Item cluster-based assessment: Modeling and design

by

Amy Arneson

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

Committee in charge:

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

Spring 2019
Abstract

Item cluster-based assessment: Modeling and design

by

Amy Arneson

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This three-paper dissertation explores item cluster-based assessments, first in general as it relates to modeling, and then, specific issues surrounding a particular item cluster-based assessment designed

There should be a reasonable analogy between the structure of a psychometric model and the cognitive theory that the assessment is based upon. Specifically, for item response theory (IRT) models in educational assessment scores, the structure of dependencies among items that are designed as item clusters (groups of items that share common stimulus material, etc) should be reflected in the model. This type of designed local item dependence (LID) can be modeled in many different ways. The literature on the existence of LID and models developed to account for this LID is somewhat extensive, though there is little work to unify and organize these different approaches. The first paper presents a general framework to guide modeling decisions for item cluster-based assessments by first formalizing some of the terminology used in the context of LID, providing an overview of methods for detecting LID, and discussing general modeling approaches for response data that is theorized to exhibit LID.

Recent pushes for increased rigor and focus on complex constructs (such as critical thinking) in K-16 education highlight a need to develop assessments that measure these complex constructs. The second paper explores these issues in the context of a particular complex constructs in statistics education, that of Linking Data to a Claim (LDC), Meta-Representation Competence (MRC), and Formal Inference (FoI). We present a multidimensional treatment and analysis of field test data for the Critical Reasoning for College-Readiness (CR4CR) Assessment, an item cluster-based assessment. We found that the LDC and FoI items as written can provide a mapping of student ability estimates to the construct map levels as defined, but that the MRC items do not. Further, as expected, we found moderately strong correlations among the three constructs.

The third paper describes the design of selected response items based on open-ended counterparts for the CR4CR Assessment, and the empirical comparisons of these different formats. It is commonly thought that multiple choice (or selected response) items on tests do not provide useful information to educators regarding higher level thinking skills such as argumentation or critical thinking. However, there is also a need for diagnostic assessments
to provide educators with timely feedback on student performance so that instruction can be adapted or interventions administered based upon student needs. We found that though existing literature suggests that selected response item types are easier, in general, than constructed response item types, this may not be the case for all constructs. We found that, for the LDC and FoI constructs, multi-select multiple choice items behaved similarly to their constructed response counterparts.
for Neko
# Contents

1 A framework for modeling item cluster-based assessments 1  
1.1 Introduction ............................................................... 1  
1.2 Local item dependence .............................................. 6  
  1.2.1 Definitions ......................................................... 6  
  1.2.2 LID and dimensionality ........................................ 9  
  1.2.3 Detecting LID ..................................................... 11  
  1.2.4 Early methods for dealing with LID ....................... 13  
1.3 Cluster design structures .......................................... 14  
1.4 Approaches to modeling LID ....................................... 18  
  1.4.1 Testlet models ................................................... 20  
  1.4.2 Bundle scoring conventions .................................. 23  
  1.4.3 Theory-based models .......................................... 25  
  1.4.4 Issues ............................................................... 27  
1.5 Conclusion ............................................................. 32  

2 Developing assessments for college ready critical reasoning in statistics 34  
2.1 A framework for critical reasoning for college readiness in statistics .... 36  
  2.1.1 Linking Data to a Claim (LDC) ............................... 40  
  2.1.2 Meta-Representational Competence (MRC) .............. 43  
  2.1.3 Formal Inference (FoI) ........................................ 46  
2.2 Assessment tasks ..................................................... 51  
2.3 Empirical analysis .................................................. 52  
  2.3.1 Methods ........................................................... 52
# List of Figures

1.1 The Creativity Development Inventory, reproduced from Wang et al. (2005). .................................................. 2
1.2 Path diagram showing an item cluster-based assessment with six items and two clusters. ........................................... 10
1.3 Sample item cluster, *Philip Glass Passage*. ................................. 15
1.4 Sample item cluster, *Plant Species*. ................................................................. 16
1.5 Sample item cluster, *Seatbelt Law*. ................................................................. 17
1.6 Random effect testlet model path diagram. .................................................. 21
1.7 Path diagram for second-order model. .................................................. 22
1.8 Path diagram showing an interpretive and an incidental dimension. ........ 23
1.9 Testlet-like approach for a two-dimensional fifteen-item test containing a five-item cluster. .................................................. 30
1.10 Three sum score approach choices for a two-dimensional test containing a multidimensional cluster. .................. 31

2.1 Framework triangle for the CR4CR framework. ................................. 36
2.2 The Linking Data to a Claim (LDC) construct map. .............................. 40
2.3 Sample item from AP Comparative Government & Politics: *Birth Rates*. 43
2.4 The Meta-Representational Competence (MRC) construct map. ................ 44
2.5 The Formal Inference (FoI) construct map. ................................................. 48
2.6 Sample CR4CR Assessment item cluster: *Miles per Gallon (MPG)*. ........ 53
2.7 LDC Wright Map. .................................................................................. 59
2.8 MRC Wright Map. .................................................................................. 60
2.9 FoI Wright Map. .................................................................................. 61

3.1 The *P-Value* item. ............................................................................... 67
3.2 Framework triangle for college-ready statistical thinking. 71
3.3 The Linking Data to a Claim (LDC) construct map. 74
3.4 The Formal Inference (FoI) construct map. 78
3.5 The selected response version of the Twitter item cluster. 82
3.6 Scoring guide for Twitter 2. 83
3.7 Scoring guide for the Paternity Leave 2 item. 84
3.8 The original constructed response version of the Twitter item cluster. 86
3.9 The revised constructed response version of the Twitter item cluster. 87
3.10 Path diagram for LDC (Form G) analysis. 89
3.11 Covariance matrix for the FoI model. 90
3.12 Covariance matrix for the LDC model. 90
3.13 Path diagram for FoI (Form I) analysis. 91
3.14 Scatter plot of comparison item thresholds. 98
3.15 Wright Map for the LDC (Form G) analysis. 100
3.16 Wright Map for the FoI-S items from the FoI (Form I) analysis. 102
3.17 Wright Map for the FoI-P items from the FoI (Form I) analysis. 103
A.1 The constructed response version of the Paternity Leave item cluster. 115
A.2 The selected response version of the Paternity Leave item cluster. 116
A.3 The constructed response version of the Miles per Gallon item cluster. 117
A.4 The selected response version of the Miles per Gallon item cluster. 118
A.5 The constructed response version of the Chocolate item cluster. 119
A.6 The selected response version of the Chocolate item cluster. 120
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A framework for modeling item cluster-based assessments</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>The CR4CR constructs</td>
<td>39</td>
</tr>
<tr>
<td>2.2</td>
<td>Formal Inference sample performances</td>
<td>49</td>
</tr>
<tr>
<td>2.3</td>
<td>Form design by item for the CR4CR field test</td>
<td>55</td>
</tr>
<tr>
<td>2.4</td>
<td>Respondent demographics</td>
<td>56</td>
</tr>
<tr>
<td>2.5</td>
<td>Substantive dimension results</td>
<td>57</td>
</tr>
<tr>
<td>2.6</td>
<td>Covariance and correlation estimates among substantive dimensions</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Formal Inference sample performances</td>
<td>79</td>
</tr>
<tr>
<td>3.2</td>
<td>CR4CR Assessment field test test form design</td>
<td>88</td>
</tr>
<tr>
<td>3.3</td>
<td>Information about empirical analyses</td>
<td>94</td>
</tr>
<tr>
<td>3.4</td>
<td>Model estimates and fit statistics</td>
<td>95</td>
</tr>
<tr>
<td>3.5</td>
<td>Item rater reliability for CR items</td>
<td>95</td>
</tr>
<tr>
<td>3.6</td>
<td>Item parameter estimates for comparison items</td>
<td>96</td>
</tr>
<tr>
<td>3.7</td>
<td>Step parameter estimates for comparison items</td>
<td>97</td>
</tr>
<tr>
<td>B.1</td>
<td>Details on simulation variance/covariance generating values</td>
<td>121</td>
</tr>
<tr>
<td>B.2</td>
<td>Details on simulation form designs</td>
<td>122</td>
</tr>
<tr>
<td>B.3</td>
<td>Simulation study results, reliability estimates</td>
<td>122</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to express gratitude to the following people who made this work possible. First, to my family who are friends and friends who are family for the support and continuous encouragement: Dad, Mom, Kristen, Jeffrey, Barbara, Matt, Elise, Ray, Sara, Julia, and Courtney. Second, to the faculty and students, past and present, of the Quantitative Methods and Evaluation program at the University of California, Berkeley for the academic support, potluck lunches, and Friday game nights: Mark Wilson, Sophia Rabe-Hesketh, David Torres Irribarra, Laura Pryor, Joonho Lee, Diah Wihardini, Nicholas Sim, Rebecca Freund, Shruti Bathia, Jinho Kim, Karen Draney, and Linda Morrell.

This dissertation would not have been possible without the pre-doctoral training and financial support provided by the Research in Cognition and Mathematics Education grant funded by the Institute of Education Sciences and the financial support from the National Math and Science Initiative.
Chapter 1

A framework for modeling item cluster-based assessments

1.1 Introduction

A common practice in educational test design is to write sets of items, or *item clusters*\(^1\) that all refer to the same stimulus material. This is done, often, in the interest of testing efficiency – both in terms of time and the cognitive load on the part of the respondent. For a respondent to read a lengthy passage and only be asked a single question about it could be frustrating for the respondent (as considerable time and mental energy was spent reading the passage), could mean a significant cut in instructional time (if a test was made up of many of these lengthy passages), or be viewed as a wasted opportunity to have a respondent engage deeply with a topic. Aside from the time and cost efficiency of using item clusters, asking *multiple* questions about lengthy and/or cognitively demanding stimulus material also allows for scaffolding, incorporation of higher-level thinking skills (e.g. coordinating multiple perspectives), producing information on complex constructs that are not likely to be revealed in responses to single items, and a more realistic reflection of “real-life” situations. Due to the administrative and educational advantages of item clusters on tests (not limited to those listed here), it is often recommended that tests contain item clusters, as is the case for assessments under development for the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) by the Science Assessment Item Collaborative framework (WestEd & Council of Chief State School Officers, 2015) and the National Research Council (Pellegrino, Wilson, Koenig, & Beatty, 2014).

This seems straightforward and uncontroversial. Classroom teachers and item writers for large-scale assessments often employ item clusters for reading comprehension and analytical reasoning tests, performance-based science and mathematics tasks, and scenario-based measures (Bradlow, Wainer, & Wang, 1999; Wang, Cheng, & Wilson, 2005). So, how does this established and standard practice find its way into a dissertation? The issue does not

\(^1\)They are also commonly referred to as *item bundles* (Rosenbaum, 1988) or *testlets* (Wainer & Kiely, 1987).
lie in the test design practice, but in the statistical modeling of scored response data from tests that contain item clusters. Item response theory (IRT) models that are used for educational assessment data, as with all regression models, come with a collection of underlying assumptions which are made often for model identification or ease of estimation. One of the basic convenience assumptions of traditional IRT models is that of \textit{local independence} (a.k.a. conditional independence). That is to say that, on a unidimensional test, for a student with a given ability ($\theta$) the responses to any item is independent of all others. In the case of responses to items within the same item cluster, that assumption is likely violated. Violation of this assumption is called \textit{local item dependence} (LID). One can readily identify the effects of such LID: (a) locally dependent items are \textit{redundant} in that they provide less information on the dimension(s) of interest than if they were statistically independent. In some sense, locally dependent items create a “shorter” test. Also (b), when present and ignored, LID contaminates the meanings of the latent variable(s) in the model. The greater the dependence present among items on an instrument, the greater the contamination (Wang et al., 2005). As well (c), carryover effects from previous items may artificially increase or decrease a respondent’s location on a latent trait. For example, the Creativity Development Inventory provided in Wang et al. (2005) and reproduced in Figure 1.1 has respondents rating importance of different personal characteristics in creativity development \textit{and} how much they possess these characteristics. If they first judge importance highly, they may inflate (or deflate) their attitude toward possession. In this case, a carryover effect from the importance responses are expected in the possession responses.

Why care about violation of this (or any) assumption? Statistical modeling can only yield meaningful and useful results if the model, which includes its assumptions, is grounded in existing knowledge and theory about the phenomena it is purported to model. For any statistical investigation, model assumptions should be made explicit and critically questioned. Are we comfortable with these assumptions? How robust are our results to the violation of these assumptions? What are the consequences (statistical and practical) of violating these assumptions? When are the effects of assumption violation too small to worry about? Consider a specific context in educational testing—a statistical model is applied to scores on a placement test for incoming college students: a statistical consequence of violating the
assumptions would be something like “inconsistent parameter estimates” whereas a practical consequence would be “incorrect course placement.”

What makes the case of item clusters interesting is that LID is induced by the test design itself—in a way, it’s an intentional violation of an assumption. Even if test designers are not aware of the modeling implications (violation of the local independence assumption), we still refer to this as designed or intentional LID though it could be argued that in some cases this is done unwittingly by the item writers.

Having common stimuli is not the only test characteristic that can lead to (expected) LID. Yen (1993) offers a fairly comprehensive list of situations in which dependence among items may be expected. These situations can be classified as either designed LID or unintentional LID. Designed LID arises from item clusters or from controlled test administration conditions. Examples (adapted from Yen, 1993) of this include:

- passage dependence: multiple items all relate back to a common stimulus;
- speededness: tests in which time management is an issue, items near the end of the test tend to be positively, locally dependent;
- practice: exposure to item types may improve performance later in the test;
- item chaining: items are organized in sequential steps;
- item or response format: items that have a similar format may have higher score correlations than items of different formats;
- scoring rubrics or raters: items scored with the same rubric may exhibit higher score correlations as well as those scored by the same rater; and
- content, knowledge and abilities: if an item cluster is set in the context of a rugby game, respondents differing knowledge about rugby may affect their responses.

Unintentional LID arises from situations that are out of the control of the test designer such as:

- external interference: classroom disruptions, faulty materials, inaccurate information from test facilitators or fellow respondents;
- respondent fatigue; and
- facilitated items: external assistance from a test facilitator or other respondent.

---

2Yen (1993) does not make this distinction. The categorization of elements from her list is done here to provide examples in each category.

3It is possible that items can be considered a cluster even if they are not administered as a set. Items scattered throughout an assessment could refer back to a common stimulus. For simplicity of discussion, we stick to items delivered as a set in immediate proximity to each other.
Some of the statistical consequences of ignoring LID have been documented in the literature. Estimates of item and person locations can be biased due to model misspecification. These erroneous parameter estimates can then lead to incorrect inference and decisions on any analysis based on those estimates, such as testing for correlations among latent variables or testing for differences among respondent groups (Wang et al., 2005). And, in terms of assessment design, not modeling LID may be a limitation when making item design decisions. By ignoring LID in the modeling, in essence, there will be missing elements of the model. Decisions about including, excluding, or revising items then cannot be informed by the nature of dependence among items. If items are locally dependent, deciding to change one item may affect another in a meaningful way. The validity of using model estimates to inform test construction and scoring is based in the local independence assumption (W.-H. Chen & Thissen, 1997).

As for violation of the independence assumption of any regression model, the estimators of standard errors of the parameter estimates will be inconsistent. In the case of LID, the standard errors for item parameters tend to be underestimated\(^4\), and this overstatement of precision can lead to incorrect inference. Specifically in the context of computerized adaptive testing (CAT), standard errors for \(\hat{\theta}\) are an element of stopping rules (determining when the respondent has provided enough information to estimate their ability). If these standard errors are underestimated, the test may stop before it should (Bradlow et al., 1999). Further, it has been shown that test reliability is overestimated when LID is present and ignored (for tests of the same length built from all independent items) as it is artificially inflated by the increased correlations among clustered items (Sireci, Thissen, & Wainer, 1991). From this overestimation of reliability, practical ramifications may be that a test is deemed “reliable enough” and put into use when it truly is not. This is especially unacceptable in instances where changes and growth are being measured (Wainer & Thissen, 1996). It is important in any test data analysis situation, however, to consider the effect size and practical implications of an overestimation of reliability in relation to the test use. It may not be of practical concern if small changes in scores (relative to chance variation) would not translate to, for instance, students being treated differently.

This paper is focused on the issue of designed LID on assessments that contain items intentionally grouped by a common context. The context could be a reading passage, a graph or table, or a multi-stage task. Many of the approaches described herein can be applied to other LID circumstances, but the discussion will be limited to explicit clustering of items by common stimulus. The choice of the term cluster is intentional and meant to mirror its use in the multilevel modeling literature that focuses on clustered data.

Creating item clusters is common practice in both classroom and large-scale assessments. In fact, it has been recommended to develop cluster-based assessments in recent years, especially for tests aligned to the latest wave of academic standards such as Common Core State Standards (CCSS) and the Next Generation Science Standards (NGSS) (National Research Council, 2001; WestEd & Council of Chief State School Officers, 2015). This move toward “larger more coherent item groupings” by large-scale test developers started well before the

\(^4\)There are situations in which dependence among dependent variables can lead to overestimated standard errors, but underestimation is a greater practical concern in terms of inference.
introduction of even this latest round of more rigorous standards (Wainer & Thissen, 1996, p. 26). These recommendations come from respected and established figures in psychometrics who are well aware of the ramifications such as reduced test reliability. One reason for this is that there has been a move toward more “authentic” tests (Wainer & Thissen, 1996). Large, multi-step tasks as opposed to small, independent (often multiple choice) items carry more face validity with the educational community because they are, arguably, closer to real-life problems. Further, the attributes that have been deemed important to measure by the adoption of standards that emphasize deep understanding over discrete bits of knowledge may require collections of interdependent items rather than independent ones. Item clusters are not the only path to greater authenticity: extended essay, portfolio, and other assessment structures that require human judges/raters also introduce noise in scores, reducing test reliability. It is also widely known that variation among raters reduces test reliability. For both cases mentioned here, item clustering and rater effects, methods have been developed to combat but not necessarily eliminate the negative empirical effects. It is the general consensus that test authenticity is often worth the reduced reliability (Wainer & Thissen, 1996).

That item clusters violate an important assumption of many IRT models has been recognized by model developers for some time (Cureton, 1965; Kempf, 1977; Rosenbaum, 1988). This issue has been dealt with by modifying existing models, developing new models, and exploring methods from other disciplines that deal with similar modeling issues. What there is still room for in the literature is the exploration of how these methods presented as solutions to the problem of LID impact the interpretation and use of analysis results and test scores. Models that account for LID are a particular class of IRT models that enjoy a robust literature. However, to my knowledge, there exists no general framework or guidance for which of these methods to use in different situations. There has been some work in organizing models that account for LID into categories based on the model structure or to match to test designs (Rijmen, Tuerlinckx, de Boeck, & Kuppens, 2003; Wang et al., 2005; Wilson & Adams, 1995). This chapter builds upon that organizational work by situating a number of these methods in a framework to guide meaningful modeling.

Meaningful modeling has a stated purpose (e.g., predicting, explaining) and is transparent about its limitations, usually arising from assumptions of the model made for convenience. This framework for modeling item cluster-based assessments aims to provide a structure for meaningful modeling of item cluster-based assessment data. This first section summarizes existing literature and proposes definitions and classifications of terms to unify the literature and the framework itself. Subsequent sections describe item cluster design structures, organize modeling approaches into a framework, and finally concludes with a short discussion. This work is focused on how to relate the situational context of item clusters to statistical models, and not the technical details of the statistical estimation methods. It is intended to provide the conceptual framework to offer guidelines for modeling assessment data – not a compendium of this class of models. It may serve as the starting point for a compendium in future research.

The study herein is organized into sections based upon these summarizing questions:
1. What is the language (including specific terms) that is used in the context of LID in the existing literature? What overlaps or discrepancies exist in the use of this language?

2. What are the differences and similarities among item cluster structures used in educational assessment?

3. What are the commonly used approaches to modeling LID? How might we organize these approaches into a framework?

1.2 Local item dependence

1.2.1 Definitions

**Item cluster.** As in the previous section, we define *item cluster* as *a set of items that all refer to the same stimulus material*. These sets are defined by the item designers, not empirically by examining the data for local dependence, and are administered as a set. Each item in an item cluster receives a unique score. These sets may be static or adaptive, may be presented at the same time (on the same paper or screen) or individually, and may be restricted in that respondents cannot review and/or change previous responses once they have moved on to the next item. It is expected that these groups of items will produce locally dependent responses, but statistical dependence is something to be tested for, not a characteristic to define these sets. We have adopted what we feel is a fairly broad definition. The concept of groupings of items has been labeled elsewhere in the literature as *item bundles* or *testlets*, most papers defining and using these terms in the specialized context of the paper. These two terms, *item bundles* and *testlets*, are probably the most commonly used terms and are often used interchangeably. We use *item cluster* as opposed to either of these terms, because the existing terms are also commonly used to name and describe models and/or modeling approaches. We want to differentiate the test design process from the response data modeling process.

Terms for grouped items can be defined either from the *design* perspective (items that are grouped based on item characteristics) or from the *empirical* perspective (items that exhibit unexpectedly high score correlations). The definition adopted in this work is from the design perspective. It is helpful to note that even the fundamental term *item* is used in design, empirical, or both senses in any given work in the literature, often without a clear definition. *Item* in the empirical sense is the element of a test that receives a score – the data used in the analysis – and is thus, crossed with the respondent, the unit of analysis in the statistical modeling. In the design sense, an item is an indexed element of a test for which a response is provided. The responses vary in lengths and may, in some cases, receive multiple scores.

Andrich (1985) uses the term *subtest* which he describes in two (unidimensional) contexts from the design perspective. First, he states that a subtest is a group of items, determined

---

5These scores may be combined or transformed before feeding data to a model estimation engine.
a priori, that are deemed to “belong together” (p. 255). For the specific example that is used in his paper, items are grouped by math content area (i.e. whole numbers, decimals, fractions, ratio and proportion). This broad definition could include any a priori grouping of items. The other context he described is when multiple scores are assigned to a single response, usually an extended written or essay response. Different “pieces” of the response each receive their own score. So, in the design sense this is a single item with multiple scores, and in the empirical sense, these are multiple items. In this paper, all of the scores are assumed to be on the same latent trait\(^6\). Both of these contexts would constitute an item cluster as defined above.

Rosenbaum (1988) defines an item bundle, also in the design perspective, in terms of common stimulus but with narrow conditions on other item design characteristics: “a small group of multiple choice items that share a common” stimulus or “a small group of matching items that share distractors” (p. 349). These groups of items may violate the conditional independence assumption, but this violation is not a requirement of the definition. In this sense, an item bundle is a design element of an instrument, not a way to describe an empirical finding of LID. Though Rosenbaum’s initial definition is quite narrow, the discussion elaborates that the interesting thing about these groups of items is that they share (some but not all) cognitive tasks. This aspect of item sets without restrictions on the items being forced choice is a more helpful defining characteristic for furthering work on item bundle modeling – especially as innovative item types are being used in classroom and wide-scale assessments.

Wainer and Kiely (1987) define a testlet as “a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow” (p. 190). Here, they are describing the development of a CAT in which previous responses to items determine which item will be presented next to an examinee. These sequences of items create a specialized test that could end up being unique to each examinee. They advocate the creation of testlets, for one, to reduce the number of possible tests that can be administered to a set of respondents. Testlet paths may be linear or hierarchical. They describe the “testlet framework” as a guiding principle for designing a CAT. Thus, the CATs developed in this framework are expected to be built entirely of testlets.

Literature written for an audience of practitioners tends to use the term task (or, often, performance task) to describe long-form item responses or sets of item responses situated under a shared context. These often have the aim of being authentic tasks, defined by the NRC’s Knowing What Students Know report (2001, p. 30) as “tasks that involve the application of combined knowledge and skills in the context of an actual project.” Because performance tasks take more time to complete, it is often (but not always) the case that they produce multiple scores. Even for tasks that ask for a long-form, but singular, response may receive multiple scores. For example, an essay-type item may be assigned a score on each of six elements of writing: content, conventions, diction, fluency, structure, voice. Any time multiple scores are situated within a common context, it is an item cluster.

---

\(^6\)For discussion and analysis on multiple scores in multiple dimensions, see Draney (2005).
In the discussion above, there are three possible units of analysis (crossed with persons) that have been discussed either explicitly or implicitly, here: the item, the item cluster, and the test. A test is a collection of items that may or may not be organized into item clusters. An independent item may be considered a cluster of size 1, but we instead refer to them as independent items. When designing tests, one may consider the “basic unit of construction” as the test item, the item cluster, or both. Thus, one can categorize any test into one of the following four categories:

1. scenario-based: the test is comprised of a single item cluster (all items belong to the same cluster);
2. cluster-based: the test is comprised of item clusters (all items belong to a cluster of at least size 2 and there are at least 2 clusters);
3. item and cluster-based: the test contains both singular items and item clusters (of at least size 2);
4. item-based: there are no item clusters (no items share stimulus material).

The modeling framework described in this paper does not apply to scenario-based assessments in which all items are situated under the same scenario. From a data analysis standpoint, scenario-based and item-based assessments would be analyzed in equivalent ways. If all items on a test refer back to the same stimulus material, there will be no way to disentangle a “scenario-effect” as it is a test-level characteristic. This work is limited to instruments that fall into categories 2 and 3.

**Local item dependence (LID).** A pair of items on an instrument exhibit *local item dependence* if there is a non-zero correlation of scores after conditioning on the ability of the respondent. LID can be positive or negative, though in the case of item clusters, we expect it to be positive among the items in the same cluster. This can also be conceptualized as *residual dependence* in that there will be remaining dependence among the residuals that is not explained by a model that does not allow/account for them.

Rosenbaum (1984) reasons that the term *conditional* independence should be used over *local* independence in IRT, citing the Schweder (1970) argument that *local* should be used in the context of time and space. Since temporal or spatial locations are not the only ways that dependence may be induced among sets of items, using *conditional* is more appropriate. However, because we are dealing within the context of item clusters, which are administered temporally and spatially close together, we have adopted the term *local* independence for this phenomenon.

LID is the statistical result of a *context effect* which Wainer and Kiely (1987) define as any influence an item may acquire as a result of its relationship to other items on the test (p. 187). Examples of characteristics that may give rise to context effects are item location within a test, cross-information, and unbalanced content. Measures of LID aim to quantify these context effects. One specific type of effect, a *carryover effect*, occurs when a response is affected by previous responses. This happens particularly when items are
chained or in instances where respondents make a concerted effort to have their responses to similar items match. Recall Figure 1.1 from Wang et al. (2005) in which respondents rated personality characteristics in two ways: importance and the degree to which they possessed them. Carryover effects are anticipated in situations such as this.

W.-H. Chen and Thissen (1997) break LID into two types: underlying and surface. In this context, underlying LID in essence introduces a latent dimension unique to each item cluster. This dimension represents a context effect and is theoretically independent of all other latent dimensions being modeled. The “ability” for each item becomes a linear combination of the value of dimension(s) being measured (i.e. the dimension(s) of interest) and the dimension unique to the cluster. We recognize that the independence of the context effect and the dimension of interest may be a constraint done for convenience in estimation or interpretation or because they are actually theorized to be independent. Surface LID is a very special case to describe when a pair of items are perfectly locally dependent. Here, the response to an item is completely determined by the responses to another item on the test. These are the most redundant of locally dependent items and would be odd in the context of educational assessment. This classification of LID into types is what we refer to as the nature of the dependence in the sections that follow. By this, we mean how the dependence among items can be described and/or explained in more qualitative terms beyond noting that there is a correlation.

