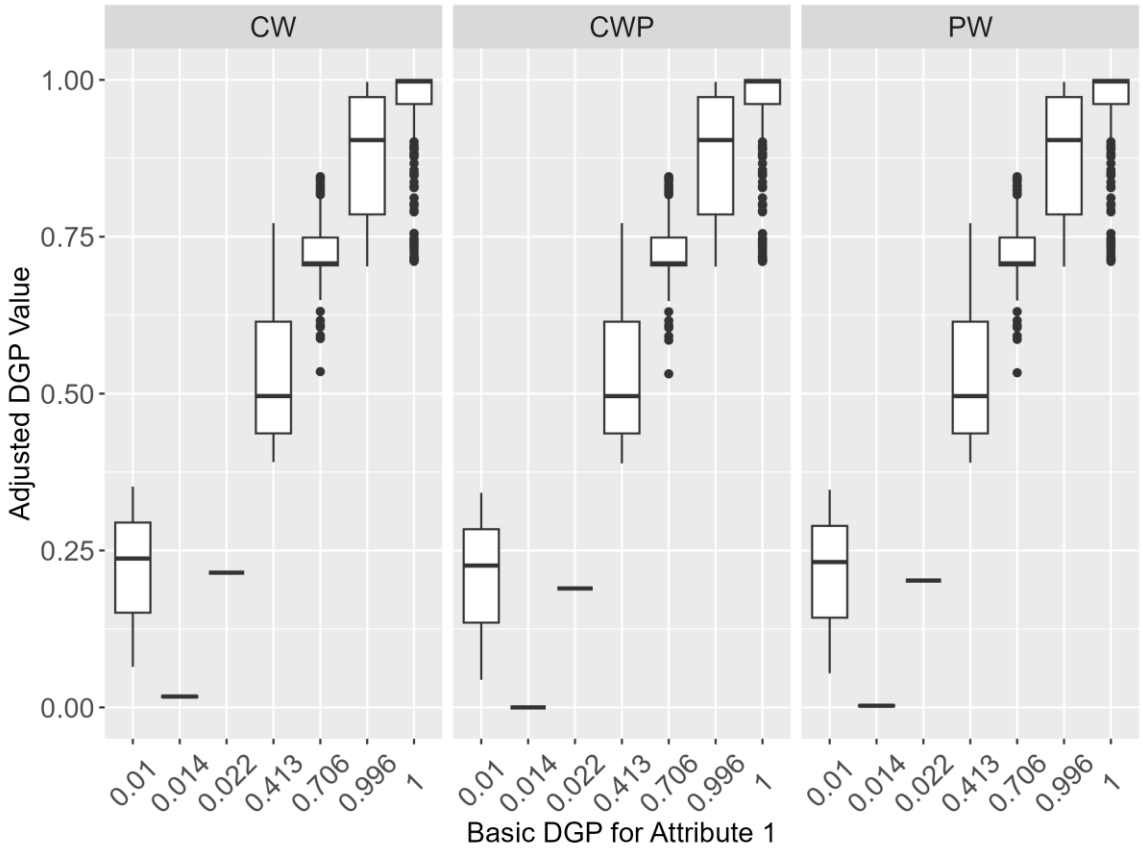


Figure 65

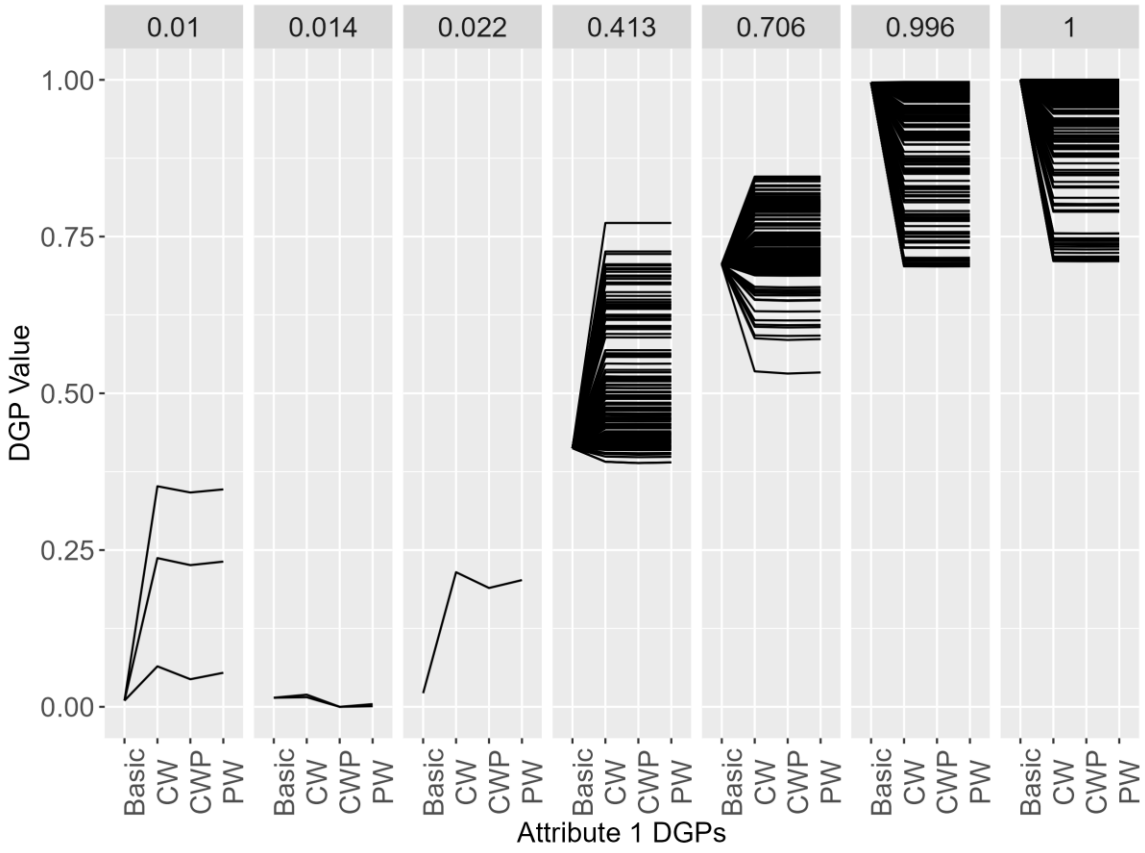
Adjusted DGP Boxplot for the Empirical Data Analysis with One Three-Category Attribute



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting.

Figure 66

Student's DGP Plots Split by Basic DGP for the Empirical Data Analysis with One Three-Category Attribute



Note. CW = Adjusted DGP with complete weighting; CWP = Adjusted DGP with complete weighting and a penalty for forgetting; PW = Adjusted DGP with partial weighting.