1.2.2 LID and dimensionality

We limit this review, initially, to models with a single θ dimension, referred to as the unidimensional context. LID and multidimensionality manifest themselves mathematically in test data in the same way - as residual dependencies. It is important, then, to distinguish between interpretive dimensions (a.k.a. psychometric, substantive, general, or conceptual dimensions) which are the targeted phenomenon an instrument measures and incidental dimensions (a.k.a. specific dimensions) that are caused by the test design characteristics that induce LID. The interpretive dimensions are often defined by the test designer as a specific skill, construct, or trait (e.g.s. math ability, scientific argumentation sophistication, happiness) while the incidental dimensions are unique to an item cluster. Especially for random effect models (discussed later), the dimensions associated with an item cluster are referred to as testlet effects or testlet dimensions.

Consider a test that is designed to measure multiplication ability and help a teacher identify students for a multiplication intervention. The dimension under study, ability to multiply, is shown in the top center oval in Figure 1.2. Here, we assume that multiply causes the responses to six items, indicated by the arrows leading to each of the item responses ($y_i$). In addition, suppose the items are clustered by two common contexts: there are items about finding the area of a rectangle and items about a rugby game. A path diagram, a tool commonly used in causal modeling and structural equation modeling to specify a model, is used here to illustrate the cluster structure of this hypothetical test and borrows some of the conventions of these more formal path diagrams. In this paper, they will be used to describe cluster structures and illustrate differences among models, but should not be interpreted as
fully specifying any model. Whether an \textit{area}_p or \textit{rugby}_p dimension is included in the IRT model used for scores produced by this test is a decision that must be made based upon the cognitive theory driving the test construction and analysis and/or the research question the data analysis is meant to address.

One question of cognitive theory is whether the item cluster effects are interpretive or incidental dimensions. In this case, a student’s \textit{multiply}_p and \textit{area}_p abilities are probably both of consequence to her math teacher when making a decision about a multiplication intervention. We might consider treating \textit{area}_p as a dimension in the IRT model. The other item cluster’s common context, \textit{rugby}_p, is meant to illustrate a nuisance dimension. The student’s familiarity with and knowledge of rugby may influence her responses, but not in a way that will be helpful to the teacher making a decision about the multiplication intervention. The clustering should be accounted for in the model, but the teacher likely doesn’t care about a student’s location on a rugby dimension.

The terms used to differentiate these dimensions is not meant to imply that incidental testlet dimensions should always be considered a nuisance. The incidental dimensions that arise as an artifact of item clustering may be of importance in the context of the test use or important in understanding how the items on an instrument function together. These incidental dimensions could be the focus of a research question driving the analysis. In other cases, they may be recognized as important dimensions, but not in the context of the research questions or test use. In yet other cases, they may be regarded simply as nuisances.

We contend that LID arising from item clustering is too often considered a nuisance. Common “solutions” to the “problem” of LID that assume it is a nuisance fall into either the design realm – redesign the items to reduce or eliminate the residual dependency (Hoskens
& de Boeck, 1997) – or the modeling realm – toss orthogonal random effects into the model to soak up the residual dependency. However, for many of the practical reasons discussed in the introduction, eliminating item clusters is often not a feasible move on the part of the test designers. Item clusters are often designed intentionally for reasons related to validity, the nature of the construct being measured, and administration constraints like time and money. That the resulting mathematical dimensions of item clusters are nuisances is an issue that should be considered in terms of the research question(s) an analysis of the response data intend to answer and in terms of the test use. How testlet dimensions are regarded is the central issue in determining which modeling approach to take. Some possible approaches are discussed in a later section and constitute this framework for modeling item cluster-based assessments.

1.2.3 Detecting LID

Papers in the LID literature usually have one of two objectives: either detecting or modeling LID. There are some model formulations that can be used to test for LID – usually by comparing a model that accounts for LID to one that does not by using likelihood-ratio tests or an information criterion like Akaike Information Criterion (AIC). However, this approach does require that the analyst has some theorized or expected LID in mind in order to build the less restrictive model and that the models being compared are nested. This would be a confirmatory approach. LID indexes have also been developed for more exploratory identification of (pairwise) LID in instruments.

When LID is anticipated, there are “predictable and testable consequences” when a test includes clusters of items (Rosenbaum, 1988, p. 350). Empirical checks can be employed to test whether there is statistical dependence. These often take the form of comparing observed item characteristic curves (ICCs) or score frequencies to predicted ICCs or frequencies. Rosenbaum (1984) states that methods for detecting LID should be sensitive to the violations of local independence, but not to parametric assumptions about the item characteristic curves. Often classical methods for testing hypotheses about dependence meet this criterion. For instance, Mantel-Haenszel tests (Mantel & Haenszel, 1959; Rosenbaum, 1984; Braeken, Tuerlinckx, & de Boeck, 2007) can be performed on pairs of dichotomous items to see if the data is inconsistent with the local independence assumption. Because of concerns regarding multiple hypothesis testing, these methods should be used for leads on potential local dependence to be investigated, not for conclusions. The multiple comparison problem only exists if inferences are to be made from the data, but using multiple comparisons in an exploratory way is acceptable. If significance tests are to be used to make decisions, however, corrections for multiple testing are needed. It is important to try to develop effect sizes for LID effects. Some other methods for investigating the presence of LID are summarized in the following paragraphs.

Indices developed for detecting LID are often summaries of model misfit. Yen (1984) discusses a number of $Q$ statistics that do this for the three-parameter logistic (3PL) model. Calculating $Q_1$ (Yen, 1981) involves splitting respondents into deciles and, for each item, comparing the observed and expected counts of correct responses (of dichotomously scored
items) within each decile. These comparisons are summed over the deciles to produce the \(i\)th item’s fit \((Q_{1i})\), and then the item fit statistics are summed to get the fit of the instrument, \(Q_1\). The \(Q_2\) statistic involves finding a fit for each item pair \(i, j\) using expectations based on the Rasch model (which Yen (1984) generalized to the 3PL). Yen’s \(Q_3\) statistic for an item pair \(i, j\) is the correlation of the residuals over all respondents between the residuals\(^7\) for item \(i\) and the residuals for item \(j\). This is not an exhaustive list of developed methods for detecting LID via model misfit. See also Reiser (1996); Cai, Maydeu-Olivares, Coffman, and Thissen (2006); Reiser (2008).

W.-H. Chen and Thissen (1997) describe four indexes to be used for dichotomous items modeled with IRT that are routinely used for investigating association in \(2 \times 2\) tables. In their method, \(2 \times 2\) marginal tables are constructed for each pair of items. Because marginal maximum likelihood (MML) estimation methods conveniently allow derivation of expected frequencies for response patterns within each pair of items, they can be compared to the observed frequencies. If there is substantial difference in the expected versus observed frequencies, the IRT model has induced more or less dependence than was observed. They discuss the use of Pearson’s \(\chi^2\), \(G^2\), the standardized \(\phi\) coefficient difference, and the standardized log-odds ratio. Both the standardized \(\phi\) coefficient difference and the standardized log-odds ratio are signed and give an indication of the direction and strength of the local dependence. In this way, they can be considered effect sizes if one is interested in the strength or direction of LID. The first two indexes can be used in formal statistical tests to detect whether LID is statistically significant, but do not yield information on strength or direction.

Tuerlinckx and de Boeck (2004) outline a method in which the empirical log odds ratios of pairs of item scores are graphically compared with simulated empirical log odds ratios. Even after detecting LID, the items design and/or administration circumstances should be revisited in order to theorize the nature of that dependence. This will inform modeling choices and/or revisions to items in iterative test construction approaches like the BEAR Assessment System (BAS) of Wilson (2005). This graphical approach shares its base logic with that of posterior predictive checking (Rubin, 1984) which is a commonly used in the Bayesian framework for model checking and model diagnostics. Sinharay, Johnson, and Stern (2006) discuss this method as it applies to IRT.

LID is, as defined, correlations among scores beyond those explained by the dimensions included in the measurement model. These correlations could, in fact, be explained by another dimension. However, it is important from a test use and interpretive standpoint to distinguish between LID and “departures from unidimensionality” though they manifest themselves in mathematically equivalent ways in IRT modeling (Bao, Gotwals, & Mislevy, 2006). In terms of testing for LID, however, many tests of multidimensionality can be used because of their mathematical equivalence. Model misfit due to multidimensionality also happens when LID is present. When both LID and multidimensionality are present, it is difficult to distinguish between them empirically, so it is helpful and important to have qualitative insight about which groups of items may be dependent (Rosenbaum, 1988).

\(^7\)The residual is the difference between a respondent’s score and the predicted probability of a “correct” answer (a score of 1) from the IRT model.
Rosenbaum (1988) gives a method for determining if lack of fit is due to multidimensionality or due to conditional independence violations. Essentially, one must search for violations that cannot be attributed to particular format characteristics of a test. The violations that occur in items that are grouped do not “typically suggest the existence of a new fundamental human ability” (p. 349) and may be considered a dimension only applying to the particular grouping context. These dimensions are typically not found to be important characteristics of respondents.

1.2.4 Early methods for dealing with LID

In true score theory (a.k.a. classical test theory), item dependencies are not an issue when the unit of analysis is the test. Issues only arise when interpreting item or test statistics that assume all items are independent. For instance, in using the Spearman-Brown formula to calculate the change in reliability that occurs when the length of the test is changed, it is assumed that the added or deleted items have the same properties as the existing items on a test. Thus, Spearman-Brown breaks down with the removal or addition of items that are locally dependent. Cronbach’s alpha ($\alpha$), largely interpreted as a measure of an instrument’s internal consistency or the degree to which the instrument is unidimensional, should be interpreted with caution especially when LID is suspected based on the existence of item clusters. As discussed earlier, LID and multidimensionality manifest in the same empirical way. It has been shown that tests with differing underlying dimensional structures can produce the same value of Cronbach’s alpha.

There exist warnings about “inflated reliability estimates” in early measurement texts (Guilford, 1936; Thorndike, 1951; American Psychological Association, 1966) when the independence assumption is violated, but without statistical methods to account or correct for LID (as cited in Sireci et al., 1991). Instead, it seems the initial recommendation for dealing with LID was to try to not have items with common stimulus material in an instrument in the first place: when LID is found, it should be designed out. More recently, the existence of LID was not seen as a design problem, as long as it could be modeled which was the argument made by Rosenbaum (1988). And now, in many cases (as described in the introduction) test designs that will likely lead to LID are encouraged and recommended for some educational purposes.

The central issue that LID causes for most traditional IRT models is that conditional probabilities are multiplied in the likelihood equation used in estimation. These multiplications are only valid if the observations are conditionally independent. Models that take the cluster structure into account in the calculation of the likelihood equation solves the problems of dependence. For example, the likelihoods for multidimensional models are conditioned on all dimensions included in the model. So, if the LID is considered and treated as multidimensionality in the IRT model, there is no longer an issue in forming the likelihood so long as the specific assumptions are correct. Because this approach remains popular and valid in many settings, it is described later in Section 1.4.
1.3 Cluster design structures

Item clusters can have different structures depending on the intentions of the item designer or the data analyst tasked with modeling the scored response data. They may be used for test administration efficiency in terms of test-taking time or cognitive load. Take the commonly used structure for reading comprehension tests in which a group of items are presented following a lengthy reading passage. In the simplest of cases, the items are not intended to be ordered and could theoretically be asked in any order in that a response to any one item should not affect a student’s response to any other item. Item order effects can always be tested for in the data, but we mean to refer to the case in which the items do not build upon each other, as in the Philip Glass Passage item cluster provided in Figure 1.3, a sample item set for the Graduate Record Examination (GRE; Educational Testing Service, n.d.). The three questions could theoretically be re-ordered without changing a respondent’s understanding of each one. The items in the Fruit Flies cluster from the Advanced Placement (AP) Biology exam (College Board, 2015a), provided in Figure 1.4, on the other hand, are operationally ordered in that a response for item (a) is required in order to answer item (b). It would not make sense to change their order as the respondent is asked to explain a previous response. Item clusters where a respondent is asked to choose or identify something and then asked to explain their choice are common.

As was discussed in the Definitions subsection above, it is important especially in cases with follow-up items as in Plant Species to differentiate between items in the design sense and empirical items. Parts (a) and (b) in Figure 1.4 are presented as two items in that they ask for responses separately. However, the scoring guide will determine whether this is an item cluster or not by whether one or multiple scores are assigned. If the two parts are scored together – a single score is given for the choice of data and the follow-up explanation – then Plant Species is not an item cluster at all, but a single empirical item. And, one could do both. However, if two or more scores are assigned, and thus two or more empirical items are modeled, it is an item cluster.

Another example of designed vs. empirical item clustering can be considered in item 2 of Figure 1.3. In terms of design, it is a multi-select multiple choice item. A scoring guide could be written that would make this item an empirical item cluster in and of itself! Each option could receive a score, resulting in three scores (treating each in a way as a “True/False” item) for the cluster. Alternately, the collection of choices could be scored wholistically by considering each possible combination of selections.

With a lessened focused on recall of facts and algorithms in educational testing and a greater focus on more complex constructs and skills (e.g. problem solving, critical thinking, argumentation, reasoning, application of knowledge in real-world contexts), classroom and large-scale test writers alike often design complex assessment tasks as item clusters in order to elicit rich information about these complex constructs. Consider the Seatbelt Law example shown in Figure 1.5. In item 1, a student’s basic skills in interpreting a data display are assessed. Items 2 and 3, on the other hand, aim to gather information about a student’s critical thinking skills as the score comes not from their initial “yes or no” answer, but from their construction of an argument in responding to the “Why or why not?” portion. Note
Questions 1 to 3 are based on this passage.

Reviving the practice of using elements of popular music in classical composition, an approach that had been in hibernation in the United States during the 1960s, composer Philip Glass (born 1937) embraced the ethos of popular music in his compositions. Glass based two symphonies on music by rock musicians David Bowie and Brian Eno, but the symphonies’ sound is distinctively his. Popular elements do not appear out of place in Glass’s classical music, which from its early days has shared certain harmonies and rhythms with rock music. Yet this use of popular elements has not made Glass a composer of popular music. His music is not a version of popular music packaged to attract classical listeners; it is high art for listeners steeped in rock rather than the classics.

Select only one answer choice.

1. The passage addresses which of the following issues related to Glass’s use of popular elements in his classical compositions?

(A) How it is regarded by listeners who prefer rock to the classics
(B) How it has affected the commercial success of Glass’s music
(C) Whether it has contributed to a revival of interest among other composers in using popular elements in their compositions
(D) Whether it has had a detrimental effect on Glass’s reputation as a composer of classical music
(E) Whether it has caused certain of Glass’s works to be derivative in quality

Consider each of the three choices separately and select all that apply.

2. The passage suggests that Glass’s work displays which of the following qualities?

(A) A return to the use of popular music in classical compositions
(B) An attempt to elevate rock music to an artistic status more closely approximating that of classical music
(C) A long-standing tendency to incorporate elements from two apparently disparate musical styles

3. Select the sentence that distinguishes two ways of integrating rock and classical music.

Figure 1.3: Sample item cluster, Philip Glass Passage.
Populations of a plant species have been found growing in the mountains at altitudes above 2,500 meters. Populations of a plant that appears similar, with slight differences, have been found in the same mountains at altitudes below 2,300 meters.

(a) **Describe** TWO kinds of data that could be collected to provide a direct answer to the question, do the populations growing above 2,500 meters and the populations growing below 2,300 meters represent a single species?

(b) **Explain** how the data you suggested in part (a) would provide a direct answer to the question.

Figure 1.4: Sample item cluster, *Plant Species.*

the subtle difference of items 2 and 3: item 2 is about whether the choice of data display is appropriate for a claim about trend, and item 3 is about the limits of making a causal claim. Both of these constructs are of interest in the analysis of student responses and scores (note that the assessment was intended to be multidimensional), and it was determined that asking only one of these questions in this context led to confusion for respondents. Because there is a subtle difference in the target construct, but the questions appear to be quite similar, it was found that respondents would give a response on the unintended construct unless both were included so that the subtle difference was highlighted. Measurement of complex and nuanced constructs may require some scaffolding on the part of the test designers in order to lead respondents not to the most correct answer, but to a response on the intended construct.

The design structure of an item cluster may be dictated by practical constraints on time or limitations of the test administration system. As discussed previously, they are often employed when there is lengthy and/or cognitively taxing stimulus material. Further, as in the case of the *Seatbelt Law* item cluster in Figure 1.5, the design may be influenced by the complexity of a construct or multiple constructs. In some cases, however, the structure may not reveal itself in these ways. In many cases, decisions about structure are made for more “artistic” or idiosyncratic reasons. Recall the Creativity Development Inventory shown in Figure 1.1. What is shown is what Wang et al. (2005) call the *parallel design.* The item labeled 1 is actually two items as two responses are provided and scored: an importance item and a possession item. Another way these could be structured is *sequentially.* In this structure, the 15 importance items are asked first with a respondent providing only one response for each personal characteristic followed by the 15 possession items. So, each of the personal characteristics would be presented to a respondent twice. This specialized structure of the Creativity Development Inventory is also an example where the definition of the item cluster is not straightforward. Are the item clusters defined by the personal characteristics (there are 15 of them), or by what is being rated (there are two of them: importance and possession)? Further, could the items be *cross-classified* and belong to multiple item clusters? For the simplicity of the upcoming modeling discussion, consider items that are
The following display shows the total number of drivers killed in each month from 1969 through 1984 in the United Kingdom (UK).

1. Summarize the overall pattern of driver deaths in 1-2 sentences, based on the above display.

On January 31, 1983, compulsory wearing of seatbelts became law in the UK. The display from the previous screen is repeated, here, but now with a red vertical line drawn when the law went into effect.

2. Do you think the display above is effective as evidence for the claim "There were fewer deaths after the seatbelt law went into effect"? Why or why not?

3. Do you think the display above is effective as evidence for the claim "The seatbelt law saves lives"? Why or why not?

Figure 1.5: Sample item cluster, Seatbelt Law.
Table 1.1: A framework for modeling item cluster-based assessments

<table>
<thead>
<tr>
<th>Modeled outcome</th>
<th>Item cluster dependencies</th>
<th>Not interpreted</th>
<th>Interpreted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster score</td>
<td>Polytomous IRT model</td>
<td>N/A</td>
<td>Theory-based models; Saturated bundle model</td>
</tr>
<tr>
<td>Vector of items scores</td>
<td>Testlet model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.4 Approaches to modeling LID

If LID is detected or if there is an a priori expectation of LID, the local independence assumption can be relaxed by extending the statistical model to incorporate the hypothesized LID. As the focus of this paper is on instruments that contain item clusters, we focus on models that are used when there is some knowledge or expectation of LID among these sets of items. Table 1.1 shows three general approaches organized into a framework in the form of a $2 \times 2$ table with one empty cell. Each dimension of the table represents a fundamental decision that a modeler must make about a cluster-based assessment before analyzing the score data: (1) whether the modeled outcomes are the cluster scores or the full vector of scores and (2) whether to explicitly model the structure of the dependence or not. This decision can be made per cluster, so Table 1.1 is phrased for a single cluster. It is possible, for example, for a modeler to determine that some, but not all, clusters on an assessment will undergo the scoring convention approach and include random testlet effects for others.

When creating the data matrix of item scores for an IRT analysis, one must first decide on what the empirical item is. This is the task of choosing what produces the score that is then the outcome in the IRT model: the cluster score or the individual item scores. This is often determined at the time the scoring rules and rubrics are developed. Again, consider the Plant Species task in Figure 4. It asks the respondent to first describe and then explain. If a rubric is designed to assign a single score to the full response to both parts (a) and (b), then there would be a single empirical item. Alternately, a score could be assigned for parts (a) and (b) separately, yielding up to two empirical items. It could be broken up even further, assigning two scores each for parts (a) and (b), one for each kind of data the respondent describes. There are many other scoring schemes that could be developed for this task, and any that gives rise to more than one score makes the task an item cluster.

Scored item data is often presented in wide format with a row for each respondent and a column for each item. The cell entries are the item scores. If a modeler chooses to sum or otherwise combine clustered item scores to create a single empirical item, the wide-form data
matrix will have fewer columns as the columns representing items with a given cluster will be collapsed together. When this is done, the cluster score crossed with the person becomes the unit of analysis; this is shown in the top row of Table 1.1. If the scores for the designed items are used as originally assigned, then the designed items are also the empirical items, and the cluster score vector is the unit of analysis (the bottom row Table 1.1).

The second decision, corresponding to the columns of Table 1.1, that must be considered is whether the nature of the dependence among items within a cluster is of consequence to the data analysis. If it is not, then it does not need to be interpreted. However, if modeler has or wants to test a theory about the structure of the dependence, then an approach which explicitly models the dependence should be used. Note that it is not possible to incorporate any dependency structure into a model if the cluster score vector is the modeled outcome which is why the top right cell is empty.

The answers to these two binary questions lead to three approaches labeled in the cells of Table 1.1. Starting in the top left corner, for the cases in which someone wants to have the item cluster as the empirical item, then a scoring convention approach can be used. When there is no theory of dependence structure, people often choose to model a cluster score, often the sum score of the clustered items\(^8\). Moving down, testlet approaches should be used when the modeler wants to have the designed item scores crossed with persons remain the units of analysis but is not concerned with accounting for any specific dependence structure in the modeling. In the testlet approach, a random testlet effect is included in the model as a way to relax the local dependence assumption, but it does not make any assumptions or restrictions on the structure of the local dependence past conditional independence given the testlet latent variable. If there is a desire to account for a specific theory-based dependence structure in the modeling, then a theory-based modeling approach should be used. In the case that the modeler has no a priori theory or assumption about the nature of the dependence, the saturated bundle model can be used in which each unique response vector constitutes a different outcome (Wilson & Adams, 1995). Each of these approaches will be described in detail, with its own subsection.

The scoring convention, testlet, and saturated bundle approaches are fairly generalized, while the theory-based approach includes many highly specialized models can that have been proposed for dealing with specific item cluster structures and will be noted. As this paper cannot cover them all, select models from existing literature have been chosen that we feel can be helpful when making initial decisions about modeling.

For simplicity, each approach will be considered only in the case for which response data is unidimensional first. LID is both conceptually and mathematically the presence of another dimension, so many of the models discussed are technically multidimensional models. In fact, many tests for unidimensionality parallel test for LID. However, we limit the initial discussion to essentially unidimensional (Nandakumar, 1991) contexts in which there is only one substantive dimension measured by the instrument that produced the data. Extensions to multidimensional settings and other considerations to take into account when making

\(^8^\)Note there are some implicit assumptions being made when scores are simply summed, some of which are described in detail, later.
modeling decisions are discussed after the presentation of each of the three core approaches.

1.4.1 Testlet models

In the contemporary IRT literature, most IRT models are already random effect models in that $\theta_p$, person $p$'s location on the scale being measured by an instrument, is included in the model as a random intercept. The models discussed in this section are classified as testlet random effect models because additional random effects are included in order to account for the LID expected among items belonging to item clusters (a.k.a. testlets). In fact, random effects are often regarded as the “basic modeling tool for correlated data” (Tuerlinckx & de Boeck, 2004, p. 290) outside of psychometrics. Technically, these testlet models are multidimensional based on the clear observation that a testlet model contains multiple random effects. However, we continue to limit the discussion to the unidimensional context when there is one interpretive or substantial dimension that is of interest.

This discussion of random effect approaches for item cluster-based assessments will start with the bifactor model (Holzinger & Swineford, 1937; Gibbons & Hedeker, 1992), not because it is the most popular random effect approach for testlets, but because many of the widely-used approaches (Bradlow et al., 1999; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005) have been shown to be special cases of the bifactor models for categorical data (Li, Bolt, & Fu, 2006; Rijmen, 2009, 2010). Use of the bifactor model is widespread in the factor analytic tradition, but not as popular in the IRT literature (Rijmen, 2010). Both literatures, however, continue to benefit from the work of others showing equivalencies and close relationships among factor analytic and IRT methods.

In applying the bifactor model to dichotomous response data, a random testlet effect is included for each item cluster. These testlet effects are often assumed to be independent (orthogonal) to the dimension of interest ($\theta$, referred to as the general factor in factor analysis) and to all other testlet effects in the model ($\gamma$s, referred to as the specific factors in factor analysis). So, we have for item $i$ belonging to cluster $c$:

$$\Pr(Y_{ip} = 1|\theta_p, \gamma_{cp}, \alpha_i, \alpha_{ci}) = \frac{\exp(\alpha_i\theta_p - \delta_i + \alpha_{ci}\gamma_{cp})}{1 + \exp(\alpha_i\theta_p - \delta_i + \alpha_{ci}\gamma_{cp})}$$

(1.1)

where $\delta_i$ is the item intercept parameter, $\alpha_i$ is the loading of item $i$ on $\theta_p$, $\alpha_{ci}$ is the loading of item $i$ on the item cluster effect for cluster $c$, and $\gamma_{cp}$ is the random effect for person $p$ associated with cluster $c$. In this convention, the magnitude of the loadings from the $\theta_p$ dimension are $\alpha$s with a single subscript ($\alpha_i$), and those from the testlet dimensions have two subscripts for both the cluster and the item ($\alpha_{ci}$). For an item $i'$ that does not belong to cluster $c$, $\alpha_{ci'} = 0$. This is illustrated in Figure 1.6. The assumed independence of the random effects $\theta_p$ and $\gamma_{cp}$ is represented in the path diagram in Figure 1.6 by the absences of arcs connecting the random effects (the ovals). That there is no arrow representing, for example, $\alpha_{14}$ means that item 4 is not part of item cluster 1 and thus that $\alpha_{14} = 0$. For model identification, some parameters must be constrained. It is standard in bifactor analysis to constrain the variances of the random effects to one, $\text{var}(\theta_p) = 1$ and $\text{var}(\theta_{cp}) = 1$ for all clusters $c$. When this is done, you will obtain estimates of the loadings (the $\alpha$s).
To obtain the 2-parameter logistic (2PL) testlet model (Bradlow et al., 1999; Wainer et al., 2007), some constraints are placed on the loadings (the αs in Equation 1.1). Specifically, the loadings on the general dimension are constrained to be proportional to the loadings on the specific dimension within an item cluster:

$$\beta_c = \frac{\alpha_i}{\alpha_{ci}} = \frac{\alpha_{i'}}{\alpha_{c'i'}}$$  \hspace{1cm} (1.2)

for items $i$ and $i'$ belonging to cluster $c$. When this is done, the numerator of Equation 1.1 can be written to incorporate this proportionality constraint, $\beta_c$, as $\exp(\alpha_i(\theta_p) + \beta_c \gamma_{cp}) - \delta_i$.

In the 1-parameter logistic (1PL) testlet model, loadings are further constrained. Here, all loadings within dimension, general or specific, are set equal to each other. For items $i$ and $i'$ in testlet $c$:

$$\alpha_i = \alpha_{i'} \text{ and } \alpha_{ci} = \alpha_{c'i'}.$$  \hspace{1cm} (1.3)

For the Rasch testlet model (Wang & Wilson, 2005), all loadings are set equal to 1. When this is done, the variances of the random effects are now estimated (instead of constrained to 1, as done previously). The equation for the Rasch testlet model, then, does not require any αs, shown below in Equation 1.4.

$$\Pr(Y_i = 1|\theta_p, \gamma_{cp}) = \frac{\exp(\theta_p - \delta_i + \gamma_{cp})}{1 + \exp(\theta_p - \delta_i + \gamma_{cp})}$$  \hspace{1cm} (1.4)

Another possible approach for random effect modeling of testlet effects is to use a higher-order model. A second-order model is shown in Figure 1.7. In this formulation, the higher-order $\theta$ acts on the responses through the $\gamma$ dimensions, the testlet effects. The dimension of interest only acts on the responses indirectly.
As the random effect approach to modeling LID is exactly a multidimensional model, extending to multiple interpretive dimensions may not be so much an extension as a modification. In running a model with only one interpretive dimension, the covariance matrix for all random effects must be fixed to have off-diagonal entries of only 0. When more than one interpretive dimension is included, the interpretive dimensions should be allowed to covary. Thus, care must be taken in whatever estimation software that is being used to make sure that covariances between substantive dimensions are estimated, and covariances between any of the testlet effects and any other dimension are anchored to 0. Which dimensions that are allowed to covary should be based upon which are deemed to be interpretive and which are not by both the aims of the data analysis and the meaningfulness of their interpretation, as discussed above. A covariance matrix for a test with one interpretive dimension and two testlet dimensions is provided below. This is the case depicted in the multiplication assessment discussed earlier, in Figure 1.2.

\[
\begin{bmatrix}
\text{var}(\theta) & 0 & 0 \\
0 & \text{var}(\gamma_1) & 0 \\
0 & 0 & \text{var}(\gamma_2)
\end{bmatrix}
\]  

(1.5)

Interpretation of item parameters in random effect testlet models follows the same convention as in any multidimensional Rasch family IRT model. Interpretational issues specific to these testlet models arise in the interpretation of person locations within the interpretive and/or incidental dimensions due to the orthogonality constraints. Consider again Figure 1.2. If this was a path diagram for a testlet model, all correlations among dimensions would be constrained to zero, given the absence of arcs connecting them. Then, the interpretation of the multiply parameter would be a student’s multiplication ability orthogonal to her area and rugby abilities. It is arguably not a meaningful to interpret only the portion of
multiplication ability that is completely independent from area knowledge, or vice versa. It would be meaningful, on the other hand, to interpret a multiplication dimension that removes rugby knowledge – in this circumstance, we do not care about rugby knowledge, nor with how it interacts with other dimensions. In this case, it may make the most interpretational sense to allow non-zero correlation between \( multiply \) and \( area \), but constrain \( rugby \) to be orthogonal to all others. This situation is shown in Figure 1.8 with the addition of a dotted arc between \( multiply \) and \( area \). The associated covariance matrix would be as follows. If there were instead two interpretive dimensions, the covariance matrix would take the following form.

\[
\begin{pmatrix}
\text{var}(\theta) & \text{cov}(\theta, \gamma_1) & 0 \\
\text{cov}(\theta, \gamma_1) & \text{var}(\gamma_1) & 0 \\
0 & 0 & \text{var}(\gamma_2)
\end{pmatrix}
\]  

(1.6)

### 1.4.2 Bundle scoring conventions

One of the earliest suggestions for dealing with expected LID for items that share stimulus material is to change the unit of analysis altogether. Wainer, Sirici, and Thissen (1991) argue that if an interrelated and integrated group of items is to be administered as a single unit, it should be analyzed in that way. Instead of feeding the measurement model scores for each scored item on the test, the scores within an item cluster are aggregated and a total cluster score is used. These cluster scores are usually polytomous (even if the base items are dichotomous), commonly obtained by simply summing the scores of the items that make up the item cluster. This approach, then, does not require the assumption of local item independence, but instead local bundle independence. The empirical item is the item
cluster as it receives the score. This shortens the (empirical) test, but allows for the direct applications of non-specialized polytomous IRT models such as the nominal model (Bock, 1972), graded response model (Samejima, 1969), partial credit model (Masters, 1982), or rating scale model (Andrich, 1978).

This is not truly modeling the LID, though it has been claimed that changing the scoring unit is “explicitly modeling the testlet structure” (Wainer, 1995, p. 171): we contend that this is not the case. It is indeed a clever way to meet the independence assumption of IRT models, but it does so by changing the unit of analysis and thus avoiding the issue. Though the testing experience is unchanged for the respondent, the information it provides to the model is reduced since the number of empirical item scores (observations, essentially) used in the model estimation is reduced.

If a scoring convention approach is used, there are nontrivial decisions to be made about how to obtain the aggregate score. There are three common conventions to summarizing or combining multiple responses to create the unit of analysis.

- **Sum score convention.** As mentioned above, one of the most common approaches is to score each item within the cluster and then to simply sum them. In doing this, all response vectors that yield the same sum are essentially assumed to be equivalent. Consider an item cluster that consists of three dichotomous items. There are three different response vectors that yield a sum of two: (011), (101), and (110). Using this convention, each of these response vectors is assumed to mean the same thing in terms of the construct being measured. Any polytomous IRT model can then be applied. The formulation for Master’s partial credit model (PCM) is shown below for the sum score

\[ W_c = \sum_{k \in K} Y_k \]

where \( K \) is the vector of the indices of the items in cluster \( c \) and \( Y_k \) is the score on item \( k \). The probability of a score \( w \) is then

\[
\Pr(W_c = w; \delta_c | \theta_p) = \frac{\exp \sum_{j=0}^{w} (\theta_p - \delta_{cj})}{\sum_{l=0}^{L} \exp \sum_{j=0}^{l} (\theta_p - \delta_{cj})}
\]

(1.7)

where \( \theta_p \) represents person \( p \)’s location on the dimension of interest, \( \delta_c = (\delta_{c1}, \delta_{c2}, \ldots, \delta_{cL}) \) is the vector of item cluster parameters, and \( L \) is the number of possible sum scores for cluster \( c \). For the baseline case, when \( W_c = 0 \), we must define \( \sum_{j=0}^{0} (\theta_p - \delta_{cj}) \equiv 0 \).

- **Response vector convention.** If one believes that different response vectors mean something different in terms of the construct being measured, the response vector pattern itself can be the unit of analysis. In this convention, a sum score is still calculated, but a different parameter for each response pattern is estimated. This convention has been referred to as the bundle approach (Rosenbaum, 1988; Wilson & Adams, 1995). The outcome is the entire response vector \( \mathbf{V}_c \), each unique vector constituting a unique singular outcome. Let \( w(\mathbf{v}) \) be the total score of the cluster (the sum of vector \( \mathbf{V}_c \)) and \( \xi \) be the vector of parameters associated with cluster \( c \). The following expression is the probability of a response vector \( \mathbf{v} \) on cluster \( c \), given a person location of \( \theta_p \). Note that \( v \) is used as an index to indicate a given response vector from the \( 2^M \) unique response vectors of the item cluster of size \( M \) and \( c = 1, 2, \ldots, M \).
identifies the vector.

\[
\Pr(V_c = v; \xi | \theta_p) = \frac{\exp(\theta_p \cdot w(v) - \delta_{cv})}{\sum_{m=1}^{2^M} \exp(\theta_p \cdot w(v) - \delta_{cm})}
\]

(1.8)

There are many special cases of this model including ones that group response vectors together corresponding to specific LID hypotheses, and, in particular, ones that incorporate item parameters. The most general form of this approach, shown in Equation 1.8, is termed the saturated bundle model. Wilson and Adams (1995) offer a formulation of this model in the Multidimensional Random Coefficients Multinomial Logit (MRCML) framework (Adams, Wilson, & Wang, 1997) which allows this model to be run without any alteration of the data matrix by specifying specialized design and scoring matrices.

- **Holistic scoring convention.** One way to reduce the complication of combining multiple scores into a single score is to assign a single score in the first place. Holistic scoring of item clusters is simply the case where all responses are considered simultaneously and are scored based on a rubric for the combined responses. This type of scoring is often decided upon in the item design, often before any modeling decisions are made. Thus, it has not been labeled or discussed in the modeling literature. For example, the \(2^3 = 8\) possible response vectors of the cluster described above could be ordered and each receive a unique score from 0-7. In particular, multi-select multiple choice item types could be scored holistically as a single empirical item if each possible combination of response choices is considered in turn. However, as previously discussed in relation to Figure 1.3 item 2, each option could be scored as if they were “True/False” items.

Though it has its limitations, the scoring convention approach may be appropriate in certain situations. In the case of item clusters, it was the test designer’s intention that the set of items form a sort-of unit. If that unit-ness is more important than the individual items for score interpretation, it should be investigated whether modeling the unit is empirically supported (by, for example, comparing model fit). Wainer and Kiely (1987) advocate making this determination – the item or the cluster as the unit of analysis – empirically. Clustered items can be treated as independent if the fit is good enough, but if not, the response vectors should be collapsed into a single score. Again, note that this approach is not helpful for instruments in early phases of an iterative test design approach. Deciding which items within a cluster to alter or remove is difficult when scores are collapsed.

### 1.4.3 Theory-based models

When the nature of the dependence is deemed an important thing to model, fixed effect models are a likely candidate as they can do this explicitly. Hoskens and de Boeck (1997) consider two types of dependencies: **ordering** and **combination**. Combination dependencies are most relevant to the context of item clusters as described here as they are symmetrical,
independent of ordering, and come about from the shared stimulus material of item clusters. Order effects, on the other hand, may be present within item clusters so models dealing with this type of dependency are discussed in this section. As Hoskens and de Boeck (1997) discuss, order effects among items may arise from the item order on the test, historical ordering (the order in which students are exposed to material), or conceptual ordering (a developmental order). Wainer and Lewis (1990) actually make a case to use testlets to reduce item ordering effects across a test, especially for computerized adaptive testing, because on an item-cluster based test, ordering effects will be localized to the cluster (though cluster ordering effects may still be an issue). The general approaches for fixed effect modeling of item cluster effects follow.

- **Cluster covariates.** One potential approach to building a model for an item-cluster based assessment is to include indicator covariates for each cluster. The linear logistic test model (LLTM; Fischer, 1973; an extension of the Rasch model) could be used to estimate overall cluster difficulties. The LLTM constrains item difficulties to a linear combination of item characteristics. Including a single characteristic for each item, a cluster membership indicator, would then produce an estimate for the average difficulty of the items in that cluster. Alternately, a modeler could also include item indicators (as is done in the traditional Rasch model) along with cluster indicators. This would produce an estimate for the change in difficulty for each item belonging to a given item cluster.

- **Interaction terms.** The relationship between two items within an item cluster can be modeled with a multiplicative interaction term in the model. The componential models of (Embretson, 1984) and (Jannarone, 1986) that are based on decomposing some complex cognitive task (usually in the form of an item cluster) into its more basic elements in that they include interaction terms to model response patterns. The constant interaction model of Hoskens and de Boeck (1997) in it’s simplest form considers a pair of dichotomously scored items. The model then includes a main effect for each item as well as an interaction term that quantifies the difficulty change in one item if the other is correct. So, for two dichotomous items the possible response patterns are the set \( K = \{(0,0), (0,1), (1,0), (1,1)\} \). The probability of a response pattern \((y_1, y_2)\) is given by

\[
Pr(Y_1 = y_1, Y_2 = y_2 | \theta_p) = \frac{\exp[y_1(\theta_p - \delta_1) + y_2(\theta_p - \delta_2) + y_1y_2 \cdot \delta_{12}]}{\sum_k \exp[k_1(\theta_p - \delta_1) + k_2(\theta_p - \delta_2) + k_1k_2 \cdot \delta_{12}]} \quad (1.9)
\]

where \( \theta_p \) represents person \( p \)'s location on the dimension of interest, \( \delta_1 \) and \( \delta_2 \) are the item main effects, and \( \delta_{12} \) is the interaction effect. If \( \delta_{12} = 0 \), then this model reduces to the standard, dichotomous Rasch model. A positive value of \( \delta_{12} \) means that getting

---

9Interactions among more than two items can be modeled with higher-order interaction terms as well.
10The term “constant” is used because the interaction effect is constant across examinees. Hoskens and de Boeck (1997) also formulate a model for dimension-dependent interactions which are allowed to vary across examinees.
11The original formulation of Hoskens and de Boeck (1997) is in the 2PL with a discrimination parameter. It has been constrained here to be a Rasch family model.
both items correct is more difficult than would be expected if the items were solved independently. A negative $\delta_{12}$ indicates the opposite, that getting both items correct is easier than would be expected.

Other examples of interaction models are discussed in other works (Wilson & Adams, 1995; Tuerlinckx & de Boeck, 2004), but the Hoskens and de Boeck (1997) constant interaction model is illustrative of the approach.

- **Lagged dependence.** Another way to incorporate response dependencies into a model is to condition the probability of an item responses on one or more previous responses. This approach is (usually) accompanied by the implicit assumption that items are strictly ordered. In some cases, previous responses are used as covariates in the linear model, as in Tuerlinckx and de Boeck (2004). The lagged dependence model of Ackerman and Spray (1986) illustrates this general approach for the case where all items are dichotomous and each response is dependent only on the response immediately previous to it—thus, it is a one-item lag model. So, for any item $i$, the response $Y_i$ is conditioned on the response $Y_{i-1}$ to item $i - 1$ (excepting $i = 1$, the first item). Thus, there are four probabilities for every item ($i \neq 1$) to model: $P(Y_i = 0|Y_{i-1} = 0)$, $P(Y_i = 0|Y_{i-1} = 1)$, $P(Y_i = 1|Y_{i-1} = 0)$, and $P(Y_i = 1|Y_{i-1} = 1)$. Those associated with response vectors in which the $i^{th}$ response differs from the $i-1^{th}$, $P(Y_i = 1|Y_{i-1} = 0)$ and $P(Y_i = 1|Y_{i-1} = 0)$, are transition probabilities\footnote{Ackerman and Spray (1986) note this is a Markov chain with non-stationary transition probabilities.} and notated by $\alpha_{ij}*$ and $\beta_{ij}*$, respectively. This set of probabilities for item $i$ are summarized in the following table.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i - 1$</td>
<td>0</td>
<td>$1 - \alpha_{ij}^*$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$\beta_{ij}^*$</td>
</tr>
</tbody>
</table>

In summary, the fixed-effects approach is most appropriate when the analyst has a grounded theory to describe the nature of the dependence, and especially when the interpretation of a parameter can answer a research question. The selection of fixed effects to include in the model then is based on that theory and the research question that motivates the analysis.

1.4.4 Issues

Some of the common issues in IRT analyses in general deserve special consideration when item clusters are present on an instrument. Differential functioning, multidimensionality, and sparseness are discussed in turn in this subsection.

**Differential functioning.** If accounting for LID by collapsing multiple items into a single score using a scoring convention, we lose item-level parameter interpretations and fit
statistics. The ramifications of this loss of item-level estimates and information is different based upon the stage of development of the instrument. For thoroughly vetted and stable instruments with unidimensional item clusters, this may not be a big deal. For those in initial stages of development, however, considering the entire item cluster as the unit of analysis may be less than helpful. Ill-fitting items may be undetectable if other items within the cluster fit well enough to mask the poor fit of a “misbehaving” item or, possibly, if another item within the cluster “misbehaves” in the opposite direction, canceling it out. Alternatively, if poor fit statistics are found at the cluster-level, the way forward is not clear. The approach does not allow for investigating which item(s) in the cluster are misbehaving and thus should be altered or removed. Should the entire item cluster be thrown out for potentially one bad item? Additionally, since the designed item is no longer the unit of analysis, differential item functioning (DIF) cannot be investigated; differential cluster functioning (DCF), on the other hand, can be.

It should be noted here, however, that though DIF can be masked in the cluster aggregate, there is more statistical power in DCF (Wainer, 1995). If differential functioning among items is balanced within a cluster, this does in fact reduce the advantages and disadvantages overall between groups. For this reason, Wainer (1995) argues that small amounts of DIF that cancel within an item cluster (leading to a conclusion of no DCF) are “perfectly acceptable.” However, if changes to a cluster are to be made, it is important to uncover which items exhibit DIF and in which direction. It would be bad practice to alter the cluster by removing an item that affects the overall balance of differential functioning. Further, administrators and users of the test may decide to drop items from the test administration or from the analysis for any number of reasons. Unless they understand this point that finding no DCF does not necessarily mean there is no DIF, they risk losing this balance when item clusters are altered. There would need to be assurances in place that the item cluster is administered and analyzed in a consistent way. In our opinion, any indication of DIF should be investigated and attempts made to remedy it. Identified poorly-behaving items should not be included on a test or in an item bank.

**Missingness and sparseness.** One of the advantages of taking an IRT approach is its handling of missing data. When assigning a score to a cluster of items, dealing with missingness within a cluster is not straightforward. In general, IRT approaches can treat missing data by “simply” not including that item for that person in the estimation. In the CTT approach, missing responses must almost always be scored as something in particular, usually 0. This can also be done within the IRT framework, but coding missing responses as 0s makes the implicit assumption that missing is equivalent to incorrect. The validity of this assumption should be examined within the context of the test – what it is measuring and what the scores will be used for. When a missing response occurs within an item cluster, how should that be reflected in the aggregate score? In any of the conventions outlined above – sum score, response vector, or holistic scoring – the cluster score could be considered missing if any of the item responses are missing. In that case, all information that could have been used from the responses that were present is lost.

If the sum score convention is used, the missing scores could be coded as 0s in an attempt to retain more information for model estimation. In doing this, however, note that the
assumption that missing is identical to incorrect is still being made. If the response vector
convention is used (e.g. the saturated bundle model), including a “missing” category will
increase the number of possible response vectors as the number of possible scores on each item
within the cluster is increased by 1. Consider an item cluster containing three dichotomous
items. With no missingness, there are $2^3 = 8$ possible response vectors. By allowing missing
responses on each of the component items, there would be $3^3 = 27$ possible response vectors.
Though 27 may be a manageable number, there are other cases in which the increase may
seem more drastic. Consider an item cluster with three polytomous items, each with four
score categories. The number of possible response categories increases from $4^3 = 64$ to
$5^3 = 125$. Note that there is an implicit assumption in this solution that the pattern of
missingness itself is information about the construct being measured. It may also complicate
interpretations of cluster parameters as there is summing of scores involved, and missing
responses will be treated as 0s in assigning the cluster score.

Whether or not missingness is an issue when using the saturated bundle model, the
test designers and analysts should always consider the number of possible response vectors
within a cluster. To ensure a large enough cell size for all score patterns large sample sizes
are needed when there are a lot of possible response vectors. Sparseness in the data matrix
(caused by few respondents having a given score pattern within a bundle) can bias item
parameter estimation. Note, however, that if detecting LID is the aim of the analysis and
parameter estimation is not important, sample size and sparseness issues do not matter as
much (Wang et al., 2005).

In terms of lagged dependence models that condition on previous scores or score vectors,
choices need to be made for the handling of missing data as well. Missing responses could be
treated as “incorrect,” treated as their own score category, or thrown out. Note that decisions
about handling missingness could subtly change the interpretations of parameters, so this
should be taken into account when making decisions about how to treat missing responses
(Mislevy & Wu, 1996; Patz & Junker, 1999).

**Multidimensionality.** Further complications may arise using for some approaches when
item clusters are multidimensional. Consider, as an illustrative example, a 15-item test
containing a 5-item cluster and two substantive dimensions $\theta_1$ and $\theta_2$. Assume three of
the clustered items load on $\theta_1$ and the other two on $\theta_2$ and that all items are dichotomous.
Figure 1.9 shows a path diagram for this test. This figure also represents the testlet approach
for these data. In this case, a multidimensional IRT model can be used with constraints on
the covariance matrix. Because $\theta_1$ and $\theta_2$ are allowed a nonzero covariance, it is not a bifactor
or testlet model, but because the conditional independence is modeled with a random testlet
effect, we can call this the “testlet-like” approach. In the language used previously, this is a
model with two interpretive and one incidental dimension.

Extensions of fixed effect, theory-based models from unidimensional to multidimensional
is straightforward for the models discussed here as they all classify as item-explanatory
Rasch family models. This is discussed in Rijmen and Briggs (2004). In particular, the
multidimensional extension of the constant interaction model is considered by Hoskens and
de Boeck (2001).
Figure 1.9: Testlet-like approach for a two-dimensional fifteen-item test containing a five-item cluster.

Though the testlet and theory-based approaches have straightforward extensions to multidimensional contexts, the scoring convention approach does not. Though the scoring convention approach is fairly generalized, care must be taken in applying it especially when item clusters are multidimensional. Consider again Figure 1.9. If we instead model the sum score of the item cluster, this becomes an 11-item test with the new (empirical) item formed by collapsing the cluster having possible scores of 0, 1, 2, 3, 4, or 5. This situation is illustrated in the top panel of Figure 1.10. In doing this, we end up with a slight mismatch between the theoretical data generating model and the reality of the assessment design in that it equivocates response patterns that are not theoretically equivalent if the model truly is two-dimensional. Though the model is identified and will produce item parameter estimates, the interpretation of those parameters may be problematic. In this model, it is the assumption that a sum score of, say, 2 is equivalent, no matter the response pattern. This model assumes that any score of 2 supplies the same information to the model within both dimensions $\theta_1$ and $\theta_2$. In the most extreme case, that score of 2 could have been concentrated completely inside of the $\theta_1$ dimension, with a score of 0 in the other. Yet, using this approach treats all scores of 2 identically though different response patterns supply distinctly different information within each dimension.

Of course, the scoring convention approach could be applied here, but with some care around how the scores are produced. The middle panel of Figure 1.10 shows the situation of creating two scores from the item bundle by summing the three items that load on $\theta_1$ and the two that load on $\theta_2$ separately. However, using this approach, we still have two clustered scores. One remedy for this might be to include a testlet effect for those two empirical items. This is shown in the bottom panel of Figure 1.10. There are models that can be built, as this one, that blends the approaches. It might be best in these multidimensional situations, however, to not use the sum score approach. Modeling the response vector instead of the
Figure 1.10: Three sum score approach choices for a two-dimensional test containing a multidimensional cluster.
**Interpretational Issues.** For models that incorporate fixed parameters to account for LID, care must be taken for the interpretation of other parameter estimates in the model, as they usually must include conditioning on the LID parameters or previous responses. In many cases, the straightforward item difficulty parameters that are produced from Rasch family models can no longer be interpreted as such. For lagged dependence (a.k.a. recursive) models that incorporate terms for previous responses, the item parameter can often be interpreted as a “learning parameter” (see Tuerlinckx & de Boeck, 2004) as it is no longer an “item difficulty.” Note that this learning parameter is not learning in the traditional sense that a respondent moves up in construct being measured, but instead in a local sense; it is a change in $\theta$ due to exposure to a previous item.

The interaction models carry the same caution for an item main effects interpretation—it is *not* the item difficulty in the usual sense, but the item difficulty for the particular case when the other item in the pair has a score of 0. The formulations presented above were for dichotomous item pairs for simplicity, but higher-order dependence terms and models for polytomous data yield more complicated interpretations for parameters. Further, when the LID modeled is based upon item ordering (as in lagged dependence models), care must be taken to ensure that order effects can safely be assumed to be constant across persons. This may not hold in situations in which respondents are allowed to change earlier responses during test administration.

### 1.5 Conclusion

This paper aims to serve as a general framework to guide modeling decisions for item cluster-based assessments. Terminology around item clusters was formalized, methods for detecting LID were discussed, as were early recommendations and methods for dealing with LID. The three general approaches were presented for accounting for LID due to common stimulus material: random testlet effect models, bundle scoring conventions, and theory-based (often fixed-effect) models. Each of these were broken down into more specific conventions or methods, where appropriate, and some modeling issues related to these approaches were discussed.

As estimation capacity grows with the availability of flexible modeling software and advancements in statistical modeling in general, the old recommendation to avoid LID need no longer apply. There are validity and practical efficiency concerns that make item clusters (and the resulting LID) an attractive option when designing a test. LID no longer needs to be considered a deficiency of test data, but now just a characteristic of it. The test design doesn’t need to be restricted by what could once be considered a shortcoming.

There is no “one size fits all” modeling approach for assessments that contain item clusters; modeling decisions should be guided by the research question (the reason for the data analysis), the theory of the phenomenon under study (the construct map), and the structure of the item clusters. The considerations discussed in relation to the different modeling
approaches in this framework should help an analyst decide which approach to take.
There has been in recent years an increasing demand for educational assessments focused not solely on factual knowledge and algorithms, as they traditionally have, but also on more complex constructs such as critical thinking or so-called “non-cognitive” skills. Alongside this call for assessment of complex constructs and skills, there has also been growing emphasis on preparing college-ready students in K-12 education. A very significant example is that the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) explicitly names college (and career) readiness as its goal. These circumstances have motivated a call for increased rigor and explicit focus on college readiness (CR) during K-12 schooling and in turn the development of valid, reliable, and psychometrically sound measures of CR. Though the ACT and SAT have long been considered indices of CR, the content of these entrance exams is very limited and coarse, and they cannot be considered complete measures of CR. “College readiness” is most often defined multidimensionally, cross-cutting all academic domains and including many complex critical thinking and non-cognitive skills (Conley, 2007; Wiley, Wyatt, & Camara, 2011). Different levels of coarseness in these definitions are helpful to different stakeholders in education. Whereas ACT and SAT score reports may be helpful to post-secondary institutions assessing the overall academic readiness of an applicant, those same score reports are likely not helpful to a classroom teacher in planning instruction and intervention a student. These traditional measures of college readiness are difficult to use in a formative way for students still in high school. Finer, more specific definitions of college readiness in particular domains are needed for formative purposes.

In a response to this, the Berkeley Evaluation and Assessment Research (BEAR) Center has partnered with the National Math and Science Initiative (NMSI) to develop domain-specific measures of college readiness in mathematics called the Critical Reasoning for College Readiness (CR4CR) Assessments. One strand of this work is described in this paper, the development of three constructs and accompanying assessment items to measure data-based and statistical critical reasoning.
Laying the groundwork for the CR4CR Assessment involved analyzing mathematics and statistics standards documents\textsuperscript{1}, a survey of Advanced Placement (AP) coursework\textsuperscript{2}, and a survey of standards in applied disciplines\textsuperscript{3}.

In the *Standards for Success* research (Conley, 2005), college professors, especially those of the natural and social sciences, indicated that prerequisite knowledge of basic statistical concepts and techniques plays an important role in entry-level courses. Conley reports that students who fail initial coursework because they lack a prerequisite skill will avoid majors in that area of study, “closing off entire avenues of the curriculum and career pathways” (p. 114). It is interesting to note, however, that Conley observed that statistical skills are less important in entry-level coursework for mathematics majors while it is imperative for success in other majors in the natural sciences (e.g., biology, ecology, physics), social sciences (e.g., economics, psychology, journalism), professional degrees (e.g., nursing), and even in some parts of the humanities (e.g., history).

We have defined three constructs within the field of critical reasoning in data-based and statistical contexts and have developed test items to measure them.

1. Linking Data to a Claim (LDC) describes the sophistication of a student’s argument for (or against) the use of particular data to make a data-based claim.

2. Meta-Representational Competence (MRC) is development through critically thinking about the use of data displays to answer questions.

3. Formal Inference (FoI) describes student understanding of statistical and practical significance.

The CR4CR assessment forms have items aligned to each of these constructs, which are elements of but do not fully define critical statistical reasoning. Though they are defined distinctly, we recognize that these constructs are related to each other; they are not independent. The test is being designed for both formative and summative use. Especially for teachers using the assessment results formatively, it will be important for the score report to give information on student performance within each construct, instead of a single, composite score across all constructs. For these reasons, they are treated as distinct dimensions in


\textsuperscript{2}AP courses were surveyed as they can be considered reflective of entry-level college coursework and provide an easily accessible source of widely-used documents. Twenty-nine of the 38 courses were surveyed and statistical skills were either referenced in the course objectives or included in sample items in all but three of the surveyed courses.

\textsuperscript{3}College Board Standards for College Success\textsuperscript{TM} for Science (Science Standards Advisory Committee, 2009), the Next Generation Science Standards (NGSS Lead States, 2013), and the KSUS Standards in all other disciplines (English, Natural Sciences, Social Sciences, Second Languages, and Arts)
the measurement model, as opposed to running a unidimensional model that would produce a single, composite score. Further, there are many constructs in the CR4CR framework, more than the three described in this paper. In the field testing that produced the data analyzed herein and in future data collections, test forms needed to be structured so that they could be completed by students in an hour or less. It will not be feasible to include many items on every construct on every form. As such, a student may respond to as few as three items on a construct. Understanding how these constructs are correlated strengthens the measurement, especially in situations like this. A multidimensional model can use responses to items aligned to other constructs to provide some information on correlated constructs. This paper uses CR4CR field test data and the results presented herein are to be used to further revise and refine construct definitions, assessment items, and scoring procedures.

The methods used for test development follow the BEAR Assessment System (BAS; Wilson, 2005), a principled approach to the development of measurements. The first section lays out our framework for college-ready critical reasoning in statistics and defines the three constructs in the form of construct maps, the second section describes the assessment items developed for the CR4CR assessment and provides an example item cluster, next the quantitative methods are described and results presented, and finally there is a discussion of the results and directions for future research.

2.1 A framework for critical reasoning for college readiness in statistics

Combining the information garnered from the mathematics and statistics standards documents, the AP course description survey, and the college readiness literature, we propose the
following Framework for College-Ready Statistical Thinking. Existing standards and frameworks informed this framework, but none on its own provided both the breadth and depth we required for the statistical thinking constructs that the CR4CR Assessment targets. Each, though, illuminated elements of statistical thinking that are incorporated into this framework. Specifically, we considered the contexts of (1) reading and critiquing reports on data or statistical investigations, and (2) planning a statistical investigation, two common tasks for which we posit parallel thinking processes. The work of Conley, Drummond, de Gonzalez, Rooseboom, and Stout (2011) reports that the standard “Evaluate reports on data” of the CCSS-M was rated as one of the most important standards in the statistics strand by instructors of first-year courses. Lajoie and Romberg (1998) state that K-12 students should learn to both critique as well as produce reports of statistical results “as required” for their future roles as consumers and produces in society. This is all the more relevant for their specific roles as college/university students for those who choose that path. The CR4CR framework includes the societal demands for statistical literacy as the demands for college coursework include, but move beyond, those. For students to critically consider results of a data analysis—their own or another’s—there are three broad essential questions that they must be able to answer, regardless of the course content or research context:

1. What decisions need to be (were) made based on the data? (Decision Making)
2. Where did the data come from? (Producing and Selecting Data)
3. What was done to the data? (Summarizing Data)

Figure 2.1 places each of those questions at the vertex of a triangle. Although it makes sense to see each vertex as being part of a sequence of steps in a statistical investigation–get the data, analyze it, make a decision based on it—the three parts need to also be considered simultaneously. For example, it makes little sense to think about the qualities of data collection without also considering the use(s) of that data for decision making. Statistical investigation starts with its motivation - What decisions need to be made based on data? This is shown in the bottom left of Figure 2.1. From there, it moves to the top left - the selection and/or production of data - What are the data and where did it come from? Once data has been obtained, it is presented and/or summarized in a way that informs the decision to be made - What was done to the data? Often, the investigation stops here. However, some may iterate through this critical questioning process by either refining steps of the investigation or adapting them. We conceptualize critical reasoning in data-based or statistical contexts as asking oneself these essential questions and then basing an investigation (or critique of one) on the answers. Note, however, that these are very broadly defined, and the critical thinker does not stop with these. A cascade of questions will arise based on these broad, starting questions. What the follow-up questions will be is largely determined by the material presented to the student and the domain in which the statistical investigation originated.

Underlying all statistical skills is an understanding of basic probability theory, so that is represented at the center of the triangle. So far, this area is not an explicit focus of this work. Though probability theory underlies nearly all concepts in the statistics discipline, the focus is instead on how probability is applied when statistics are used in other (non-mathematical)
fields. The mathematical study of probability often removes the context of the numbers. We aim to study statistical thinking in applied contexts only. Further, the relevant aspects of probability theory will show up in the other areas. Thus, to start, constructs and items have first been developed in the three areas that make up the vertices of the triangle in Figure 2.1.

This categorization of statistical skills is not new. Scheaffer, Watkins, and Landwehr (1998) organize the statistical content they felt should be in the K-12 curriculum into these same categories with the addition of “number sense.” Their “Planning a Study and Producing Data” strand is Producing and Selecting Data, “Data Analysis” is Summarizing Data, and “Inferential Reasoning” is Decision Making. Our framework is focused on statistical reasoning, not the whole of the study of statistics. From a different perspective, the conception of critical questioning for statistical literacy (Watson & Callingham, 2003) underlies the selection and illustration of constructs included in this framework. Watson’s work is grounded in a three-tier framework for statistical literacy (Watson, 2006, Chapter 1) in which each tier describes increased sophistication in thinking about a statistical problem. This complements the construct modeling approach of BAS which also defines hierarchies of development in critical statistical literacy, though we aim to flush out more specific elements of statistical literacy and with a refined focus on college-readiness, not general citizen statistical literacy.

The three constructs in Table 2.1 are intended to be an initial step in developing the learning progression (Wilson, 2009) for college-ready statistical thinking. As this is an incomplete list, future development and research may lead to constructs being altered, split apart, combined, added, or removed. For now, no conjectures are made about the ordering of these constructs. A fully-defined learning progression would include construct maps as well as theorized ordinal relationships among the levels of the different constructs.

In general, each construct represents a respondent’s predilection to ask and answer a certain critical question. We used the idea of critical statistical questioning (Watson & Callingham, 2003) as a way to define our critical reasoning constructs and expanded upon it with more detailed questions to contextualize each construct that has been and will be defined as an element of a vertex in Figure 2.1. The critical questions associated with each of the three constructs defined in the following subsections are provided in Table 2.1. Note that most of the critical questions are written as if a respondent is reading the work of someone else. There is an analog for each question in the context of planning and conducting a statistical study, but to avoid redundancy, they are not presented here. The analog question can be formed in most cases by adding the word “should” or “will” in the appropriate place.

---

4 For instance, students who conduct formalized statistical tests will need to understand that a p-value is a statement about conditional probability.

5 We do not intend to communicate that understanding probability and interpretations of probability statements is any less important than understandings in the other three areas by omitting a theoretical structure for the center of the triangle. In fact, probabilistic intuition underlies many of the contracts within the framework sections outlined so far. We have first undertaken identifying the constructs in the other three areas. Once there exists a more detailed and complete picture of each of the vertices, a structure for college-readiness in terms of probability theory can be developed. In short, it will be difficult to define the “prior knowledge” in probability without a solid understanding of the knowledge and skills that contribute to statistical thinking.
Table 2.1: The CR4CR constructs.

<table>
<thead>
<tr>
<th>Construct name</th>
<th>Associated critical question</th>
<th>Performance description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linking Data to a Claim (LDC)</td>
<td>Are the data the right data?</td>
<td>identify and justify the link between the claim made and the data collected to reflect it</td>
</tr>
<tr>
<td>Meta-Representational Competence (MRC)</td>
<td>What does the display show and hide about the data?</td>
<td>think critically about how design decisions are related to the purpose of the data collection</td>
</tr>
<tr>
<td>Formal Inference (FoI)</td>
<td>Is the result statistically significant and important?</td>
<td>determine practical and statistical significance of the results of data analysis</td>
</tr>
</tbody>
</table>

Note that these critical questions and accompanying constructs do not focus on the rote application of statistical methods. This framework is not focused on further promoting an algorithmic approach to statistics. We focus on statistical reasoning and thinking (and not specific statistical methods) because (1) there is no “standard” methodology across the applied disciplines encountered in first-year college coursework and (2) good data-based and statistical reasoners should be prepared to deal with statistical reporting, no matter the statistical method used. Most of the reports a student will encounter will be situated in the frequentist framework, but it is likely that students may be exposed to results from Bayesian analysis (or the methods themselves) in college coursework as Bayesian methods are becoming more common in research. The CR4CR framework and constructs are defined such that they can be applied to tasks using any statistical approach. Though none of the tasks written for the CR4CR Assessment so far include any Bayesian inference statistics or procedures, the CR4CR framework has been defined broadly enough to encompass many of the foundational concepts of a Bayesian approach.

The goals for the CR4CR Assessment include both breadth and depth. Regarding breadth, we have defined multiple constructs and have more planned for the future. For depth, the focus of this paper is on those three specific constructs listed in Table 2.1. Detailed descriptions of these constructs along a developmental continuum are provided in the following subsections. In narrowing the focus to specific constructs, it is intended that the assessment results provide informative diagnostic information about student learning, insights into the structure of the identified construct, and lay the groundwork for the construction of a larger statistical thinking learning progression that also includes the relationships among the constructs involved. Each of the three constructs are treated in turn.
2.1.1 Linking Data to a Claim (LDC)

It is important to note that critical statistical thinking is not exclusively “quantitative.” Yes, quantitative statistical procedures are conducted to test hypotheses, but they are abstractions of real-world phenomena. An important task in either planning a statistical study or critically evaluating one is to justify the link between the real-world object under study and the variables or measurements that are purported to reflect what we want to know about that object. These measured variables make up the data and decisions about those data need to be justified relating to the question under investigation and the claim resulting from statistical analysis. Explicit consideration of this connection may not be a concern in traditional statistics courses when students learn statistical procedures. In these settings, it can usually be taken for granted that the data collected are the right data to test the
hypothesis. In contrast, in applied settings, this concern that the selected data truly reflect the phenomena about which a claim is made is of central importance!

The AP course survey showed that many of first-year courses expect students to design plans for and then conduct data collection. AP Biology, Chemistry, and Physics each list “justifying the selection of data” in course objectives. Students in those courses will be expected to choose data, plan collection, and execute that plan (College Board, 2015a, 2014a, 2014b). AP Environmental Science and Human Geography also stipulate that students should be given the opportunity to collect data (College Board, 2013, 2015b). Inherent in putting together these collection plans are the levels of thought embodied in LDC. These thought processes are similar whether a student is planning the data collection for a study or reading the works of others the content of both the question and the claim must be reflected in the observations selected for the purposes of the study. This may be more straightforward for tasks that investigate physical phenomena than for tasks such as the example above that investigate latent phenomena like attitudes.

The LDC construct in Figure 2.2 describes development in statistical thinking about this link between the data and the content of the research question (and thus, the content of the claim that answers that question) of a statistical investigation. Other popular frameworks for statistical investigation and statistical learning include a step for formulating a research question (e.g. Scheaffer et al., 1998), so it is important to note that the levels of this construct are conditioned on that question already having been formulated. In fact, we focus the discussion and definitions of the construct levels on the claim at the end of a statistical investigation. However, linking the data to either the question or the claim is the same critical thinking process. Note that the word “claim” could be replaced with “question” for any of the descriptors of LDC levels.

The LDC construct levels are presented in Figure 2.2; the qualitative descriptions of each level are provided on the left side with sample responses in the right-most column. The sample responses are for Miles per Gallon (MPG) item 4 provided in Figure 2.6. The construct is read from the bottom up, as that is the (metaphorical) direction of increasing sophistication within the construct. Note that each level represents a qualitatively different type of student response to an item, indicating a different level of sophistication in thinking. Students tend to develop gradually from lower levels to higher levels, and these “ideal-points” are milestones along a continuum of development: not all students will fit distinctly onto one of these points, the majority will be somewhere in-between them. All construct maps in this document follow this convention.

The LDC construct is an integration of the argumentation construct map developed by Henderson, Osborne, MacPherson, and Szu. (2014) and the SOLO taxonomy (Biggs & Collis, 1982). The SOLO taxonomy outlines understanding of some concept with increasing levels of coordination, from the ability to simply identify elements to the ability to integrate those elements. For LDC, the levels of SOLO were further disaggregated and the (Henderson et al., 2014) map provided guidance for structuring that disaggregation. Their model for argumentation is structured around the three elements of an argument (claim, evidence, and warrant), the coordination of these three elements, and then of evaluating competing arguments. In the LDC construct, the choice of data is the evidence and the link between
the data and the statistical claim is the warrant.

The bottom of the construct, LDC0, characterizes a respondent that does not identify a statistical claim. Moving up to the external level (LDC1), there are two unordered sublevels that may characterize students at this level. At LDC1A, the respondent identifies or provides a claim, but without accompanying evidence; students fail to recognize there is a question about the choice of data at all. At LDC1B, respondents may judge the appropriateness of the data collected by relying only on an external (contextually irrelevant) authority, their own beliefs about the subject at hand, or they may hold an “always or never” attitude—those who seem to think that the inclusion of a number, any number, in a claim is evidence enough for that claim, or those who trust no number. The sublevels have some qualitative distinctions, but they were found to span the same range of empirical difficulty and thus responses at either LDC1A or LDC1B both receive a numerical score of 1 when preparing the data for model estimation. In essence, it was no more difficult for a respondent to provide a weak link than to not provide a link at all in the LDC items developed for the CR4CR Assessment.

LDC2 linked is the first in which a valid and appropriate link between the data and the claim is explicitly drawn by the respondent (invalid or inappropriate links are considered “weak” as in LDC1B). However, this link is merely identified and there is no justification or statement for it as a choice. Respondents may not recognize that there is a choice of observations to represent some phenomena of interest. This recognition may be taken for granted, especially for tasks that involve physical/manifest phenomena that are easily measured. Consider question 4 of the MPG task in Figure 2.6. It may be taken for granted that average miles per gallon represents a car’s impact on the environment, but we expect higher level students to think about the bigger picture and recognize that other characteristics can affect a car’s environmental impact.

Respondents at LDC3 begin to realize that sometimes a link is justified and sometimes it is not. This is where the credibility of the data becomes a prime issue—the warrant is not only predicated on the existence of a link, but also that the data themselves are credible in this context. Of course, some data, such as randomly generated numbers, may be credible in only unusual cases, so that they may be unsuitable for a whole range of claims. However, students at LDC3 provide single-sided justifications only, focusing only on the choice that was made. Responses at the highest level, LDC4, provide more complex and balanced justifications of the choice of data to represent the phenomena of interest. They compare choices and provide reasoning in favor of the choice and against other choices. The distinction between LDC4 and LDC5 is the inclusion of a critique. LDC3 is simply an argument for (or against) a choice of data. LDC4 furthers the argument by critiquing alternate choices by contrasting them with the respondent’s choice of data. An LDC3 response is an argument that may be of the structure “It is good because . . .”. Whereas, at LDC4, the argument takes a comparative form: “It is better than the alternatives because . . .”.

42
2.1.2 Meta-Representational Competence (MRC)

Virtually all first-year coursework requires students to interpret data displayed graphically or in tables. Most AP course descriptions include in the topic outline or learning objectives a statement that students are to create their own data displays from their own data collections and/or critically evaluate presentations of data in historical documents or mass media. An evaluation of the sample items for the AP exams, however, showed that they were aimed primarily at basic reading of information from a graphic display as in the Birth Rates sample item from the AP Comparative Government & Politics exam (College Board, 2014c) shown in Figure 2.3. This is not to say that higher level critical thinking skills are not nurtured in AP classrooms, but that the items are not measuring the explicit learning objectives for the course. While much of the K-12 mathematics and statistics curricula include creating, understanding, and interpreting elements of graphical displays, mere technical knowledge of graphs is insufficient for statistical literacy. Scheaffer et al. (1998) and Bright and Friel (1998) discuss the importance of choosing a display that is appropriate given the type of data at hand. However, we contend that in addition to this, it is necessary to also consider the question at hand that gave rise to the statistical investigation-the reason for the data collection.

Figure 2.4 is the MRC construct map. Levels MRC1 through MRC5 are borrowed from the Modeling Data project (Lehrer, Jones, Kim, Pfaff, & Shinohara, 2014) though some slight adaptations and additions to the level descriptions have been made. The original
Figure 2.4: The Meta-Representational Competence (MRC) construct map.
MRC construct was written to accompany a middle school statistics curricular program (although the upper levels were seldom reached by middle-school students without a lot of scaffolding by teachers). We extend the MRC work of the Modeling Data project by refining the MRC5 level (because more qualitative differentiation from MRC4 was needed) and adding the MRC6 level. Note that levels 2, 4, and 5 each have sublevels A and B. These are levels that are related and qualitatively distinct, but have been found to be empirically indistinguishable in our studies so far. The problem context and item will generally restrict scoring into either levels A and B because all of the levels A described responses about a single data display, while the levels B describe responses that compare two or more displays. In order for a response to be scored at a sublevel B, at least two displays must be provided.

At the no competency level (MRC0), no information about the data is given. At emerging competency (MRC1), a respondent recognizes that the display represents some data, but misinterpretations of the display are made. At the next level, elementary competency (MRC2), interpretations are correct and features of displays are correctly compared, but the relationship of the choices made to display the data and the purpose of the data collection is not yet explored. Even at the level of concrete competency (MRC3) in which responses do begin to recognize that displays may highlight some aspects of the data while masking others, these observations are not explicitly tied back to the statistical question under investigation.

The novice-expert continuum developed by Conley and others for the College-readiness Performance Assessment System (C-PAS) guided the refinement of the upper levels of the MRC construct. The levels of the novice-expert continuum describe development of thinking from procedural to interpretive to creative. The MRC construct spans the procedural (MRC1 and MRC2) and interpretive (MRC3 and MRC4) levels. However, we should hope that at least some college-ready students would be emerging experts and are capable of creative approaches to producing or critiquing data displays. At the simultaneous competency (MRC4), respondents can make appropriate decisions about a data display (or critique the appropriateness of another’s). This is an instance of the “strategic thinker” on the novice-expert continuum. The defining difference between the MRC4 and MRC5 levels is the contextualization of the interpretation/observation in relation to the argument being made (or question being posed) in MRC5. MRC4 merely recognizes that different displays tell different stories.

MRC5 (display fluency) could be classified as “emerging creative thinking” in that a respondent at least recognizes that different choices about a data display exist and can weigh those choices against each other. Respondents make display decisions for their data to effectively match their argument. What was missing from the original MRC construct, however, is that, within each display format, more sophisticated choices can be made. There is more to data display production (and critique) than determining which format (e.g. histogram versus pictograph) to use.

This is not a deficiency of the original MRC construct as it was targeted at middle schoolers. However, by the start of college, we should expect student fluency with different formats to develop and improve over time. At MRC5, a student has a solid understanding and familiarity with the standard displays of discipline (e.g. histograms, pie charts, scatterplots, ... ) and can make appropriate choices regarding the parameters of that display (e.g. histogram bin size). The distinguishing feature of a student at MRC6, however, is the move beyond
the standards. Respondents acknowledge that data can be displayed selectively, and that this may or may not be misleading. Outliers might be removed to make patterns more discernible. We may select to only plot data points from a 10-year range, even though we have data over 30 years. There may be missing variables. The whole story behind a scatterplot may include a third variable (see Simpson’s paradox; Simpson, 1951). Note that student skepticism starts around MRC4, when respondents begin to consider what is not shown by a particular display. MRC5 and MRC6 continue this development of skepticism, with more complicated and sophisticated changes between considered displays.

In the Birth Rates item (Figure 2.3), students must read information from a double bar graph and make literal interpretations of that information in the given context. This is an important and useful skill, and a correct answer would be scored into MRC2 as it does not require any thought about either the form or function of the graph. From the course material survey, this type of item is illustrative of the sophistication expected of students regarding data displays. This is not to say it is a poor item, just that it will only give information about the lower end of the construct. We do note that, although the stem of the item suggests that the scientific interest here is in the relationship between the two variables represented in the graph (i.e., between the percentage of women in the parliament, and the birthrate), not one of the options presented reflects that concern.

2.1.3 Formal Inference (FoI)

In nearly every AP course included in the survey, students are expected to be able to read and critique reports on data. These may be study results, experimental or otherwise, published in a research report or in the media. Some coursework has students conducting and reporting on their own studies, and students may be expected to provide feedback to their peers or reflect on their own projects after they’ve finished. A very important part of interpreting statistical results is determining significance (a.k.a. conducting inference). Students need to be critical of the use of the term “significant” in articles and reports, because it could be used to refer to statistical significance, practical significance, or the common (non-technical) definition of significance. Statements of statistical significance usually accompany the results of formal statistical tests (e.g. t-tests, confidence intervals); practical significance is often quantified by an effect size (Agresti & Finlay, 2009). This is especially important in the natural and social sciences. Recent work on science standards have recognized this explicitly. One of the NGSS goals is that “By grade 12, students should be able to analyze data systematically, either to look for salient patterns or to test whether data are consistent with an initial hypothesis” (NGSS Lead States, 2013).

The logic of statistical significance testing is a particularly difficult concept for students to understand, as it requires reasoning about conditional statements, their understanding of which is often flawed. Typically, the reasoning of statistical significance tests starts with the assumption that some effect of interest is not present in the population (the condition). Evidence for the effect (or, evidence against the absence of an effect) is established by a low probability of obtaining something at least as extreme as the observed result under the assumption that there is no effect. This type of reasoning using a conditional statement
in which negation of the conclusion leads to negation of the condition is called *modus tollens*, which has been found to be a difficult type of reasoning for people to master (Evans, Newstead, & Byrne, 1993). The logic of inference from confidence intervals is similar. Even though the logic/reasoning is difficult, tests of significance are widely used, and first-year college students are expected to understand the results of statistical tests that they perform or that are reported by others.

While the interpretation of statistical significance is generally uncontroversial in the statistics literature, guidelines for defining, using, and interpreting practical significance in decision making processes is less well agreed upon (Kelley & Preacher, 2011). It is widely recommended by a number of research organizations that effect sizes be reported alongside determinations of statistical significance (e.g.s. American Psychological Association, 2010; Task Force on Reporting of Research Methods in AERA Publications, 2006; National Center for Education Statistics, 2002) and we take the stance that university coursework should be held to the same standard. An understanding of practical significance is important for reporting and critically reading statistical results, and the higher levels of our proposed construct include the formalization of this understanding in terms of an effect size. The definition for effect size given by Kelley and Preacher (2011): “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest” (p. 140). This definition is suitable for the inference construct as it describes development through the thinking process of considering the context of the statistical investigation (i.e. the question being addressed) in the choice of effect size that is reported. The highest level of the construct is the integration of practical and statistical significance, where a critical thinker balances the information provided by each and integrates both sources of information about the research question at hand, especially when they seemingly provide conflicting information (e.g. if a result is statistically, but not practically, significant or vice versa). It is not expected that an incoming university freshman student is familiar with a large collection of effect sizes but rather to understand the logic behind effect size and, as Kelley and Preacher (2011) describe, be able to choose “relevant effect sizes” with subjective input from field experts. The CR4CR Assessment assesses whether and how students are thinking about these choices, and in doing so emphasizes the importance of training statistical thinkers who do not make ad hoc choices about these things and who think critically about the choices that are made by others.

The FoI construct map is shown in Figure 2.5 and sample responses to some of the items shown in the MPG item (see Figure 2.6) are given in Table 2.2; it’s structure differentiates the progression of student development of claims concerning significance related to practical significance (effect size) and statistical significance (frequentist statistical tests and confidence intervals). The lower levels in both branches are identical with the split occurring after the level FoI1. It is important to note, then, that after the split, it is not necessarily expected that students progress on the two branches at equal rates. At this point, there is no claim being made about the ordering of the levels across the two branches. However, the structure does assume that the integration of the two branches constitutes the highest level. The focus is not on the procedures of conducting statistical tests or calculating effect sizes but, rather, the use and interpretations of the results that these procedures produce. This construct instead maps development of informed decision making using the results of
Formal Inference

(4) Integrated
Interprets statistically significant results in terms of an effect size. This is the basic task in understanding the limitations of statistical significance.

(3S) Informed
Interprets inference statistics correctly and can explain the underlying logic of the process. Attention paid to the appropriateness of the procedure.

(3P) Informed
Determines "large" & "small" effect sizes in the context of the problem and can explain the underlying logic of an effect size.

(2S) Algorithmic
Calculates and reports inference statistics. Statistical tests may be inappropriately applied. Interpretations of inference statistics may be incorrect.

(2P) Algorithmic
Calculates and reports effect sizes. Interpretations of effect sizes may be incorrect.

(1) Naive
(A) Determines significance on the basis of personal ideas about magnitude. (B) Over-influenced by the details of the current data collection. "Significance" is ignored. May interpret at the level of the experimental unit (not at the population level).

(0) No Inference
No conclusion is made.

Figure 2.5: The Formal Inference (FoI) construct map.
Table 2.2: Formal Inference sample performances.

<table>
<thead>
<tr>
<th>Level</th>
<th>Performances</th>
<th>FoI-S (Statistical Significance)</th>
<th>FoI-P (Practical Significance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Integrated</td>
<td>“Assuming we find a p-value less than 0.05 which is likely, the difference is statistically significant. However, I’d want to find out how this difference actually effects drivers in terms of the amount of money they spend at the gas pump. If they don’t feel the difference, even if it’s statistically significant, it doesn’t mean much.”</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Informed</td>
<td>“Probably since the confidence intervals don’t overlap. However, a confidence interval for the difference in means would be more helpful in this context (or a t-test for a difference in means).”</td>
<td>“There is a 2.6 MPG difference in the means. The variability around the mean in all of the samples is a lot smaller than that. The effect size will probably end up being large.”</td>
</tr>
<tr>
<td>2</td>
<td>Algorithmic</td>
<td>“The confidence intervals overlap from 2012 to 2013 and then again from 2013 to 2014.”</td>
<td>“I need more information to compute a Cohen’s $d$ to determine practical significance.”</td>
</tr>
<tr>
<td>1</td>
<td>Naive (B)</td>
<td>“There is only about a 2.6 MPG difference in the means. This isn’t very significant.”</td>
<td>“Yes. MPGs are higher in 2013 and even higher in 2014.”</td>
</tr>
<tr>
<td>1</td>
<td>Naive (A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>No inference</td>
<td>“I don’t know.”</td>
<td></td>
</tr>
</tbody>
</table>
inference procedures. Informally, it asks “To what extent is a conclusion (or claim) informed by the statistical procedure?”

FoI0, no inference, characterizes a response that does not contain a claim about inference. Level FoI1 contains two unordered sublevels. Though FoI1A is shown below FoI1B, we are not claiming that it is empirically lower than FoI1B. Responses at the FoI1A level where an inference is made do not consider the idea of significance - the result of the data analysis is directly used to make the claim. For example, if the treatment group in a randomized drug trial has a lower mean blood pressure than the control group, this difference is reported and interpreted as such. The localized result is all that is reported, and it may be reported at the population level.

At FoI2B, respondents rely on their own personal intuitions about the magnitude of the results to make a determination of significance. For instance, an FoI2B response may be along the lines of “One inch isn’t that big, so the difference is not significant.” This response fails to contextualize their conclusion about how important one inch is, relative to the phenomenon under study. For example, one inch of growth in the height of a toddler is relatively more meaningful than a one-inch difference in the distance of a daily commute to work; hence, context matters for practical significance. Respondents who use the term significance colloquially (and don’t differentiate between statistical and practical significance) would be placed here. Gross misinterpretations of inference procedures would also be placed here, such as a respondent determining practical significance based upon the width of a provided confidence interval. In that case, there is little to no understanding of what a confidence interval means and arguably no differentiation of practical and statistical significance, yet the respondent is showing some reasoning skill (albeit incorrect) in coming to a conclusion.

At this point, the construct branches into the statistical (S) and practical strands. The levels within each branch are discussed here together, as they parallel each other in explanation. At the algorithmic levels (FoI2S and FoI2P), a respondent can execute the algorithm to produce an inference statistic. \( P \)-values or standardized effect sizes are reported, but there may be subtle misinterpretations of these values such as interpreting a \( p \)-value of 0.02 as “The probability of my data is 2%” or a 95% confidence interval as an “interval that has a 95% probability of containing the true value.” It is important to note, again, that respondents are not expected to be at these levels on the S and P branches simultaneously. They may very well know the algorithm to produce a 90% confidence interval (placing them at FoI2S), but not know a thing about Cohen’s \( d \) (not quite reaching FoI2P).

At FoI2S and FoI2P, procedures may be inappropriately applied and inference may be drawn about incorrect populations (overgeneralization, usually). However, at the informed levels (FoI3P and FoI3S), attention is paid to the procedure because this is when respondents can explicate the logic behind statistical tests. In order to do so in context, the data generating model of the null hypothesis is understood and thus the correct procedure to match the data structure and research question must be chosen. Finally, the S and P branches reconnect at FoI4, where respondents integrate statistical and practical significance. Here, the limitations of \( p \)-values are recognized in the context of effect sizes. Both are reported (with correct interpretations) and they are coordinated. Situations where the statistical and
practical significance results conflict are especially important to reveal the level of critical thinking at FoI4. The task of recognizing the limitations of statistical significance in terms of an effect size is the highest level of critical thinking.

Another important element of this work that is outside the scope of this paper, but important to mention and clarify, is that of determining at what level in each construct constitutes college-ready. For now, the constructs are defined as fully as possible for what they are, and we do not claim that mastery of the top level of each construct is what defines college-ready. In all likelihood, college-ready is at the high-mid levels of these constructs. For example, The FoI construct map is intended to describe the full developmental progression through the two branches of formal inference. It may not be necessary for students to reach the top-most level of this construct to be prepared for first-year coursework. Correct interpretation of statistical tests, confidence intervals, and effect sizes may be sufficient for the critical evaluation of reports on data analysis that is expected in first-year coursework. Deeper investigative research into the content of coursework is needed to determine which level of mastery indicates that a student is college-ready in terms of statistical thinking. Analysis of textbook content and problems as well as interviews with professors who teach introductory coursework will provide insight. Further, different types of post-secondary institutions may have different expectations of first-year students. The college-ready boundary may vary depending on institution, subject matter, or degree program.

2.2 Assessment tasks

Both constructed response (CR) and selected response (SR) items were developed for the CR4CR Assessment as two main goals of the assessment design are to have both rich feedback and efficient scoring. Though SR items are often considered to be poor indicators of higher-level thinking skills, offering information only on rote memorizations, algorithms, “test-taking skills,” and other low-level knowledge and skills, the CR4CR Assessment SR items are written to give information at both the low and high ends of constructs. This is done by writing options at multiple levels of sophistication so that some choices are more sophisticated than others, but the “most correct” option is not readily apparent to respondents at lower levels. These items are then scored polytomously as opposed to the traditional dichotomous scoring (correct or incorrect) of multiple choice items. Thus, none of the options are “distractors,” instead each option is an “attractor” aligned to a particular level of the construct. This is quite a challenge for item developers (see the third paper of Arneson, 2019). Qualitative data gathered in cognitive interviews (a.k.a. “think alouds”) and pilot tests response data informed the options of the SR items so that they run the spectrum of the target construct, allowing for students to demonstrate high-level thinking.

The CR4CR Assessment is an item cluster-based assessment. Because the constructs are elements of critical thinking, it is necessary for respondents to meaningfully engage with complex contexts such as a potentially lengthy description of a statistical study or an excerpt from a newspaper article. Because these contexts are cognitively demanding and time consuming, multiple items are situated within them. These groups of items that share
Figure 2.6 provides a sample FoI-focused item cluster from CR4CR Assessment, the *Miles per Gallon* (MPG) item cluster. The test was delivered on a computer, using the BEAR Assessment System Software (BASS; Torres Irribarra, Frueand, Fischer, & Wilson, 2015). Figure 2.6 shows the rendering of the item cluster as it would be on the screen of a student taking the test. In this cluster, items 1-3 are aligned to FoI and item 4 is aligned to LDC. The majority of the item clusters that were developed for CR4CR Assessment are multidimensional in this way, containing unidimensional items scored on different constructs. In particular for the FoI construct, as respondents may be at different levels of sophistication in thinking about statistical versus practical significance, items explicitly targeted to each side of FoI are included in the same cluster context. Further, item clusters especially give respondents at the *integrated* (FoI4) level an opportunity to demonstrate this as they have to consider both statistical and practical significance separately, and then integrate their considerations. These steps would be difficult to measure in a single written item response (including selected response items), unless the item asked for an essay length response, which was not feasible for the CR4CR Assessment.

### 2.3 Empirical analysis

#### 2.3.1 Methods

As discussed in the introduction, the CR4CR framework is intentionally multidimensional and we acknowledge that our constructs are interrelated. To reflect this, we used a multidimensional Rasch-family model to analyze the scored response data produced in a field test of the CR4CR Assessment. Further, since the assessment is item-cluster based, we used the Rasch Testlet Model (RTM; Wang & Wilson, 2005) because we expect local item dependence (LID) among the items within each item cluster. Because our focus for this analysis is on the assessment performance as a whole and as we do not yet have a theory about how the relationship of the items within clusters is structured, a testlet model is appropriate. A principled guide for choosing a model for an item-cluster based assessment is provided in the first paper of Arneson (2019).

The polytomous RTM is an extension of Masters’ partial credit model (PCM; Masters, 1982). The RTM includes a random effect for each item cluster. These random effects are fixed to be uncorrelated with all other random effects in the model, including the \( \theta \) dimensions of interest (one for each of our constructs, LDC, MRC, and FoI). The ConQuest software was used for model estimation (Wu, Adams, Wilson, & Haldane, 2007).

The formulation of the polytomous multidimensional RTM is given in Equation 2.1 below. In this equation, \( p_{nij} \) and \( p_{n(i−1)} \) are the probabilities of person \( n \) scoring \( j \) and \( j−1 \) (respectively) on item \( i \) belonging to cluster \( d \), \( \theta_n \) is a vector of person \( n \)'s “abilities” on each dimension, \( Q_i \) is the loading vector of the item on the dimensions, \( b_{ij} \) is the item step parameter for step \( j \) of item \( i \), and \( \gamma_{nd(i)} \) is the random effect for item cluster \( d \). \( \gamma_{nd(i)} \) is equal
An automotive industry lobbyist wishes to claim that cars have been getting more energy efficient over time, and thus the federal government should relax some of the environmental regulations they impose in the industry.

To back up this claim, they use Environmental Protection Agency (EPA) fuel economy data to compare the “city” miles per gallon (MPG) ratings of random samples of cars from 2012, 2013, and 2014. Using statistical software, they found the mean MPG along with 95% confidence intervals. These are reported in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Mean MPG</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>19.41</td>
<td>(18.01, 20.81)</td>
</tr>
<tr>
<td>2013</td>
<td>20.73</td>
<td>(19.40, 22.06)</td>
</tr>
<tr>
<td>2014</td>
<td>22.03</td>
<td>(20.84, 23.22)</td>
</tr>
</tbody>
</table>

[1] The lobbyist makes the claim "There was a statistically significant improvement in fuel efficiency from 2012 to 2013." Do you agree with this statement? Why or why not?

Write your answer here.

[2] The lobbyist also makes the claim "There was a statistically significant improvement in fuel efficiency from 2012 to 2014." Do you agree with this statement? Why or why not?

Write your answer here.

[3] Is the observed 2.62 MPG improvement from 2012 to 2014 in fuel economy large enough to be of importance to an average driver? Why or why not?

Write your answer here.

[4] Recall that the lobbyist wants to use this analysis as evidence for a claim that government regulations can be relaxed. Without regard to the analysis results, do you think that he chose the right data to analyze? Why or why not?

Write your answer here.

Figure 2.6: Sample CR4CR Assessment item cluster: Miles per Gallon (MPG).
to 0 if item $i$ does not belong to cluster $d$; $d(i)$ is used in the subscript to indicate this. Note that the CR4CR Assessment items were written to load on exactly one construct/dimension each, so the $Q_i$ vector will be $(100)$, $(010)$, or $(001)$, depending on which construct the item is aligned to\(^6\).

$$\log \left( \frac{p_{nij}}{p_{nij(j-1)}} \right) = Q_i \theta_n - b_{ij} + \gamma_{nd(i)} \quad (2.1)$$

The field test data collection for the CR4CR Assessment contained three parent forms\(^7\), each targeted to one of the constructs but containing items across all three constructs. For example, Form G contained mostly LDC items, but also had MRC and FoI items to facilitate linking and produce estimates on all three constructs for each respondent. Table 2.3 provides information on all items included in this analysis and how they were distributed to each form. An entry of “SR” in a form column indicates that the items was present on the form and it is a selected response item and an entry of “CR” indicates it is a constructed response item. Some items have “SR/CR” and this means that either the SR or the CR version of the item was delivered to a given respondent (but not both)\(^8\). Note that there were eight common items across all three forms (Schooling, Birthday, Swim Club 1, Swim Club 2, Correlation 1, Correlation 2, Vocabulary, and Boards). In addition, the Cats item cluster was on both Forms G and I and P-Value was on both Forms H and I. And the SR versions of the MPG and Chocolate clusters were included on Forms G and H and half of the Form Is (see footnote).

The form design has direct ramifications on the analysis results, particularly the reliability indices for each dimension, which are largely affected by the number of items each respondent takes on each dimension and the spread of the person location distribution. We recognize that this form design is a limitation of the multidimensional analysis of the CR4CR field test data. A simulation study was conducted to investigate the effect of our form design on the reliability. The methods and results of the simulation study are provided in Appendix B.

\(^6\)If we were estimating a three-dimensional model without testlet effects, we would say that the items exhibit between-item multidimensionality, as opposed to within-item multidimensionality. Because the testlet effects in the RTM are essentially more dimensions, we could say that there is within-item multidimensionality. The testlet effects, however, are not the dimensions of interest for this analysis and we do not interpret them as meaningful, stand-alone dimensions. Thus, we intentionally avoid attaching either of these labels to this context to reduce confusion.

\(^7\)Each parent form was split into two child forms. This was done as one of the research questions for the project was about the empirical behavior of item clusters that were adapted from CR items to SR items. The details of this form structure, the analysis performed to investigate that research question, and the results are provided in the third paper of Arneson (2019). For simplicity in this paper, we will refer to and describe the design of the parent forms only.

\(^8\)As noted earlier in a previous footnote, there were two child forms for each parent form. This is not a focus of this paper, but it is important in order to understand how forms were linked, especially as it comes to the MPG item cluster.
### Table 2.3: Form design by item for the CR4CR field test.

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Construct</th>
<th>Form G</th>
<th>Form H</th>
<th>Form I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paternity Leave 1</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paternity Leave 2</td>
<td>LDC</td>
<td>SR/CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paternity Leave 3</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cats 1</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td>SR</td>
</tr>
<tr>
<td>Cats 2</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td>SR</td>
</tr>
<tr>
<td>Cats 3</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td>SR</td>
</tr>
<tr>
<td>Twitter 1</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter 2</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter 3</td>
<td>LDC</td>
<td>SR</td>
<td></td>
<td>SR/CR</td>
</tr>
<tr>
<td>Air Quality</td>
<td>MRC</td>
<td>SR/CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Titanic</td>
<td>MRC</td>
<td>SR/CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seatbelts 1</td>
<td>MRC</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seatbelts 2</td>
<td>MRC</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seatbelts 3</td>
<td>MRC</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seatbelts 4</td>
<td>MRC</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seatbelts 5</td>
<td>MRC</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 2</td>
<td>FoI</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 3</td>
<td>FoI</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 4</td>
<td>FoI</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 5</td>
<td>FoI</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 6</td>
<td>LDC</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 1</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR/CR</td>
</tr>
<tr>
<td>MPG 2</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR/CR</td>
</tr>
<tr>
<td>MPG 3</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR/CR</td>
</tr>
<tr>
<td>MPG 4</td>
<td>LDC</td>
<td>SR</td>
<td>SR</td>
<td>SR/CR</td>
</tr>
<tr>
<td>Chocolate 3</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR/CR</td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR/CR</td>
</tr>
<tr>
<td>Schooling</td>
<td>LDC</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>Birthday</td>
<td>MRC</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>Swim Club 1</td>
<td>MRC</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>Swim Club 2</td>
<td>MRC</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>GPAs</td>
<td>MRC</td>
<td>SR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-Value</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td></td>
</tr>
<tr>
<td>Correlation 1</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>Correlation 2</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>Confidence Interval</td>
<td>FoI</td>
<td>SR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>FoI</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>Boards</td>
<td>LDC</td>
<td>SR</td>
<td>SR</td>
<td>SR</td>
</tr>
</tbody>
</table>
Table 2.4: Respondent demographics.

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Number</th>
<th>% Prob/Stat</th>
<th>% Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School</td>
<td>82</td>
<td>95%</td>
<td>43%</td>
</tr>
<tr>
<td>University</td>
<td>273</td>
<td>40%</td>
<td>64%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>360</td>
<td>52%</td>
<td>59%</td>
</tr>
</tbody>
</table>

2.3.2 Data

There are 360 total respondents in the data set from multiple testing sites and this includes both high school students (23%; from Kansas and Ohio) and first year college students (76%; from California and Ohio). Five students (1%) did not provide their education level. The number of respondents at each education level is provided in Table 2.4 along with distributions of other demographics. The third column shows the percentage in each group who reported taking a probability or statistics course (at any level) in high school. Note that all of the high school participants in Kansas were enrolled in AP Statistics, but the assessment was given at the start of the school year. So, it is likely that many had no formal introduction to statistical inference at the time of testing. However, it is important when validating the structure of a construct to ensure that data is collected across all levels of the construct, low and high. Though these students may not have had exposure to statistical reasoning in their coursework, they have most likely had exercises in data-based reasoning by high school. The CR4CR Assessment contains items rooted in both. In an effort to get the higher ends of the construct, we also recruited college students to take the assessment.

One-hundred twenty-five respondents took Form G, 92 respondents took Form H, and 143 respondents took form I. Within each of these forms, note that approximately half of the respondents took the child form 1 (G1, H1, or I1) and the others took child form 2 (G2, H2, or I2).

2.3.3 Results

All of the models were calibrated using ConQuest (Wu et al., 2007). Graphics and tables were produced using R (R Core Team, 2017); ConQuest estimation output was read into R using the WrightMap package (Torres Irribarra & Freund, 2014).

Variance parameter estimates and reliability indices for the three substantive dimensions in the model (LDC, MRC, and FoI) are provided in Table 2.5. As seen here, the reliabilities are quite low – well below the commonly used threshold for acceptability of approximately 0.8. As noted above, our form design was developed in the interest of comparing subsets of items, and did not prioritize individual measurement of persons. The simulation study briefly described in Appendix B shows that these low reliabilities are largely driven by the narrowness of the person distributions – i.e. their relatively low variances.

We believe that our sample is censored, likely at both the low and high ends. Our sample
Table 2.5: Substantive dimension results.

<table>
<thead>
<tr>
<th></th>
<th>LDC</th>
<th>MRC</th>
<th>FoI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated variance</td>
<td>0.292</td>
<td>0.522</td>
<td>0.087</td>
</tr>
<tr>
<td>EAP reliability</td>
<td>0.429</td>
<td>0.390</td>
<td>0.349</td>
</tr>
</tbody>
</table>

Note: The EAP/PV reliability is a measure of test reliability or person separation reliability. It is calculated by dividing the variance of the individual EAP ability estimates by the observed person variance (Adams, 2005), and is equivalent to traditional reliability measures such as Cronbach’s alpha. There are as many EAP/PV reliability indexes estimated as dimensions in the model. Only the reliability indexes for the dimensions representing the target construct are presented here. Note that each model also included random item cluster effects that have an associated index, but they are not reported here.

was a convenience sample, and in an attempt to collect data from students at all levels of our constructs, we recruited both high school students (who had only just started AP Statistics and hadn’t yet reached much of the advanced content) and university students. Thus, we likely did not include in our sample students at the lower levels of our construct maps as they were either high school students who were enrolled in college-level coursework during high school or first- and second-year college students. Also, our sample has not captured the higher levels of our construct maps either. Though we reasoned this would be the case by recruiting students from a selective public university, we recruited students from non-specialized math courses and applied science courses such as astronomy and introductory biology. Fewer than half of the college students in our sample reported having taken formal statistics coursework, and for those that did, we did not ask them how well they performed in that coursework nor how long ago it was. Had we recruited students enrolled in more advanced math or applied coursework, we speculate that our person distributions would have been wider. In our small simulation study (described in the Appendix B), we found that distributions with standard deviations of 1.5 logits produced acceptable reliability indexes for all three dimensions, whereas simulated response data with distributions as we found in our data (all with standard deviations lower than 1.0 logit) did not even when the length of the test was more than doubled. One of the highest priorities of continuing research is to obtain a more variable sample. As for this study and one concerned with empirical comparisons of CR and SR items (reported in the third paper of this dissertation) for which the test forms were designed, the low reliability indexes are not expected to affect the main conclusions.

The peaked nature of the person distributions is evident in Figures 2.7 through 2.9. These figures provide the Wright Maps for each dimension modeled. The left sides of each panel provide the distribution of person location estimates as a histogram, and the right side of the panels provide the item thresholds, the point at which a respondent at the same location has a 50% of scoring into that level or higher, organized by the associated construct level. These thresholds are nonlinear transformations of the step parameters estimated by statistical software ($b_{ij}$ in Equation 2.1). Consider the point in the lower left of the LDC Wright Map in Figure 2.7 labeled Drug Trial 6), the second from the bottom point at “Level
1” along the horizontal axis. It is located at approximately −3 on the vertical axis. This means that a student who has a location of −3 on the LDC continuum has a 50% probability of scoring into at least LDC1 on the item named Drug Trial 6. A student with a location greater than −3 has more than a 50% probability of scoring into at least LDC1, and a respondent below −3 would have less than a 50% probability of doing so.

Note, however, that the ConQuest software does not estimate standard errors for them. So, though interpretations may be more meaningful, we cannot perform formal statistical tests to determine observed differences are statistically significant. However, we can look for meaningful patterns and differences in the threshold estimates using the Wright Maps as exploratory graphical devices (Figures 2.7 through 2.9). Each will be discussed in turn in the following paragraphs.

The LDC Wright Map in Figure 2.7 shows a steadily increasing pattern of item thresholds through the LDC construct levels. As an internal structure validity concern, we should hope to see a Wright Map organized in this way by construct level to show approximate “banding” of levels – a generally increasing pattern of thresholds without much overlap between levels. We should hope to see little overlap – that, for example, LDC2 thresholds are located mostly lower than LDC3 thresholds. For the most part for our LDC items, this looks to be the case. LDC1 thresholds are below LDC2 thresholds which are below LDC3 thresholds. This does not necessarily mean there is no overlap among these levels, however. One limitation of these Wright Maps is that there is no depiction of uncertainty around the threshold estimates. There are no points plotted at LDC4 as no one in our sample scored into that level.

Figure 2.8 plots the item thresholds and person estimate distribution for the MRC dimension. Here, we do not see the clear banding that was seen for the LDC construct. Levels MRC1, MRC2, and MRC3 overlap quite a lot, while thresholds for MRC4 and MRC5 are well separated.

In Figure 2.9, the thresholds for FoI-S items are plotted in blue and those for FoI-P items are plotted in yellow. In many cases, as many yellow points are above the blue points within the same level, the FoI-P items seem to be more difficult than the FoI-S items. However, there are “easy” FoI-P thresholds within each of the levels FoI1 through FoI3. Again, as no one scored into Level FoI4, there are no thresholds plotted there. Though there looks to be a bit more overlap than we saw for LDC in Figure 2.7, the FoI thresholds do exhibit a steadily increasing pattern as the construct levels increase along the horizontal axis. Perfect separation is not common, and that we do not see entire levels contained within each other (as we did at the lower end of MRC in Figure 2.8), this banding result is fairly good.

Though it is important, especially from an item development point of view, to inspect estimated item and person locations for each of the substantive dimensions on their own, the multidimensional model also provides estimates of correlations between dimensions (see Table 2.6). We see that the correlations between pairs of constructs are all moderately strong, as expected, at approximately 0.7.
Table 2.6: Covariance and correlation estimates among substantive dimensions.

<table>
<thead>
<tr>
<th></th>
<th>LDC</th>
<th>MRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRC</td>
<td>0.742</td>
<td></td>
</tr>
<tr>
<td>FoI</td>
<td>0.684</td>
<td>0.718</td>
</tr>
</tbody>
</table>

Figure 2.7: LDC Wright Map.
Figure 2.8: MRC Wright Map.
Figure 2.9: FoI Wright Map.
2.4 Discussion

The multidimensional treatment of the CR4CR response data collected in the field testing of the assessment is not without its limitations, but has provided clear guidance for the next steps of the research into defining and assessing college ready data-based reasoning. Within the LDC and FoI constructs, we achieved desirable banding of item thresholds in that our construct levels are generally separated (allowing for some overlap at each end). In this way, we can construct a mapping of student scores on the CR4CR Assessment to levels of the construct to make student scores meaningful, especially for formative purposes, for teachers. The MRC construct, however, had a lot of overlap in the lower levels, so revisions to the items, scoring guides, or the construct itself may be necessary before we can do that for MRC.

One important next step will be to obtain a more variable sample – to ensure that we capture both the lower and upper ends of our constructs. We will do this by recruiting from more advanced mathematics and statistics courses at both the high school and college levels, in addition to the types of courses we recruited from for the field testing. With our current data set, we can’t make any conclusions about the highest level of the LDC construct nor the highest level of the MRC construct as no student in our field test sample scored into either of them. It will also be important in future data collections to optimize the form design for individual measurement. This will include better balancing of items across the dimensions across the forms.

As one of the priorities of the CR4CR Assessment project is to design an assessment that can be administered in one hour or less, we are concerned with strategies to make the test forms short and still obtain estimates for the target constructs. With the moderately strong correlations among these three dimensions, we can continue to design forms that may have fewer items on a dimension than might be preferred. We can still obtain estimates for constructs that have as few as three to four items by using the scores on items aligned to related constructs. Another strategy we have undertaken to make the CR4CR Assessments more efficient, as described in the third paper of this dissertation, is to design selected response items that behave in similar ways to constructed response items, which can take students considerably longer to construct responses for (and scores considerably longer to score!).
Chapter 3

Developing selected response items for a critical statistical thinking instrument

There is a tension in educational testing between (1) authentic assessment of higher-order thinking skills and (2) efficient and reliable feedback on student performance, especially in the case of formative assessment that is to be used by teachers to plan instruction that is responsive to student needs. Higher-order thinking skills (such as critical thinking, argumentation, reasoning, problem solving, and metacognition) are thought by some to be better measured by constructed response (CR) items, especially at their higher levels, than by selected response (SR, a.k.a. multiple choice, closed-ended) items (Lane & Stone, 2006). Regarding (2), teachers and other educators who wish to use assessments in a formative way need feedback on student performance (often in the form of test and/or item score reports) in a timely manner in order to make instructional decisions based on student performance. There has been a shift in the expectations regarding the use of assessment in education, with an increasing need to place students on a continuum of development instead of measuring knowledge of facts (Pellegrino, 2013). CR items are costly to score, in terms of time and money, and can often prevent this timeliness. Further, CR items are costly to develop in that they may undergo multiple development iterations to increase rater reliability, for either human or machine rating. For these efficiency concerns, formative test developers may want to use SR items. This paper outlines an approach to developing SR items for a complex statistical thinking construct in an attempt to alleviate the tension between the authenticity (face validity) of the item format and the efficiency of the score reporting.

One of the components of a validity argument for a test is that of response process validity (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014). This involves gathering of evidence to support that the cognitive processes used by respondents when completing an assessment are those that the assessment is said to measure. In this paper, we describe two critical statistical thinking constructs and an instrument developed to measure them. Response process validity evidence
would be that which supports that respondents are using the higher-order thinking skills described in the constructs when responding to items on our assessment. This type of validity evidence is usually collected using cognitive interviews called *think-aloud protocols* and is *qualitative* in nature. This paper outlines a complementary *quantitative* investigation into response processes in the context of measuring critical statistical thinking with selected response (a.k.a. multiple choice) items. As the focus of this paper is on the quantitative analysis of the resulting field test data, the think-aloud protocol results are not discussed in detail, though this is not to diminish their importance in the item design process and the validity argument for the instrument.

The first section summarizes some of the current literature on measuring higher-order thinking skills and the use of CR and SR items on tests. The next section covers (a) the research project that generated the data analyzed in this paper, providing the project’s motivation, (b) a high-level description of the construct and test development work, and (c) some of the literature on K-16 statistics education. The third and fourth sections outline the methods employed: the methods for designing the items for the test and the quantitative methods used to analyze the scored response data, respectively. The fifth section provides the results of the quantitative analysis, and the final section is a discussion of those results leading to a more general conclusion about the feasibility of this approach, as well as suggestions for further research.

### 3.1 Literature review

#### 3.1.1 Assessing higher-order thinking skills

There is a recent increased instructional focus on higher-order thinking skills brought about by their inclusion in, for example, the Common Core State Standards (CCSS) in Mathematics and Language Arts, and Next Generation Science Standards (NGSS), and the Organization for Economic Cooperation and Development’s (OECD) 21st Century Skills (Ercikan & Pellegrino, 2017). Higher-order thinking skills are those that enhance the construction of deeper, conceptually-driven understanding (Schraw & Robinson, 2011) and include reasoning, argumentation, problem solving, critical thinking, and metacognition skills (Schraw, McCrudden, Lehman, & Hoffman, 2011). Measurement of these skills is often domain-specific; for example, instruments to measure argumentation are situated in a content domain such as science or literature. The project described in this paper, development of the Critical Reasoning for College Readiness (CR4CR) Assessment focuses on critical *statistical* thinking and *data-based* reasoning. Performance on assessments of higher-order thinking skills depends not only on domain-general thinking skills, but also on content-specific knowledge and skills (Schraw & Robinson, 2011).

These higher-order thinking skills lead to much more complex construct definitions than other, traditionally tested skills such as procedural knowledge or recall of facts. Because these complex constructs do not yet enjoy a rich literature in the measurement field, there is a particular need for a principled assessment design and validation approach for measures of
them (Ercikan & Pellegrino, 2017). The principled approach adopted for the development of the CR4CR Assessment is the BEAR\textsuperscript{1} Assessment System (Wilson, 2005; Wilson & Sloane, 2000).

### 3.1.2 Constructed response (CR) items

Complex constructs like critical thinking involve integration and coordination of knowledge and skills and generally occur in instruction and assessment as an application to a particular context (Nichols & Huff, 2017). Critical thinking tasks invariably require some content or contextual knowledge, and thus it can be difficult to separate the critical thinking skill and the particular content knowledge. Because of this, it has been recommended that measures of complex thinking use primarily CR item types (Nichols, Ferrara, & Lai, 2015; Ercikan & Pellegrino, 2017).

CR items (and other open-ended item types such as portfolios, projects, observations) may be authentic in that they require the complexity of skills that parallel real-life problem solving (Lane & Stone, 2006) and this is one of the strongest proponent arguments for their use. They are also less restrictive than closed-ended items in that respondents can demonstrate skill levels from low to high.

Many current assessment frameworks for thinking skills employ the use of extended tasks in that multiple test items are contextualized using a common stimulus (Pellegrino et al., 2014; WestEd & Council of Chief State School Officers, 2015). Other researchers call for the use of complex problem situations that are dynamic, interactive, involve multiple interrelated variables, and may have multiple goals and paths to solutions (Herde, Wüstenberg, & Grieff, 2016). These recommendations do not come without a discussion of their ramifications. Nichols et al. (2015) recognize that creating a complex extended task, especially those that are technologically enhanced, increases the number of “design decisions” that test writers make. It is important, then, to consider which of these decisions are largely “window dressing” and those that are essential to the construct as defined. Building a flashy, interactive, technologically enhanced test item may look, on the surface, as innovative, but it is important to consider if and how the flashiness of it reflects the construct being measured. As with any CR item type that is human-scored, there will be issues of rater reliability. Machine scoring of CR items carry their own reliability and validity issues. Bejar (2017) points out that automated scoring can lead to sacrificing score meaning, especially when the scoring algorithm can be gamed.

Further, as mentioned in the introduction, CR items are more costly in terms of scoring time and resources. As tests are increasingly expected to be used formatively (Nichols et al., 2015), timely feedback on student skill levels is imperative, and CR items threaten this efficiency. Many automated scoring methods for short-answer and essay CR item types have been developed in the past decades and most are proprietary such as AutoScore from American Institutes for Research (AIR), Bookette from CTB McGraw-Hill, CRASE\textsuperscript{TM} from Pacific Metrics, e-rater\textsuperscript{©} from Educational Testing Service, Intelligent Essay Assessor (IEA)

\textsuperscript{1}Berkeley Evaluation and Assessment Research Center
from Pearson Knowledge Technologies, *IntelliMetric* from Vantage Learning, *Lexile*® Writing Analyzer from MetaMetrics, and Project Essay Grade (PEG) from Measurement, Inc. (Shermis, 2014). However, these bring limitations of their own. The scoring algorithm employed has direct effects on the validity of score meaning (Bejar, 2017) especially when that algorithm focuses on the existence of certain features in text instead of the strength or sophistication of how those features were used by a respondent. An automated scoring procedure must be grounded in “the science of student learning” in the same way that the item itself is, and this often involves designing or altering scoring engines on a by-item basis (Bejar, 2017). This approach may still not be cost-effective. Scores generated by human raters are also subject to rater effects, and steps should be taken to minimize these. Human-scored CR items should undergo rater reliability studies, which can also be costly, in order to reduce the score variance attributable to idiosyncrasies among raters.

Messick (1995), and later Lane (2011), outline two potential sources of threats to the validity of score interpretation, specifically when performance assessments are used to measure higher-order thinking skills: construct underrepresentation and construct irrelevant variance. CR items are less technically efficient to include on an assessment because they often take longer and receive fewer scores than would a collection of SR items. Performance task-based assessments have fewer items and can thus suffer from construct underrepresentation when students are not given enough or varied enough opportunities to demonstrate their ability. SR items can be structured to be faster for respondents to complete (which means more can be included and there are enough opportunities) and also can be targeted at specific levels of the intended construct(s) in order to ensure that those opportunities are varied enough so that higher ability respondents can demonstrate their high abilities and lower ability respondents can demonstrate their lower abilities. The second issue is that open-ended formats can increase construct irrelevant variance as respondents can bring in many different skills and ways of reasoning to solve a task (Leighton, 2011; Nichols et al., 2015).

### 3.1.3 Selected response items

SR items are often reserved for measuring content knowledge as they are viewed to give little information regarding higher-order thinking skills (Reich, 2015). A SR item restricts a student to demonstrating only up to the skill level that the item response choices permit.

Many of the common criticisms of SR item types such as that they are viewed as “disconnected and shallow” (Charap, 2015) seem to only consider dichotomously scored (correct/incorrect), single-select, traditional multiple-choice questions such as the *P-Value* item shown Figure 3.1. This item is intended to assess whether a student knows that the smaller the *p*-value, the stronger the evidence of a relationship. Note that a respondent, if guessing completely at random, has a 20% chance of guessing the correct answer.

Relatedly, it is important to note that SR items carry with them their own concerns about irrelevant variance. Multiple-choice item responses can be affected by a respondent’s test-wiseness, defined by (Smith, 1982) as a skill that is “logically independent of the trait being measured” (p. 11). Reich (2015) discusses this within a standardized test of his-
A researcher uses a chi-squared test to determine if there is a relationship between 2 categorical variables. Which of the following statements about $p$ values indicates the strongest evidence of such a relationship?

- $p$ value = 0.002
- $p$ value = 0.01
- $p$ value < 0.05
- $p$ value = 0.05
- $p$ value > 0.10

Figure 3.1: The $P$-Value item.

Figure 3.1: The $P$-Value item.

tery knowledge (not necessarily higher-order thinking), and found that both test-wise skills and content knowledge played a part in response processes for multiple choice questions. Multiple-choice test strategies include lower-order skills such as memorization or working backwards, though some may verge on higher-order skills such as probabilistic reasoning if a respondent successively eliminates choices (Leighton, 2011). No matter the sophistication of the multiple-choice strategy, however, it is likely not what the test intends to measure. In analyzing only response data, one cannot be sure which is employed by a respondent to arrive at the submitted response. This highlights the importance of response process validity evidence in designing and validating a measure that includes SR item types. Think-aloud protocols can give insight into what types of test-wise skills a respondent could employ on a given SR item; however these cognitive interviews cannot be held with all respondents.

Further, no information about what or how a student is thinking about the data displays is collected in their responses to items like the Strong $P$-Value item in Figure 3.1. As demonstrated in the following sections, computer delivery of test items opens allows for more flexible and adaptive response types to address some of these concerns in measuring higher-order thinking. Returning to Lane’s (2011) second caution concerning performance assessments regarding construct irrelevant variance, extending and modifying the traditional multiple-choice format may lessen the impact of construct irrelevant variance arising from test-wiseness. More varied and unfamiliar formats could mean that students are not pre-equipped with strategies for gaming them. Further, SR items can help to strengthen the match between item and intended construct by limiting the response options to options aligned to that construct.

Leighton (2011) recommends that both open- and closed-ended items be used to measure higher-order thinking. Specifically, for SR items, she advises that care is taken to develop distractors that are attractive to lower-level students so that the most correct answer is not obvious. This strategy of building “attractors” for different levels of the construct being measured also addresses the issue of test-wiseness. As SR items have been widely used to assess content knowledge and not thinking skills (Schraw et al., 2011), the recommendation to use both item types for thinking skills might be considered somewhat revolutionary. Though, elsewhere in the literature, item design approaches that extend the traditional single-select, dichotomously scored multiple choice items have been discussed and analyzed. The ordered
multiple choice (OMC) item type developed by Briggs, Alonzo, Schwab, and Wilson (2006) is a promising item format to explore for measuring higher-order thinking skills. These OMC items offer choices at different levels of sophistication (that align to different levels of the construct map) and are scored polytomously. There is not just one correct answer, but instead answers that are of different levels of correctness or sophistication.

### 3.2 Measuring college-ready statistical thinking

Both selected and constructed response items were developed for the CR4CR Assessment as their relative strengths and weaknesses complement each other. We also developed item clusters (a.k.a. bundles or testlets) with more cognitively demanding and authentic stimulus material, a recommendation made by Charap (2015) as a way to increase the construct validity of tests of higher-order thinking skills. We avoid making “single-cluster” tests, however, in which all items are associated with the same stimulus real-life context which does receive some attention in at least the historical thinking assessment literature (Seixas, Gibson, & Ercikan, 2015) as this is likely to create a test that is over-focused on just one context. Items that refer back to the same stimulus material are likely to exhibit a “testlet effect” due to a violation of the conditional independence assumption. When there are no independent items or other item clusters on a test, the testlet effect cannot be separated from the effect of the construct being measured.

There has been in recent years growing emphasis on preparing college-ready students in K-12 education. A very significant example is that the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) explicitly names college readiness as one of its goals. This is, in part a response to the fact that the United States has fallen in many global rankings of technological advancement, including The Global Innovation Index (Cornell University, INSEAD, & WIPO, 2016) and The Global Information Technology Report (World Economic Forum & INSEAD, 2015). Another indicator of this situation is that college campuses have seen increased enrollment in remedial coursework (X. Chen, 2016) although this could be due to the admission of students of increasing lower ability. These circumstances have motivated a call for increased rigor and explicit focus on college readiness during K-12 schooling. The ACT and SAT have long been considered indices (proxies) of college readiness, but definitions of college readiness extend beyond the content of these entrance exams and they cannot be considered complete measures of college readiness.

The BEAR Center and National Math and Science Initiative (NMSI) have partnered to develop valid, reliable, and psychometrically sound measures of discipline specific college readiness aligned with Common Core, AP Content, and NAGB/EPIC research. As statistical thinking is a skill that shows up across many first year applied\(^2\) college courses, one task has been to develop the CR4CR Assessment which is not focused on the mathematical mechanics of data analysis and statistics, but their applications in entry-level college coursework outside

\(^2\)By “applied,” we mean coursework outside mathematics and statistics departments. This may include courses in the natural sciences, health careers, social sciences, humanities, or arts.
of mathematics and statistics departments.

The *Standards for Success* research project (Conley, 2005), asked college professors, especially those of the natural and social sciences, to indicate the prerequisite knowledge of basic statistical concepts and techniques that play an important role in entry-level courses. Conley reports that students who fail initial coursework because they lack a prerequisite skill will avoid majors in that area of study, “closing off entire avenues of the curriculum and career pathways” (p. 114). It is interesting to note, however, that Conley observed that statistical skills are less important in entry-level coursework for mathematics majors while it is imperative for success in other majors in the natural sciences (e.g., biology, ecology, physics), social sciences (e.g., economics, psychology, journalism), professional degrees (e.g., nursing), and even in some parts of the humanities (e.g., history).

We surveyed Advanced Placement® (AP) coursework and found that moderate levels of statistical knowledge was a prerequisite for most of the courses offered. Thirty-three AP courses were surveyed and statistical skills were either referenced in the course objectives or included in sample items in all but five of the surveyed courses.4

Considerable advancements have been made in recent years to identify the skills and knowledge required for students to be considered college-ready. We reviewed some of the most notable collections of college readiness standards to inform our own framework for college readiness in statistical thinking.

- **Common Core State Standards Mathematics (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).** Conley et al. (2011) report that the CCSS-M standard “Evaluate reports on data” was rated as one of the most important standards in the statistics strand by instructors of first-year college courses. Lajoie and Romberg (1998) state that K-12 students should learn to both critique as well as produce reports of statistical results “as required” for their future roles as consumers and produces in society. This is all the more relevant for their specific roles as college/university students for those who choose that path.

- **Knowledge and Skills for University Success (KSUS; Conley, 2005).**

- **Standards for College Success Mathematics and Statistics, Adapted for Integrated Curricula (CBSCS; Mathematics and Statistics Standards Advisory Committee, 2007).** In addition to the mathematics and statistics standards already referenced, the College Board Standards for College Success™ for Science (Science Standards Advisory Committee, 2009), the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), and the KSUS Standards in all other disciplines (English, Natural Sciences, Social Sciences, Second Languages, and Arts) were also scrutinized for mention of statistical thinking. This was an especially important task as the main area of inquiry is

---

3AP courses were surveyed as they can be considered reflective of entry-level college coursework and provide an easily accessible source of widely-used documents.

4There are 38 AP courses. We did not survey the five fine arts courses (Art History, Music Theory, 2-D Design, 3-D Design, and Drawing) as they were not expected to incorporate statistical thinking skills in their course objectives nor the associated AP exam items.
in how statistics is applied in the non-pure math disciplines.

Along with these three documents, three other collections of K-12 academic standards for mathematics and statistics were surveyed.

- Mathematics Framework for the National Assessment of Educational Practices (NAEP; National Assessment of Educational Practices Governing Board, 2013). The NAEP framework is not intended to be a compendium of what should be taught, but rather an outline of what is tested. NAEP is administered to grades 4, 8, and 12 and isn’t meant to measure college readiness, but it is still an important list of topics and skills used to compare student achievement across U.S. states.

- Principles and Standards for School Mathematics (NCTM; National Council of Teachers of Mathematics, 2000). The NCTM Principles and Standards were designed to meet the need to understand mathematics in everyday life and in the workplace, and not explicitly to prepare students for college. As these standards do play a role in much curriculum development across the U.S., they indicate important considerations in creating a framework of statistical thinking.

- Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework (Franklin et al., 2007). The GAISE Report, endorsed by the American Statistical Association, lays out a loose framework for general statistical literacy with a large emphasis on variability. It uses three developmental levels to describe the maturation of statistical understanding in four “process components” (formulating questions, collecting data, analyzing data, and interpreting results) and in understanding variability. All of these elements are reflected in this framework, with greater detail, and must be read with an appreciation of the fact that progressing through the levels of development may occur at different rates for different skills and knowledge.

Our focus includes the societal demands for statistical literacy as the demands for college coursework include, but move beyond, those. Specifically, we considered the contexts of (1) reading and critiquing reports on data or statistical investigations, and (2) planning a statistical investigation, two common tasks for which we posit parallel thinking processes.

Combining the information garnered from the standards documents, the AP course survey, and Conley’s findings, we developed the Framework for College-Ready Statistical Thinking to guide the CR4CR Assessment design. This Framework consists of three high-level questions that a critical statistical thinker should ask and answer, regardless of the course content:

1. What decisions need to be (were) made based on the data? (Decision Making)
2. Where did the data come from? (Producing and Selecting Data)
3. What was done to the data? (Summarizing Data)
Figure 3.2 places each of those questions at the vertex of a triangle. Although it makes sense to see each vertex as being part of a sequence of steps in a statistical investigation - get the data, analyze it, make a decision based on it - the three parts need to also be considered simultaneously. For example, it makes no sense to think about the qualities of data collection without also considering the use(s) of that data for decision making. Statistical investigation starts with its motivation - What decisions need to be made based on data? This is shown in the bottom left of Figure 3.2. From there, it moves to the top left - the selection and/or production of data - What are the data and where did it come from? Once data has been obtained, it is presented and/or summarized in a way that informs the decision to be made – What was done to the data? Often, the investigation stops here. However, some may iterate through this critical questioning process by either refining steps of the investigation or repeating them. We conceptualize critical statistical thinking as asking oneself these essential questions and then basing a statistical investigation (or critique of one) on the answers. Note, however, that these are very broadly defined, and the critical thinker does not stop with these. A cascade of questions will arise based on these broad, starting questions. What the follow-up questions will be is largely determined by the material presented to the student and the domain in which the statistical investigation exists.

Underlying all statistical skills is basic probability theory, so that is represented at the center of the triangle. In our work so far, this area is not an explicit focus. Though probability theory underlies nearly all concepts in the statistics discipline, the focus is instead on how probability is applied when statistics are used in other (non-mathematical) fields. In short, it will be difficult to define the “prior knowledge” in probability without a solid understanding of the knowledge and skills that contribute to statistical thinking. The work developing the CR4CR Assessment seeks to build that understanding.

This categorization of statistical skills is not new. Scheaffer et al. (1998) organize the
statistical content they felt should be in the K-12 curriculum into these same categories with the addition of “number sense.” Their “Planning a Study and Producing Data” strand is analogous to *Producing and Selecting Data*, “Data Analysis” is analogous to *Summarizing Data*, and “Inferential Reasoning” is analogous to *Decision Making*. Though this may not be a completely novel structure to K-12 statistics education, this framework is focused on statistical thinking and reasoning, not the whole of the study of statistics. As such, Watson’s conception of critical questioning for statistical literacy (Watson & Callingham, 2003) underlies the selection and illustration of constructs included in this framework. Watson’s work is grounded in a three-tier framework for statistical literacy (Watson, 2006, Chapter 1) in which each tier describes increased sophistication in thinking about a statistical problem. This complements the construct modeling approach of BAS which also defines hierarchies of development in critical statistical literacy, though we aim to flesh out more specific elements of statistical literacy and with a refined focus on college-readiness, not general citizen statistical literacy.

The goals for the CR4CR Assessment include both breadth and depth. Regarding breadth, it should eventually include items for all the potential constructs, of which there are currently fifteen. For depth, the focus of this paper is on two specific constructs. In narrowing the focus to specific constructs, it is intended that the assessment results provide informative diagnostic information about student learning, insights into the structure of the identified construct, and lay the groundwork for the construction of a larger statistical thinking learning progression that also includes the relationships among the constructs involved. For this paper, we analyze items aligned to the following two constructs:

1. **Linking Data to a Claim** (LDC) from *Producing Data*;
2. **Formal Inference** (FoI) from *Decision Making*.

Two statistical thinking construct maps will be discussed in the following subsections. Each of these constructs has been defined independently as an entity that can stand on its own and be measured on its own. It is clear, however, that the CR4CR constructs may have interrelationships and connections to each other. One of the next steps of this research (not tackled in this paper) will be to define a learning progression that incorporates all of the constructs. This would entail placing structure among the constructs by coming up with a theory about how the levels on different constructs are related to each other. This theory can then be empirically tested. Because the CR4CR Assessment framework and its constructs are still subject to final confirmation, the focus is instead on defining each construct in its own right and testing the relationships among the levels within each single construct. Each construct should be validated on its own before investigating how its related to the other constructs. We make no claims about how levels of one construct are related to levels of another, and the items analysis described in later sections treats each construct on its own.
3.2.1 Linking Data to a Claim (LDC)

In this attempt to fully define college-ready statistical thinking, it is important to note that critical statistical thinking is not exclusively “quantitative.” Yes, quantitative statistical procedures are conducted to test hypotheses, but they are abstractions of real-world phenomena. An important task in either planning a statistical study or critically evaluating one is to justify the link between the real-world object under study and the variables or measurements that are purported to reflect what we want to know about that object. These measured variables make up the data and decisions about those data need to be justified relating to the question under investigation and the claim resulting from statistical analysis. Explicit consideration of this connection may not be a concern in traditional statistics courses when students learn statistical procedures. In these settings, it can usually be taken for granted that the data collected are the right data to test the hypothesis. In contrast, in applied settings, this concern that the selected data truly reflect the phenomena about which a claim is made is of central importance!

The AP course survey showed that many of first-year courses expect students to design plans for and then conduct data collection. AP Biology, Chemistry, and Physics each list “justifying the selection of data” in course objectives. Students in those courses will be expected to choose data, plan collection, and execute that plan (College Board, 2015a, 2014a, 2014b). AP Environmental Science and Human Geography also stipulate that students should be given the opportunity to collect data (College Board, 2013, 2015b). Inherent in putting together these collection plans are the levels of thought embodied in LDC. These thought processes are similar whether a student is planning the data collection for a study or reading the works of others the content of both the question and the claim must be reflected in the observations selected for the purposes of the study. This may be more straightforward for tasks that investigate physical phenomena than for tasks such as the example above that investigate latent phenomena like opinions.

The LDC construct in Figure 3.3 describes development in statistical thinking about this link between the data and the content of the research question (and thus, the content of the claim that answers that question) of a statistical investigation. Other popular frameworks for statistical investigation and statistical learning include a step for formulating a research question (e.g. Scheaffer et al., 1998), so it is important to note that the levels of this construct are conditioned on that question already having been formulated. In fact, we focus the discussion and definitions of the construct levels on the claim at the end of a statistical investigation. However, linking the data to either the question or the claim is the same critical thinking process. Note that the word “claim” could be replaced with “question” for any of the descriptors of LDC levels.

The LDC construct levels are presented in Figure 3.3; the qualitative descriptions of each level are provided on the left side with sample responses in the right-most column. The sample responses are for the Miles per Gallon (MPG) item 4 provided in Appendix A. The construct is read from the bottom up, as that is the direction of increasing sophistication within the construct. Note that each level represents a qualitatively different type of student response to an item, indicating a different level of sophistication in thinking. Students
Figure 3.3: The Linking Data to a Claim (LDC) construct map.
tend to develop gradually from lower levels to higher levels, and these “ideal-points” are milestones along a continuum of development: not all students will fit distinctly onto one of these points, the majority will be somewhere in-between them. All construct maps in this document follow this convention.

The LDC construct is an integration of the argumentation construct map developed by Henderson et al. (2014) and the SOLO taxonomy (Biggs & Collis, 1982). The SOLO taxonomy outlines understanding of some concept with increasing levels of coordination, from the ability to simply identify elements to the ability to integrate those elements. For LDC, the levels of SOLO were further disaggregated and the (Henderson et al., 2014) map provided guidance for structuring that disaggregation. Their model for argumentation is structured around the three elements of an argument (claim, evidence, and warrant), the coordination of these three elements, and then of evaluating competing arguments. In the LDC construct, the choice of data is the evidence and the link between the data and the statistical claim is the warrant.

The bottom of the construct, LDC0, characterizes a respondent that does not identify a statistical claim. Moving up to the external level (LDC1), there are two unordered sublevels that may characterize students at this level. At LDC1A, the respondent identifies or provides a claim, but without accompanying evidence; students fail to recognize there is a question about the choice of data at all. At LDC1B, respondents may judge the appropriateness of the data collected by relying only on an external (contextually irrelevant) authority, their own beliefs about the subject at hand, or they may hold an “always or never” attitude that the inclusion of a number, any number, in a claim is evidence enough for that claim, or those who trust no number. The sublevels have some qualitative distinctions, but they were found to span the same range of empirical difficulty and thus responses at either LDC1A or LDC1B both receive a numerical score of 1 when preparing the data for model estimation. In essence, it was no more difficult for a respondent to provide a weak link than to not provide a link at all in the LDC items developed for the CR4CR Assessment.

LDC2 linked is the first in which a valid and appropriate link between the data and the claim is explicitly drawn by the respondent (invalid or inappropriate links are considered “weak” as in LDC1B). However, this link is merely identified and there is no justification or statement for it as a choice. Respondents may not recognize that there is a choice of observations to represent some phenomena of interest. This recognition may be taken for granted, especially for tasks that involve physical/manifest phenomena that are easily measured. Consider question 4 of the MPG task in Appendix A. It may be taken for granted that average miles per gallon represents a car’s impact on the environment, but we expect higher level students to think about the bigger picture and recognize that other characteristics can affect a car’s environmental impact.

Respondents at LDC3 begin to realize that sometimes a link is justified and sometimes it is not. This is where the credibility of the data becomes a prime issue—the warrant is not only predicated on the existence of a link, but also that the data themselves are credible in this context. Of course, some data, such as randomly generated numbers, may be credible in only unusual cases, so that they may be unsuitable for a whole range of claims. However, students at LDC3 provide single-sided justifications only, focusing only on the choice that was made.
Responses at the highest level, LDC4, provide more complex and balanced justifications of the choice of data to represent the phenomena of interest. They compare choices and provide reasoning in favor of the choice and against other choices. The distinction between LDC4 and LDC5 is the inclusion of a critique. LDC3 is simply an argument for (or against) a choice of data. LDC4 furthers the argument by critiquing alternate choices by contrasting them with the respondent’s choice of data. An LDC3 response is an argument that may be of the structure “It is good because . . .”. Whereas, at LDC4, the argument takes a comparative form: “It is better than the alternatives because . . .”.

3.2.2 Formal Inference (FoI)

In nearly every AP course included in the survey, students are expected to be able to read and critique reports on data. These may be study results, experimental or otherwise, published in a research report or in the media. Some coursework has students conducting and reporting on their own studies, and students may be expected to provide feedback to their peers or reflect on their own projects after they’ve finished. A very important part of interpreting statistical results is determining significance (a.k.a. conducting inference). Students need to be critical of the use of the term “significant” in articles and reports, because it could be used to refer to statistical significance, practical significance, or the common (non-technical) definition of significance. Statements of statistical significance usually accompany the results of formal statistical tests (e.g. t-tests, confidence intervals); practical significance is often quantified by an effect size (Agresti & Finlay, 2009). This is especially important in the natural and social sciences. Recent work on science standards have recognized this explicitly. One of the NGSS goals is that “By grade 12, students should be able to analyze data systematically, either to look for salient patterns or to test whether data are consistent with an initial hypothesis” (NGSS Lead States, 2013).

The logic of statistical significance testing is a particularly difficult concept for students to understand, as it requires reasoning about conditional statements, their understanding of which is often flawed. Typically, the reasoning of statistical significance tests starts with the assumption that some effect of interest is not present in the population (the condition). Evidence for the effect (or, evidence against the absence of an effect) is established by a low probability of obtaining the observed result under the assumption that there is no effect. This type of reasoning using a conditional statement in which negation of the conclusion leads to negation of the condition is called modus tollens, which has been found to be a difficult type of reasoning for people to master (Evans et al., 1993). The logic of inference from confidence intervals is similar. Even though the logic/reasoning is difficult, tests of significance are widely used, and first-year college students are expected to understand the results of statistical tests that they perform or that are reported by others.

While the interpretation of statistical significance is generally uncontroversial in the statistics literature, guidelines for defining, using, and interpreting practical significance in decision making processes is less well agreed upon (Kelley & Preacher, 2011). It is widely recommended by a number of research organizations that effect sizes be reported alongside determinations of statistical significance (e.g. American Psychological Association, 2010;
Task Force on Reporting of Research Methods in AERA Publications, 2006; National Center for Education Statistics, 2002) and we take the stance that university coursework should be held to the same standard. An understanding of practical significance is important for reporting and critically reading statistical results, and the higher levels of our proposed construct include the formalization of this understanding in terms of an effect size. The definition for effect size given by Kelley and Preacher (2011): “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest” (p. 140). This definition is suitable for the inference construct as it describes development through the thinking process of considering the context of the statistical investigation (i.e. the question being addressed) in the choice of effect size that is reported. The highest level of the construct is the integration of practical and statistical significance, where a critical thinker balances the information provided by each and integrates both sources of information about the research question at hand, especially when they seemingly provide conflicting information (e.g. if a result is statistically, but not practically significant or vice versa). It is not expected that an incoming university freshman student to be familiar with a large collection of effect sizes but to understand the logic behind effect size and, as Kelley and Preacher (2011) describe, be able to choose “relevant effect sizes” with subjective input from field experts. The CR4CR Assessment assesses whether and how students are thinking about these choices, and in doing so emphasizes the importance of training statistical thinkers who do not make ad hoc choices about these things and who think critically about the choices that are made by others.

The FoI construct map is shown in Figure 3.4 and sample responses to some of the items shown in the MPG items (see Appendix A) are given in Table 3.1; it’s structure differentiates the progression of student development of claims concerning significance related to practical significance (effect size) and statistical significance (frequentist statistical tests and confidence intervals). The lower levels in both branches are identical with the split occurring after the level FoI1. It is important to note, then, that after the split, it is not necessarily expected that students progress on the two branches at equal rates. At this point, there is no claim being made about the ordering of the levels across the two branches. However, the structure does assume that the integration of the two branches constitutes the highest level. The focus is not on the procedures of conducting statistical tests or calculating effect sizes but, rather, the use and interpretations of the results that these procedures produce. This construct instead maps development of informed decision making using the results of inference procedures. Informally, it asks “To what extent is a conclusion (or claim) informed by the statistical procedure?”

FoI0, no inference, characterizes a response that does not contain a claim about inference. Level FoI1 contains two unordered sublevels. Though FoI1A is shown below FoI1B, we are not claiming that it is empirically lower than FoI1B. Responses at the FoI1A level where an inference is made do not consider the idea of significance the result of the data analysis is directly used to make the claim. For example, if the treatment group in a randomized drug trial has a lower mean blood pressure than the control group, this difference is reported and interpreted as such. The localized result is all that is reported, and it may be reported at the population level.
Algorithmic
Calculates and reports inference statistics. Statistical tests may be inappropriately applied. Interpretations of inference statistics may be incorrect.

(1) Naive
Determines significance on the basis of personal ideas about magnitude. Over-influenced by the details of the current data collection. "Significance" is ignored. May interpret at the level of the experimental unit (not at the population level).

(2) Algorithmic
Calculates and reports inference statistics. Statistical tests may be inappropriately applied. Interpretations of inference statistics may be incorrect.

(3) Informed
Interprets inference statistics correctly and can explain the underlying logic of the process. Attention paid to the appropriateness of the procedure.

(3P) Informed
Determines "large" & "small" effect sizes in the context of the problem and can explain the underlying logic of an effect size.

(2P) Algorithmic
Calculates and reports effect sizes. Interpretations of effect sizes may be incorrect.

(4) Integrated
Interprets statistically significant results in terms of an effect size. This is the basic task in understanding the limitations of statistical significance.

(0) No Inference
No conclusion is made.

Statistical Significance (Confidence Intervals, P-values)

Practical Significance (Effect Sizes)

Figure 3.4: The Formal Inference (FoI) construct map.
Table 3.1: Formal Inference sample performances.

<table>
<thead>
<tr>
<th>Level</th>
<th>FoI-S (Statistical Significance)</th>
<th>FoI-P (Practical Significance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Integrated</td>
<td>“Assuming we find a ( p )-value less than 0.05 which is likely, the difference is statistically significant. However, I’d want to find out how this difference actually affects drivers in terms of the amount of money they spend at the gas pump. If they don’t feel the difference, even if it’s statistically significant, it doesn’t mean much.”</td>
</tr>
<tr>
<td>3</td>
<td>Informed</td>
<td>“Probably since the confidence intervals don’t overlap. However, a confidence interval for the difference in means would be more helpful in this context (or a t-test for a difference in means).” “There is a 2.6 MPG difference in the means. The variability around the mean in all of the samples is a lot smaller than that. The effect size will probably end up being large.”</td>
</tr>
<tr>
<td>2</td>
<td>Algorithmic</td>
<td>“The confidence intervals overlap from 2012 to 2013 and then again from 2013 to 2014.” “I need more information to compute a Cohen’s ( d ) to determine practical significance.”</td>
</tr>
<tr>
<td>1</td>
<td>Naive (B)</td>
<td>“There is only about a 2.6 MPG difference in the means. This isn’t very significant.” “Yes. MPGs are higher in 2013 and even higher in 2014.”</td>
</tr>
<tr>
<td>0</td>
<td>No inference</td>
<td>“I don’t know.”</td>
</tr>
</tbody>
</table>
At FoI2B, respondents rely on their own personal intuitions about the magnitude of the results to make a determination of significance. For instance, an FoI2B response may be along the lines of “One inch isn’t that big, so the difference is not significant.” This response fails to contextualize their conclusion about how significant one inch is, relative to the phenomenon under study. For example, one inch of growth in the height of a toddler is relatively more meaningful than a one-inch difference in the distance of a daily commute to work, hence, context matters for practical significance. Respondents who use the term significance colloquially (and don’t differentiate between statistical and practical significance) would be placed here. Gross misinterpretations of inference procedures would also be placed here, such as a respondent determining practical significance based upon the width of a provided confidence interval. In that case, there is little to no understanding of what a confidence interval means and arguably no differentiation of practical and statistical significance, yet the respondent is showing some reasoning (albeit incorrect) skill in coming to a conclusion.

At this point, the construct branches into the statistical (S) and practical (P) strands. The levels within each branch are discussed here together, as they parallel each other in explanation. At the algorithmic levels (FoI2S and FoI2P), a respondent can execute the algorithm to produce an inference statistic. P-values or standardized effect sizes are reported, but there may be subtle misinterpretations of these values such as interpreting a p-value of 0.02 as “The probability of my data is 2%” or a 95% confidence interval as an interval that has a 95% probability of containing the true value. It is important to note, again, that respondents are not expected to be at these levels on the S and P branches simultaneously. They may very well have the algorithm to produce a 90% confidence interval (placing them at FoI2S), but not know a thing about Cohen’s d (not quite reaching FoI2P).

At FoI2S and FoI2P, procedures may be inappropriately applied and inference may be drawn about incorrect populations (overgeneralization, usually). However, at the informed levels (FoI3P and FoI3S), attention is paid to the procedure because this is when respondents can explicate the logic behind statistical tests. In order to do so in context, the data generating model of the null hypothesis is understood and thus the correct procedure to match the data structure and research question must be chosen. Finally, the S and P branches reconnect at FoI4, where respondents integrate statistical and practical significance. Here, the limitations of p-values are recognized in the context of effect sizes. Both are reported (with correct interpretations) and they are coordinated. Situations where the statistical and practical significance results conflict are especially important to reveal the level of critical thinking at FoI4. The task of recognizing the limitations of statistical significance in terms of an effect size is the highest level of critical thinking.

Another important element of this work that is outside the scope of this paper, but important to mention and clarify, is that of determining at what level in each construct constitutes college-ready. For now, the constructs are defined as fully as possible for what they are, and we do not claim that mastery of the top level of each construct is what defines college-ready. In all likelihood, college-ready is at the high-mid levels of these constructs. For example, The FoI construct map is intended to describe the full developmental progression through the two branches of formal inference. It may not be necessary for students to reach
the top-most level of this construct to be prepared for first-year coursework. Correct interpretation of statistical tests, confidence intervals, and effect sizes may be sufficient for the critical evaluation of reports on data analysis that is expected in first-year coursework. More investigative research into the content of actual coursework is needed to determine which level of mastery indicates that a student is college-ready in terms of statistical thinking. Analysis of textbook content and problems as well as interviews with professors who teach introductory coursework will provide insight. Further, different types of post-secondary institutions may have different expectations of first-year students. The college-ready boundary may vary depending on institution, subject matter, or degree program.

3.3 Items design methods

As we were tasked to measure higher-order thinking skills in this project, we initially focused our efforts on developing CR items. As discussed previously, CR items can be costly and time-consuming to score. The CR4CR Assessment is being developed for formative purposes, meaning that teachers will need reports on their students’ performance in a timely manner in order to use the results to inform their instructional decisions. So, there was a practical need in this project to develop items that could be quickly scored (preferably machine scored) while still measuring the same constructs as the CR items. This section describes how the SR items were developed using pilot response data to meet this practical need.

The SR item development for the CR4CR Assessment follows the principle that the response options are all written to be “attractors” to a specific level of the targeted construct. This follows the recommendation from Leighton (2011) that distractors should be written so that they are not obviously wrong to students at lower abilities and that the correct answer should not be so obviously right. However, our conception of SR items is more innovative than the traditionally discussed “one correct answer” single-select multiple choice item. SR items on the CR4CR Assessment are not all scored as simply correct or incorrect; most are scored polytomously as the response options reflect a spectrum of sophistication in critical thinking. Further, we decided to use multiple-select multiple choice items in many cases; these are the items that include instructions (and allow a respondent) to “choose all that apply” as in questions 1, 2, and 3a of the Twitter item cluster shown in Figure 3.5. The CR4CR Assessment was delivered online using the BEAR Assessment System Software (BASS) and all figures show how they render in the software for respondents. Note for the particular item being discussed, here, it was possible for respondents to choose contradictory combinations of choices. For example, a respondent could choose “None of these” along with any of the other available choices. Clearly contradictory responses such as this were scored at the lowest level of the construct.

The multi-select item format reduces the probability of correctly guessing the most correct response combination at random. Whereas a respondent who randomly guesses for the P-Value item has a 1 in 5 (20%) chance of selecting the most correct answer from the five choices, a random guesser has a 1 in 31 (approximately 3%) chance of selecting the most correct response in Twitter 2. Each of the options in Twitter 2 was written to be aligned
In 2012 during the US presidential race between President Barack Obama and Governor Mitt Romney, USA Today offered the "Twitter Election Meter" on its website. The image above shows the meter for Election Day, November 7, 2012: Barack Obama had a reading of 85 and Mitt Romney had a reading of 57. The following description was offered along with the graph:

This meter tracks the Twitter Polio Index, a daily measure of sentiment expressed by Twitter users about President Obama and Mitt Romney. The index, similar to a favorability rating, is calculated on a scale of 0 to 100, with 100 being the most positive.

The USA TODAY Twitter Election Meter is a physical representation of the Twitter Polio Index, totally number produced by Twitter in partnership with Topsy Labs and polling firm The Mellman Group and North Star Opinion Research. The meter is not intended to replace traditional polling, but to provide a daily snapshot of the sentiment of political conversations on Twitter. Unlike traditional polling, in which participants are carefully selected, controlled and asked a series of questions, the Twitter Polio Index is based on analyzing the volunteered opinions of people logged on to Twitter. The result is analogous to describing the general tone of political discussion overheard in a coffee shop.

(1) Which of the following statements are supported by the Twitter Election Meter readings on Election Day 2012? Choose all that apply. You must choose at least one.

- Tweets about Obama are more favorable than tweets about Romney.
- Since virtually anyone can Tweet, the views of everyone are represented.
- More people are planning to vote for Barack Obama than Mitt Romney.
- U.S. voters feel more positively about Barack Obama than Mitt Romney.

(2) For which of the following uses would it be appropriate to use the Twitter Election Meter readings? Choose all that apply. You must choose at least one.

- To describe the tone of political discussion on Twitter on a given day.
- To decide which TV stations to buy campaign ads from.
- To describe how favorably U.S. voters feel about the candidates.
- To predict voter turnout on Election Day.
- None of these.

(3a) Do you think that the Twitter Election Meter is useful for predicting the results of an election?

- Yes
- No

(3b) Which of the following statements best support your answer to question (3a)? Choose all that apply. You must choose at least one.

- There are U.S. voters who do not tweet.
- There are Twitter users who are not U.S. voters.
- It predicted the winner of the 2012 election.
- Since virtually anyone can Tweet, the views of everyone are represented.
- Social media data can't be trusted.
- Using volunteered opinions.
<table>
<thead>
<tr>
<th>Level</th>
<th>Label</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Balanced</td>
<td>Choice a “To describe the tone of political discussion on Twitter on a given day.”</td>
</tr>
<tr>
<td>3</td>
<td>Unilateral</td>
<td>Choice b “To decide which TV stations to buy campaign ads from.”</td>
</tr>
<tr>
<td>2</td>
<td>Linked</td>
<td>Choice c “To describe how favorably U.S. voters feel about the candidates.”</td>
</tr>
<tr>
<td>1</td>
<td>External</td>
<td>Choice d “To predict voter turnout on Election Day.”</td>
</tr>
<tr>
<td>0</td>
<td>No claim</td>
<td>Choice e “None of the above.” and any other choice (Clear contradiction)</td>
</tr>
</tbody>
</table>

Combinations of responses are scored at the lowest level present (unless otherwise noted). For example, a response of Choices a, c, and d would be scored into Level 2.

Figure 3.6: Scoring guide for **Twitter** 2.

to a particular level of the LDC construct. Both single- and multi-select items were scored polytomously as OMC items (Briggs et al., 2006). The scoring guide for **Twitter** 2 is provided in Table 3.6. The “most correct” response would be to choose only the first choice (choice a, “To describe the tone of political discussion on Twitter on a given day.”) and is coded at LDC2. Other combinations of choices are scored at levels LDC0 or LDC1. This scoring guide is based on the assumption that the “lowest level present” is the final score. Other scoring guides for multi-select item formats may not be as simple as this one, as each response combination must be considered in turn for how it maps back to the construct map. For the purposes of **Twitter** 2, it was decided that this simple rule was appropriate.

The scoring guide for the **Paternity Leave** items are somewhat more complex. In this case, each combination of responses was considered and aligned to a level of the LDC construct. Figure 3.7 provides the scoring guide. Note that the responses for **Paternity Leave** 1 and **Paternity Leave** 2 were considered simultaneously to provide a score for **Paternity Leave** 2. Thus, two designed items correspond to a single empirical item. **Paternity Leave** 1 was a single-select SR item which **Paternity Leave** 2 was multi-select. The choices for each item are indexed a, b, c, etc.

Both single- and multi-select SR items were developed for the CR4CR Assessment. We feel that the multi-select type, however, addresses many of the criticisms of using single-select response types for measuring higher-order thinking. They are more authentic in that they allow respondents to express some of the characteristics of higher-order thinking skills.

---

5 The full **Paternity Leave** SR item cluster is provided in Appendix A.
such as recognition of more than one solution path and coordinating results or ideas for two possible solutions. Further, the multi-select response type preserves some of the benefits of the single-select in that it can be readily machine scored, allows for test equating and linking without the complication of raters, and can help to ensure content coverage (Charap, 2015). Automatic, machine scoring can occur in the same way as it does for single-select item types, except that for multi-select item types, there are more score combinations that must be considered. Multiple choice items are often used for test equating and linking as they are often more “well-behaved” than CR item types (Kolen & Brennan, 2014). And, because SR items can be more explicitly targeted to specific content or construct levels, they can be used strategically in the test design to ensure coverage across a test.

As computer delivery also offers flexibility and innovation in SR item types. One particular feature of computer delivery that was employed to create the selected response item types was that of skip logic. Item clusters employing skip logic are adaptive in that a student’s response to one question determines which subsequent questions they are asked. It was used to (1) make the wording of item prompts more simple and natural and (2) to follow the procedural aspect of critical thinking when appropriate. Alexander et al. (2011) discuss the idea that higher-order thinking has procedural aspects, but this is not to say it is formulaic or algorithmic. High-level critical thinking is not a series of steps, but rather a move from surface-level to deep-processing of the stimulus information. The use of skip logic was to mirror the theorized thinking happening in the item response process.

Now that the SR item formats have been discussed, we will outline how the options for the items were developed. Because authenticity was a goal in the items design, it was imperative to write options for the SR items that were similar to what a respondent would have come up with on their own. This involved three main steps: (1) piloting CR versions, (2) writing
SR versions based on the pilot response data, and (3) field testing complementary CR and SR versions of each adapted item cluster for comparison. These steps are described in detail below for the Twitter item cluster which targets LDC. Three other item clusters received similar treatment and are provided in Appendix A. They are as follows (with the target construct(s) and number of comparison items): Paternity Leave (LDC, 2 items), MPG (FoI, 3 items; LDC, 1 item), and Chocolate (FoI, 2 items).

**Design Step 1: Piloting CR versions.** Initial versions of the item clusters were developed with an open-ended format. These items were subjected to a panel of experts, our College-Readiness Advisory Panel, consisting of measurement specialists, mathematics content experts, and first-year college instructors in mathematics and applied disciplines. Figure 3.8 shows the version of the Twitter item cluster that was used to collect the initial pilot data. For computer delivery, the item cluster was shown on two subsequent screens, and once the respondent advanced to the second screen, they were not allowed to return to the first. Responses collected in the pilot data collection and reactions from respondents during think-alouds showed us that students were confused by this item. They did not feel equipped to finish the description of the Meter as they did not feel they had enough information to do so. They also felt that the first two items were repetitive. In scoring the pilot data, we found that responses to the third item did not provide any valuable information about the LDC construct.

**Design Step 2: Develop complementary versions.** Based on these insights gained from analysis of the pilot data, the Twitter item cluster was revised and two complementary versions were written: the version shown previously in Figure 3.5 containing exclusively SR items and the version shown in Figure 3.9 in which the third item is open-ended (the first two SR items are identical in both versions). These will be referred to as Twitter SR and Twitter CR, respectively. The other comparison item clusters included in this analysis will be referred to with the same convention (e.g. Paternity Leave SR and Paternity Leave CR). The choices for the SR items were adapted from the responses collected in the pilot data collection.

The adaptation of CR items to SR items again involved collaboration with our Advisory Panel of experts. Responses and scores were provided to the Advisory Panel. For a given item, the Panel discussed patterns that they recognized in the responses, and developed a list of tags to describe and code individual responses. Tags like misconception, generalizes to U.S. voters, and voluntary response were assigned to responses to Twitter items that had these characteristics. The tag list for each item was specific to the item and the item’s context. Then, the Panel looked at all responses with a certain tag and drafted one sentence that exemplified the collection of similar responses and determined which level of the construct it was aligned to. To keep the number of choices manageable, only the 4-6 most popular tags received this treatment. Revised item clusters (both the SR and CR versions) were then subjected to another round of think alouds before the field test administration in an effort to further verify that the response choices in the SR versions were appropriate.

**Design Step 3: Strategic test form design for field testing.** One of the goals of the analysis for the field test data for the CR4CR Assessment was to empirically compare the behavior of the SR items with their CR counterpart. As mentioned in the introduction,
Suppose you are on the Election Meter development team. Finish the paragraph (with a few sentences) so that a visitor to the webpage fully understands the strengths and limitations of the Twitter Election Meter.

The USA TODAY Twitter Election meter is a graphic representation of the Twitter Political Index, a daily number produced by Twitter in partnership with Topsy Labs and polling firms The Mellman Group and North Star Opinion Research. The index is not intended to replace traditional polling, but to provide a daily snapshot of the sentiment of political conversation on Twitter. Unlike traditional polling, in which participants are scientifically selected, contacted and asked a series of questions, the Twitter Political Index is based on gathering the volunteered opinions of people logging on to Twitter. The result is analogous to describing the general tone of political discussion overheard in a café on a given day.

Do you believe this explanation adequately conveys the strengths and limitations of the Twitter Election Meter? Explain.

Suppose the meter registered at 25 for Governor Romney one week before the 2012 election and 57 on election day. Write a few sentences to explain what this would mean.

Do you think that the Twitter Election Meter is useful for predicting the results of an election? Why or why not?

Figure 3.8: The original constructed response version of the Twitter item cluster.
In 2012 during the US presidential race between President Barack Obama and Governor Mitt Romney, USA Today offered the “Twitter Election Meter” on its website. The image above shows the meter for Election Day, November 7, 2012: Barack Obama had a reading of 85 and Mitt Romney had a reading of 57. The following description was offered along with the graphic.

This meter tracks the Twitter Political Index, a daily measure of sentiment expressed by Twitter users about President Obama and Mitt Romney. The index, similar to a favorability rating, is calculated on a scale of 0 to 100, with 100 being the most positive.

The USA TODAY/Twitter Election meter is a graphic representation of the Twitter Political Index, a daily number produced by Twitter in partnership with Topsy Labs and polling firms The Mellman Group and North Star Opinion Research. The index is not intended to replace traditional polling, but to provide a daily snapshot of the sentiment of political conversation on Twitter. Unlike traditional polling, in which participants are scientifically selected, contacted and asked a series of questions, the Twitter Political Index is based on gathering the volunteered opinions of people logging on to Twitter. The result is analogous to describing the general tone of political discussion overheard in a café on a given day.

Tweets about Obama are more favorable than tweets about Romney.
Since virtually anyone can Tweet, the views of everyone are represented.
More people are planning to vote for Obama than Romney.
U.S. voters feel more positively about Obama than Romney.

[2] For which of the following uses would it be appropriate to use the Twitter Election Meter readings? Choose all that apply. You must choose at least one.

- To describe the tone of political discussion on Twitter on a given day.
- To decide which TV stations to buy campaign ads from.
- To describe how favorably U.S. voters feel about the candidates.
- To predict voter turnout on Election Day.
- None of these.

[3] Do you think that the Twitter Election Meter is useful for predicting the results of an election? Why or why not?

Write your answer here.

Figure 3.9: The revised constructed response version of the Twitter item cluster.
Table 3.2: CR4CR Assessment field test test form design.

<table>
<thead>
<tr>
<th>Form</th>
<th>Target construct</th>
<th>CR comparison cluster</th>
<th>SR comparison cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>LDC</td>
<td>Twitter CR</td>
<td>Paternity Leave SR</td>
</tr>
<tr>
<td>G2</td>
<td>LDC</td>
<td>Paternity Leave CR</td>
<td>Twitter SR</td>
</tr>
<tr>
<td>I1</td>
<td>FoI</td>
<td>Chocolate CR</td>
<td>Miles per Gallon SR</td>
</tr>
<tr>
<td>I2</td>
<td>FoI</td>
<td>Miles per Gallon CR</td>
<td>Chocolate SR</td>
</tr>
</tbody>
</table>

This is an element of the validity argument for the CR4CR Assessment; we want to be able to provide evidence that our SR items measure our constructs in a similar way as the CR items. In order to do this, two “parent” forms were designed, each targeting one of the three focal constructs (LDC and FoI) and are named Parent Forms G and I respectively. Parent Form G, targeting LDC, contained the two LDC item clusters for comparison - Twitter and Paternity Leave - which gives rise to the two forms G1 and G2. Form G1 contained Twitter SR and Paternity Leave CR while Form G2 contained Twitter CR and Paternity Leave SR. These item clusters were presented back-to-back in the middle of the test. The item order on Forms G1 and G2 otherwise remained constant, so an identical set of items was given prior to the comparison clusters and an identical set after. Parent Forms I was broken up into two operational forms each in a similar way. These form design factors are summarized in Table 3.2.

3.4 Quantitative methods

Because the focus of this study was to investigate the empirical behavior of the CR and SR item types within each construct, two separate analyses were conducted, one for each of the two constructs. For each analysis, the polytomous version of the Rasch Testlet Model (RTM; Wang & Wilson, 2005) was used. The polytomous RTM is an extension of Masters’ partial credit model (PCM; Masters, 1982). This model was used instead of a straightforward application of the PCM because the CR4CR Assessment is item cluster-based assessment. When an assessment contains a collection of items situated under a common context, as seen in the Twitter item clusters in Figures 3.5 and 3.9, we anticipate the local independence assumption that underlies the PCM and other IRT models to be violated. We expect item scores within an item cluster to exhibit local dependence (or residual correlations) and here we account for this in the model. The RTM does this by including a random effect for each item cluster. These random effects are fixed to be uncorrelated with all other random effects in the model, including the $\theta$ dimension of interest. There are many options for modeling this local item dependence, and the RTM model was chosen in this instance as we should account for that dependence in the model estimation, but we are not looking to explain it.

---

6Form H targeted a construct called Meta-Representational Competence (MRC). The open-ended items showed very low reliability, and as such it is not appropriate to draw conclusions about the relationship between the CR and SR items. Items and scoring guides are currently undergoing revision for future reports and are not discussed here.
nor interpret the item cluster “dimensions” (Arneson, 2019). The ConQuest software was used for model estimation (Wu et al., 2007).

The formulation of the polytomous RTM is given in Equation 3.1 below. In this equation, $p_{nij}$ and $p_{n(i-1)}$ are the probabilities of person $n$ scoring $j$ and $j-1$ (respectively) on item $i$ belonging to cluster $d$, $\theta_n$ is person $n$’s “ability”, $b_{ij}$ is the item step parameter for step $j$ of item $i$, and $\gamma_{nd(i)}$ is the random effect for item cluster $d$. $\gamma_{nd(i)}$ is equal to 0 if item $i$ does not belong to cluster $d$; $d(i)$ is used in the subscript to indicate this.

\[
\log \left( \frac{p_{nij}}{p_{n(i-1)}} \right) = \theta_n - b_{ij} + \gamma_{nd(i)}
\]  \hspace{1cm} (3.1)

In the LDC analysis, only scores for LDC items from Form G were included in the analysis in order to avoid the potential complications of using data from multiple forms that involved different proportions of LDC (and other) items. The RTM is multidimensional, mathematically, as it contains a dimension for each item cluster in addition to the single dimension of interest, in this case LDC. The item cluster dimensions (the $\gamma$s) are not meant to be interpreted but included as a way to account for the local item dependence among items in a cluster. Form G is illustrated in the path diagram in Figure 3.10. Note that there is a single dimension for Paternity Leave and Twitter, which is loaded on by both the CR and SR versions of the comparison items, as opposed to including separate item cluster effects for the different versions. Path diagrams are included to communicate information about each of the analyses such as the number of items and the dimensional structure, but note they do not fully specify the models as some more formal path diagrams are meant to
do\textsuperscript{7}. The cells of the covariance matrix for the RTM are constrained to be zero except along the diagonal (the variances of each dimension). On Form G, there were three item clusters - Paternity Leave, Twitter, and Cats - so the covariance matrix is 4 × 4. It is provided in Figure 3.12.

The FoI analysis, on the other hand, was two-dimensional due to the parallel structure of this construct, as shown in Figure 3.13. FoI-S and FoI-P were considered distinct (but correlated) dimensions, and each item was classified as yielding either an FoI-S or and FoI-P score (but none provided both). For the FoI analysis, Equation 3.1 is altered to include a vector of $\theta$ dimensions and the covariance between the two substantive dimensions (FoI-S and FoI-P) was estimated resulting in the covariance matrix in Figure 3.11. The FoI form (Form I) had four item clusters: MPG, Chocolate, Drug Trial, and Correlation.

\textsuperscript{7}For one, errors are not included.
<table>
<thead>
<tr>
<th>Drug Trial 2</th>
<th>Confidence Interval</th>
<th>P-Value</th>
<th>Correlation</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chocolate 3</td>
<td>CR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 1 CR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 2 CR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 3 CR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocolate 3</td>
<td>SR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 3 SR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>SR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocolate 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Trial 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.13: Path diagram for FoI (Form I) analysis.
The RTM is a Rasch-family IRT model which has an advantage over other IRT models when it comes to interpretation of the resulting item parameter estimates. Estimates of the item step parameters can be transformed into so-called thresholds which have a straightforward graphical interpretation. The thresholds, or transition points, are represented graphically as the location where a student has a 50% probability of scoring at least as high as that given category. So, they can be interpreted as the point of “active learning” about the transition. A respondent who is placed, for instance, at LDC3 has not necessarily mastered that level, but instead could be thought of as currently working to master LDC3. Note that thresholds can be transformed to other criterion probabilities, but here we stick with the traditional 50% interpretation. For a pair of comparison items, the estimated thresholds representing the same transition points will be compared graphically in a scatter plot and using Wright Maps which are visualizations that plots the item estimates along with the person estimates (see Figures 3.15 through 3.17 which are described in the following section). These graphical displays are used to uncover potential patterns and describe the general relationships between pairs of comparison items in the field test data. No formal tests of significance are performed, but it is important to note that all item location estimates are estimates and have an associated standard error. Taking the LDC Wright Map (Figure 3.15) as an example, the on-its-side histogram on the left side represents the distribution of respondents along the logit axis which is interpreted as both respondent “ability” in LDC and item difficulty. The thresholds are organized horizontally by construct level. So, the Twitter 1 threshold located in the lower left-hand corner at approximately −1.75 logits is the point at which a respondent has a 50% probability of scoring at LDC1 or higher. There is another threshold for the Twitter 1 item on the Wright Map at the LDC3 level at approximately 1 logit. This is the point at which a respondent has a 50% probability of scoring into at least LDC3. Both of these thresholds are for the same item, but are associated with different score levels on the LDC construct. As most of the CR4CR Assessment items were polytomous, there will be more than one threshold with the same label associated with different construct levels. There will not be any item with more than one threshold at the same construct level.

The construct maps, items, and scoring guides are still subject to final confirmation. As noted in the results section, rater reliability indexes are lower than preferred, indicating that scoring may materials still need further development work. Once these materials are more stable, appropriate tests of significance will be more useful. For now, this work is largely exploratory.

The test forms were designed to investigate the behavior of our comparison item pairs, and it is important to note that person reliability was not a major concern in designing the study. Thus, the forms were not optimized for producing high person reliabilities. Reliability indexes will be reported, but we do not expect them to be very high at this intermediate stage of assessment design. Note that low person reliabilities do not affect the estimation of item features. Another type of reliability, rater reliability of CR items, is an important feature to examine in this study, however. Low interrater reliability and/or rater agreement would be a limiting factor of our conclusions regarding item CR and SR item comparisons.

For the open-ended items, rater reliabilities were calculated. The open-ended responses were scored by two or three trained raters. All open-ended items on all forms of the CR4CR
assessment were included in this study, but only the comparison item results are presented in this paper. Fleiss’ exact kappa (Fleiss, 1971; Conger, 1980), a rater agreement index for multiple raters, was calculated for each item. Kappa coefficients have their limitations. For one, they are formulated for nominal data while our data is ordered categorical. Further, guidelines for acceptable kappa values are somewhat arbitrary, and so meaningful interpretation can be difficult. For Fleiss’ kappa, we will use the guidelines that 0.40 to 0.75 is fair and greater than 0.75 is excellent rater reliability (Fleiss, 1971). Even with their limitations, they are a widely used method for reporting rater reliability and can help to identify items or scoring guides that can be improved. Pair-wise percent agreement was also calculated for each pair of raters within an item. For the items rated by only two raters, there is one pair-wise percentage and for the items rated by three raters, there are three. These have a simpler interpretation than a kappa value and can also be helpful in targeting efforts to improve the item and scoring design. As with other indexes of rater reliability, there are differing standards of acceptability. We will consider 75% pair-wise agreement acceptable.

3.5 Empirical analysis

3.5.1 Data

The CR4CR Assessment was piloted in Fall 2018 and the two parent forms described previously were used: Forms G and I. There were items on other constructs on each of these forms, and they were linked by 12 common items that spanned the constructs. Thus a larger, linked analysis is possible in the future with these data, but as the purpose of this study is to related the SR and CR versions of particular items within each of the constructs, we performed a separate unidimensional analysis for each form/construct.

There were 254 respondents in total and Table 3.3 shows how many respondents were included in each analysis. It is important to note that each respondent took only one form. So, respondents are unique both within any given analysis as well as across the analyses. Within each analysis, approximately half of the respondents took the child form 1 (G1 or I1) and the others took child form 2 (G2 or I2). Table 3.3 also shows the number of items in the analysis as well as the number of comparison item pairs. So, the total number of items on that particular construct that any single respondent took is the difference of those two numbers. For example, Forms G1 and G2, which targeted LDC, each contained \(14 - 2 = 12\) LDC items. Approximately half of the respondents took Form G1 (recall Table 3.2) with the CR version of the Twitter cluster and the SR version of Paternity Leave cluster (each containing one comparison item), while the other half took Form G2 with the comparison versions of those clusters.

Of the 254 respondents, 77% were current college students and 21% were high school students (sophomores, juniors, and seniors) at the time they took the assessment and 1% did not provide education level information. Because we wanted information at both low and high ends of the constructs, we recruited students in high school and early college, focusing on students who were not mathematics or statistics majors. 54% percent reported having taken
Table 3.3: Information about empirical analyses.

<table>
<thead>
<tr>
<th>Parent Form</th>
<th>Target Construct</th>
<th>Num. of respondents</th>
<th>Num. of items</th>
<th>Num. of comparison item pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>LDC</td>
<td>119</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>I</td>
<td>FoI</td>
<td>135</td>
<td>19</td>
<td>5</td>
</tr>
</tbody>
</table>

or being currently enrolled in statistics or probability coursework in high school. The gender identification distribution of the sample was 61% female, 37% male, and 0.5% non-binary (0.5% did not provide this information).

3.5.2 Results

Reliability indexes and the numbers of CR and SR items in each analysis for each of the three models run are presented in Table 3.4. The reliability indexes for each analysis are lower than what is traditionally acceptable for a standardized test. This was expected as the study was not designed for the traditional purpose of measuring students but rather to study item performance and hence to investigate the validity of the proposed structure of our constructs. The average number of items taken by respondents is also included. Reliability estimates, in general, increase as respondents take more items. The second paper of this dissertation shows using simulation that these low reliability indices are largely accounted for by the low variance of the respondent distributions; a larger variance in the person distribution would increase the reliability indices to acceptable levels (Arneson, 2019). We speculate that our sample was censored, especially at the higher end of the dimensions, because we recruited relatively few students who were enrolled in or had taken advanced mathematics or statistics. Histograms for the respondent distributions are provided in the left-hand panels of the Wright Maps discussed later (Figures 3.15 through 3.17). We expect this reliability to improve with future larger and more variable data collection designed with the interest of individual student measurement.

A small rater reliability study was also undertaken to investigate the reliability of the CR items in particular. The findings of that small study are given in Table 3.5. There were three raters in total – call them A, B, C, and D – and each CR comparison item was scored by at least two of them. Each row of the table provides the number of responses rated by each rater, Fleiss’ exact kappa ($\kappa$), and the pair-wise rater agreement percentages are shown for the CR versions of the comparison item pairs. We see that all of the comparison CR items show fair reliability in terms of $\kappa$ as they are all at least 0.4 but less than 0.75. Looking at the pair-wise agreement, however, there are concerns for both of the Chocolate items (the last two rows of Table 3.5) as pair-wise agreement is below 75% for many of the rater pairs. Conclusions regarding the Chocolate 3 (FoI-S) and Chocolate 4 (FoI-P) items will carry the caveat that the interrater reliability for the CR versions did not reach an acceptable threshold.

---

8The scores used for these analyses were those of rater A, the author.

94
Table 3.4: Model estimates and fit statistics.

<table>
<thead>
<tr>
<th>Parent form</th>
<th>Target dimension</th>
<th>EAP/PV reliability</th>
<th>Num. of CR items</th>
<th>Num. of SR items</th>
<th>Avg. num. of items taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>LDC</td>
<td>0.446</td>
<td>2</td>
<td>12</td>
<td>11.9</td>
</tr>
<tr>
<td>I</td>
<td>FoI-S</td>
<td>0.263</td>
<td>5</td>
<td>6</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>FoI-P</td>
<td>0.267</td>
<td>4</td>
<td>4</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Note: The EAP/PV reliability is a measure of test reliability or person separation reliability. It is calculated by dividing the variance of the individual EAP ability estimates by the observed person variance (Adams, 2005), and is equivalent to traditional reliability measures such as Cronbach’s alpha. There are as many EAP/PV reliability indexes estimated as dimensions in the model. Only the reliability indexes for the dimensions representing the target construct are presented here. Note that each model also included random item cluster effects that have an associated index, but they are not reported as they are not useful in this study.

Table 3.5: Item rater reliability for CR items.

<table>
<thead>
<tr>
<th>Item Name</th>
<th>A&amp;B</th>
<th>A&amp;C</th>
<th>B&amp;C</th>
<th>A&amp;D</th>
<th>B&amp;D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paternity Leave</td>
<td>86%</td>
<td>81%</td>
<td>76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>77%</td>
<td>80%</td>
<td>70%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 1</td>
<td>90%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 2</td>
<td>86%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 3</td>
<td>85%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocolate 3</td>
<td>76%</td>
<td></td>
<td></td>
<td>75%</td>
<td>60%</td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>69%</td>
<td></td>
<td>72%</td>
<td>61%</td>
<td></td>
</tr>
</tbody>
</table>

N is the number of responses rated by each rater. κ is Fleiss’ exact kappa given by Conger (1980).

The interrater reliability is, on the whole, lower than is traditionally acceptable for scaling a standardized test signals that either (1) the raters need further training, or (2) them items and/or scoring guides need some revision. The rater training that occurred before this reliability study involved two 90-minute sessions: the first was focused on becoming familiar with the three construct definitions and examining sample student responses and the second focused on rectifying scoring discrepancies on a different set of student response data. This is much less training than is usually done for widely-used standardized tests and with further training, these reliability indexes are expected to increase. Recall that this project is just wrapping up its field testing. More involved rater training is planned for the future.

Table 3.6 provides the overall item difficulty estimates for the comparison items, along with their standard errors. All of these items are polytomous and the model produces a step parameter for each transition point between adjacent score categories. The estimates...
Table 3.6: Item parameter estimates for comparison items.

<table>
<thead>
<tr>
<th>Item pair name</th>
<th>Dimension</th>
<th>CR difficulty (SE)</th>
<th>SR difficulty (SE)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paternity Leave 2</td>
<td>LDC</td>
<td>-0.003 (0.079)</td>
<td>-0.330 (0.093)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Twitter 3</td>
<td>LDC</td>
<td>-0.371 (0.104)</td>
<td>0.193 (0.148)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>MPG 1</td>
<td>FoI-S</td>
<td>0.012 (0.175)</td>
<td>-0.195 (0.091)</td>
<td>0.15</td>
</tr>
<tr>
<td>MPG 2</td>
<td>FoI-S</td>
<td>-0.006 (0.163)</td>
<td>-0.021 (0.103)</td>
<td>0.47</td>
</tr>
<tr>
<td>MPG 3</td>
<td>FoI-P</td>
<td>0.379 (0.188)</td>
<td>-0.776 (0.167)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Chocolate 3</td>
<td>FoI-S</td>
<td>0.385 (0.117)</td>
<td>0.954 (0.127)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>FoI-P</td>
<td>0.063 (0.107)</td>
<td>0.551 (0.079)</td>
<td>&gt; 0.99</td>
</tr>
</tbody>
</table>

Note: The difficulty estimates presented in this table are not the thresholds plotted in the following scatter plot and Wright Maps. Instead, these are averages of the step parameters estimated from the model. They are provided here to provide comparisons at the item level (rather than the step level) along with a measure of variability/uncertainty (standard errors). The p-value is produced by a t-test of the (one-sided) alternative hypothesis $\delta_{CR} > \delta_{SR}$, that the CR overall item difficulty is different from the SR overall item difficulty.

provided in Table 3.6 are the average of the step parameters for each item and can be interpreted as an “overall item difficulty.” The higher the overall estimate, the more difficult the item is, on average. The ConQuest software’s estimation algorithm directly estimates these parameters and calculates their standard errors (Wu et al., 2007). Recall that separate analyses were run for each construct. So, the difficulty for items within constructs can be compared, but not across constructs. Thus, the Paternity Leave 2 estimates and the Twitter 3 estimates can be compared in terms of magnitude as they are on the same scale. Estimates for Paternity Leave 2 and, for example, MPG 1 cannot be directly compared as they are not.

We might expect that SR items would be easier (have lower difficulty estimates) than CR items, as much of the literature review suggested that SR items measure lower-level skills than CR items. As shown in the right-most column of Table 3.6, the only items for which this seems to be true are Paternity Leave 2 and MPG 3. T-tests were performed on each pair of difficulties with the alternative hypothesis that the overall CR item difficulty would be higher than that of the SR item, $\delta_{CR} > \delta_{SR}$. The tests on Paternity Leave 2 and MPG 3 are the only ones that produced a significant p-value (at the 5% level). In fact, for Chocolate 3 and Chocolate 4, there is evidence that those SR items were more difficult than their CR counterparts. The design goal for this study, on the other hand, was that the SR versions would be comparable in difficulty to the CR versions. And, looking across items, this is, on balance, found to be the case. For the other items - Twitter 3, MPG 1, and MPG 2 - this data collection did not produce evidence that this is not the case, at least considering the items overall.

It is important to inspect the step parameters as these items were scored polytomously. For this exercise, it is important that estimates with a comparable interpretation are paired, comparisons should be drawn between step parameters that represent the same “step” or construct level transition point. To address this, Table 3.7 is organized so that step pa-
Table 3.7: Step parameter estimates for comparison items.

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Construct Level</th>
<th>CR step estimate (SE)</th>
<th>SR step estimate (SE)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pat. Leave 2</td>
<td>LDC1</td>
<td>-0.60 (0.29)</td>
<td>-0.19 (0.32)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>Pat. Leave 2</td>
<td>LDC2</td>
<td>-0.57 (0.29)</td>
<td>-0.47 (0.31)</td>
<td>0.96</td>
</tr>
<tr>
<td>Pat. Leave 2</td>
<td>LDC3</td>
<td>1.16 (0.28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter 3</td>
<td>LDC1</td>
<td></td>
<td>0.52 (0.30)</td>
<td></td>
</tr>
<tr>
<td>Twitter 3</td>
<td>LDC2</td>
<td>-0.37 (0.15)</td>
<td>-0.13 (0.31)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>MPG 1</td>
<td>FolI1S</td>
<td>-3.05 (0.40)</td>
<td>-2.37 (0.31)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>MPG 1</td>
<td>FolI2S</td>
<td></td>
<td>0.28 (0.29)</td>
<td></td>
</tr>
<tr>
<td>MPG 1</td>
<td>FolI3S</td>
<td>3.08 (0.36)</td>
<td>1.51 (0.30)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>MPG 2</td>
<td>FolI1S</td>
<td>-2.93 (0.37)</td>
<td>-2.39 (0.37)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>MPG 2</td>
<td>FolI2S</td>
<td></td>
<td>-0.37 (0.29)</td>
<td></td>
</tr>
<tr>
<td>MPG 2</td>
<td>FolI3S</td>
<td>2.92 (0.34)</td>
<td>2.70 (0.36)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>MPG 3</td>
<td>FolI1P</td>
<td>-3.18 (0.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG 3</td>
<td>FolI2S</td>
<td></td>
<td>-0.78 (0.17)</td>
<td></td>
</tr>
<tr>
<td>MPG 3</td>
<td>FolI3P</td>
<td>3.94 (0.38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocolate 3</td>
<td>FolI1S</td>
<td>-1.15 (0.29)</td>
<td>0.80 (0.36)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>Chocolate 3</td>
<td>FolI2S</td>
<td>1.92 (0.27)</td>
<td>1.10 (0.34)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>FolI1P</td>
<td>-1.03 (0.28)</td>
<td>0.34 (0.29)</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>FolI2P</td>
<td>1.16 (0.26)</td>
<td>1.10 (0.49)</td>
<td>0.21</td>
</tr>
<tr>
<td>Chocolate 4</td>
<td>FolI3P</td>
<td></td>
<td>0.21 (0.28)</td>
<td></td>
</tr>
</tbody>
</table>

Table notes: A blank cell indicates that the step parameter was not estimable with the available data. T-test results are only reported for estimates that have a comparable interpretation in that they are aligned to the same construct level. The ConQuest software provides standard error estimates by default for all but the highest category as it is constrained in the estimation process. Estimates for these standard errors were obtained by running the model a second time in the software, but specifying that a different step be constrained.

Parameters associated with the same construct level score (though their numeric scores may differ) are in the same row. The p-values for each of the comparison pair step estimates are provided in the right-most column of Table 3.7. Only three steps are significant at the 5% level (again with the alternative hypothesis that the CR version is more difficult than the SR version) and all are on the FoI-S dimension: MPG 1 at FoI3S, MPG 2 at FoI3S, and Chocolate 3 at FoI2S.

To visualize how pairs of step parameter estimates compare, they are plotted in Figure 3.14 in a scatter plot. The unmatched steps are not included in this scatterplot, only the step pairs with the same interpretations are shown (i.e., all of the pairs that have both a CR and SR step estimate in Table 3.7). If step difficulty estimates are exactly equal within a pair, they would lie on the $y = x$ line shown as a diagonal in the scatter plot. Items for which reaching a given score level is more difficult for the CR version are in the bottom right half of the plot, and if the SR version has a higher threshold, then it is located in the top left half. Items on different dimensions are differentiated by both color and shape of the marker.
Figure 3.14: Scatter plot of comparison item thresholds.
To overcome the interpretational difficulties of the step parameters produced by the ConQuest software, the following Wright Maps instead use the item thresholds. These are a nonlinear transformation of the item and step parameters, and as such the ConQuest software does not estimate standard errors for them. So, though interpretations may be more meaningful, we cannot perform formal statistical tests to determine observed differences are statistically significant. However, we can look for meaningful patterns and differences in the threshold estimates using the Wright Maps as exploratory graphical devices (Figures 3.15 through 3.17).

If we had realized our goal of writing SR versions of items that were empirically comparable to the original CR versions, the points in Figure 3.14 would all lie relatively close to the $y = x$ line. However, this is not the case. There are two FoI-S (blue squares, near the coordinates (3, 0)) thresholds that are well below the line, indicating that at these transition points, the SR version is substantially easier than the CR version. To further explore how these relationships fall out in terms of the construct level transition points that they represent, we can use Wright Maps. For this investigation, we are primarily concerned with the item estimates. A Wright Map was generated for each dimension.

Figure 3.15 is the Wright Map for the LDC (Form G) analysis. All of the conventions described in this paragraph apply to the Wright Maps for the FoI analysis as well. In order to make the relationships among thresholds for comparison items easier to decipher, some shape and coloring conventions have been adopted. All of the thresholds for comparison items are shown in red, blue, or green on the Wright Maps. For instance, in Figure 3.15 all of the thresholds for the Twitter 3 CR and SR items are shown in blue. Again, had we realized writing SR items that perfectly match the CR items in difficulty at each level, points of the same color would overlap at the same point. There are, however, gaps between the comparison thresholds as we saw in the scatter plot in Figure 3.14 with no points on the $y = x$ line. To differentiate CR and SR items, different shapes were used for the points. This applies even to the items on the assessment that were consistent across all forms and not part of any comparison pair. CR items are shown as triangles, while single-select SR items are diamonds, and multi-select SR items are squares. We wanted to differentiated between single- and multi-select item types as we had a hypothesis that multi-select SR items might be more difficult, in general, than single-select. The left side of the Wright Map is a histogram showing the person distribution along the logit scale.

As an internal structure validity concern, we should hope to see a Wright Map organized in this way by construct level to show approximate “banding” of levels – a generally increasing pattern of thresholds without much overlap between levels. We should hope to see little overlap – that, for example, LDC2 thresholds are located mostly lower than LDC3 thresholds. For the most part for our LDC items, this looks to be the case. LDC1 thresholds are below LDC2 thresholds which are below LDC3 thresholds. This does not necessarily mean there is

---

9We can do this since each score on the scoring guide is aligned to a construct level.

10It may be important to note at this point the distinction between the use of the terms dimension and construct. The LDC analysis had one substantive dimension which corresponds to the ordinal sequence represented by the LDC construct map. The FoI analysis has two substantive dimensions, one for each side of the FoI construct map.
Figure 3.15: Wright Map for the LDC (Form G) analysis.
no overlap among these levels, however. One limitation of these Wright Maps is that there is no depiction of uncertainty around the threshold estimates. As discussed earlier, the standard errors are not available for the thresholds. It should be noted that all interpretations of these plots carry the caveat that there is uncertainty around each point. Another important observation for the LDC items is that no one in this data set scored into LDC4 on any item, so there are no estimated thresholds associated with LDC4. This could be a weakness of the sample, that the students sampled were not high enough on LDC, or with the construct map itself. LDC4 may not be different enough from LDC3 to distinguish responses. Note that for the SR items, LDC4 was not a possible score for any of them. A weakness of the form design is that the only items that allowed for a score that high were CR items, and only two CR items were administered to each respondent.

The comparison thresholds, in general, seem to be relatively close together for Twitter 3 at LDC2. For all three of the paired comparison thresholds (Paternity Leave 2 LDC1, Paternity Leave 2 LDC2, and Twitter 3 LDC2), the CR version is lower than that of the SR version. Note that all of the SR versions were multi-select. So, we do not see that SR items are generally easier than CR items, at least for the LDC construct, as might be expected based upon the literature review. This gives indication that the multi-select item type may overcome at least this one of the perceived deficiencies of the traditional single-select multiple choice format. First, the probability of guessing is reduced due to the number of possible response combinations, and multi-select formats offer the flexibility of offering more than one technically correct choice, but having them differ in sophistication.

The Wright Maps for the FoI (Form I) analysis are split across two figures, Figure 3.16 and Figure 3.17, as the S and P sides of FoI were treated as separate dimensions in the model. One major limitation to note, for both FoI-S and FoI-P, is that the item thresholds are more widely spread than the respondents. Both have a fairly narrow histogram of respondents on the left side, but items that span well below and well above the minimums and maximums of the estimated respondent distribution. The following discussions and observations all carry this caveat.

In terms of banding for FoI-S (Figure 3.16), most FoI1 thresholds are below the FoI2 thresholds, excepting Correlation 1 and Chocolate 3 SR. And, most FoI3 thresholds are above FoI2 thresholds, except for P-Value and MPG 1 SR. We rarely expect to see perfect banding, and some overlap of levels does not invalidate the structure of our construct map as overlap at the lower and upper ends of each construct level are expected and likely not problematic. Especially in our context of mixed item formats, we expected that single-select SR items would be easier than other item types, which we do see evidence for in the Wright Map. At FoI1 and FoI2, the single-select SR items (diamonds) are easier than both the CR and multi-select SR items. At FoI1, however, the MPG 1 and MPG 2 CR items are located below than their SR counterparts. Again, as with LDC, we did not have any respondents score into the top level of the FoI construct.

For the pairs of comparison thresholds, we see some that are relatively close and others that are further apart. For MPG 1 (red), they are relative close at FoI1, but further spread apart at FoI3. The MPG 1 level 3 threshold is actually within the range of the FoI2 thresholds of other items, and being an entire construct level below is a meaningful difference. Increasing
Figure 3.16: Wright Map for the FoI-S items from the FoI (Form I) analysis.
Figure 3.17: Wright Map for the FoI-P items from the FoI (Form I) analysis.
the difficulty of the MPG 1 item, possible by adapting it to the multi-select format, is a next step to explore to make the SR and CR difficulties more comparable. Note that no respondent scored into FoI2 on MPG 1 nor MPG 2, so only SR thresholds are shown there. The MPG 2 thresholds are very close to the MPG 1 thresholds in each level except FoI3. The Chocolate 3 item, the FoI1 threshold for the multi-select SR version is considerably higher than for the CR version. In this case, revisions of the SR version to make it slightly easier or rethinking of the scoring guide may be in order. Since the thresholds at FoI2 are relatively closer, the scoring guide may be a place to start. It may be that some of the response combinations that are currently scored at FoI1 may be a better empirical fit at FoI2.

Similar patterns are evident in the FoI-P Wright Map in Figure 3.17. The multi-select items tend to be more difficult, there is general (but not perfect) banding, and no one scored at FoI4. Unfortunately, none of the MPG 3 thresholds (red) are matched within levels, so no comparisons can be drawn about them. They do, however, follow the general pattern of the Wright Map, with the FoI1 threshold (the CR version) well below the FoI2 threshold (the SR version) well below the FoI3 threshold (the CR version). The Chocolate 4 item thresholds at FoI1 are somewhat separated – with the SR version approaching FoI2. Note that the Chocolate 4 SR thresholds across the levels are quite close, the FoI1 threshold is only just below FoI2 which is just below FoI3. Revisions to the item or the scoring guide may be in order to get these within-item thresholds better separated.

It is important to stress again that our sample was not variable in terms of FoI-S nor FoI-P. It is a good exercise to examine the Wright Maps for patterns, but the discussion regarding the extremes of the construct (especially the higher end) cannot reach strong conclusions as that would require a sample that includes respondents at those extremes.

The scatter plots and Wright Maps tell the same story in slightly different ways. That, for the most part, the SR versions of items are of comparable difficult to their CR counterparts, except for some within the FoI dimensions. In many cases, the SR version has a slightly higher difficulty than the CR version, even. Where we may have not accomplished this goal, however, is for the FoI-S comparison items. In three cases, the step difficulty for the CR version of the item is significantly higher than that for the SR version. These are the three blue squares below the $y = x$ line in Figure 3.14 and the rows of Table 3.7 with $p$-values less than 0.01. Both of these MPG items are single-select SR items, so a next step could be to develop multi-select versions of those and test the difficulty of those items versus the CR format.

3.6 Discussion

In sum, this study does not, of itself, either prove or disprove the validity of the CR4CR Assessment, but serves as a piece of the validity argument, especially when it comes to the inclusion of SR items on as assessment of higher-order thinking skills. We found that, with a few exceptions and caveats, our SR items behaved in a similar way to our CR item types. Further, there is evidence that multi-select item types are just as or more difficult than CR items indicating that the multi-select item type may be a viable alternative to SR items in
measuring critical data-based reasoning. They can be machine scored and seem to overcome some of the limitations of the single-select format (e.g. guessing).

This study also highlighted the importance of comparing polytomous item estimates at the step level and not just by looking at overall item difficulty. Overall comparisons are most meaningful when the scoring guides of the comparison items are well aligned (e.g. the same construct levels are represented and manifest in the data).

Though there are many limitations of the field test data analyzed in this study, it serves as a good spring board to further improve the CR4CR Assessment and further investigate the relationship between CR and SR item types. We have offered preliminary empirical tests of the widely-held belief that SR items cannot measure higher order thinking skills in the same ways that CR items can. It seems this conclusion should not be applied in general, but may be dimension-specific. For certain higher-order skills (such as that represented by FoI-S), this may be the case. However, for other dimensions of interest, like LDC, SR item types may be empirically similar to CR items.

Final validation work for the CR4CR Assessment is ongoing. Cognitive interviews (“think-alouds”) have been conducted on both CR and SR item types, but will need to happen again as items undergo revisions. The qualitative evidence from these interviews is needed to support the empirical patterns uncovered in this paper. There are currently six constructs defined and nine more theorized for this project, so similar analyses on comparison items sets written to target the other constructs are also needed. It should not be discounted, however, that more data needs to be collected (especially for FoI) in order to explore the structure and validate the higher levels of these constructs. Existing SR items must be revised or new ones developed that target those higher levels of these constructs. Even if we had sampled students at those higher levels, we still may have ended up with limited information about them.
References


Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361–372.


111


112


Appendix A

Constructed response and selected response versions of items
Suppose you work for a magazine that plans to include an article on the rise in the number of companies offering paternity leave, paid time off for fathers in the weeks following the birth of a child, for new fathers in the workplace. Your editors want to include information about public support for this practice of paternity leave.

The magazine is developing a one-question online survey that will be emailed to their readership.

[1] Suppose the editors want to make the headline “___% of readers support paternity leave”. Which of the following survey questions is best to collect the information needed for this statement? Choose only one.

- Do you support paid time off for new parents?
  - Yes
  - No

- Do you think that new fathers should get paid time off when a child is born?
  - Yes
  - No

- Rate your agreement with the following statement: New parents should receive paid leave following the birth of a child
  - Strongly Agree
  - Agree
  - Neutral
  - Disagree
  - Strongly Disagree

- Did you take paternity leave at the time your last child was born?
  - Yes
  - No


Write your answer here.

[3] Suppose the editors want to make the headline “Support for paternity leave on the rise”. What can you do (in addition to sending out the survey question you indicated above) in order to collect the information needed to make this statement? Choose all that apply. You must choose at least one.

- Ask respondents to provide their gender on the survey.
- Send out the survey again after 1-2 years to the same people.
- Calculate the difference in proportions of men who agree and women who agree.
- Find out what percent of readers supported paternity leave 10 years ago.
- Find out what percent of working adults supported paternity leave 10 years ago.

[4] The link to the survey will be included in an email to all current subscribers of the magazine. Your editor claims that this is representative of the United States adult population since the readership of the magazine is very diverse in terms of gender and ethnicity. Do you agree with your editor’s claim?

- Yes
- No

Figure A.1: The constructed response version of the Paternity Leave item cluster.
Suppose you work for a magazine that plans to include an article on the rise in the number of companies offering *paternity leave*, paid time off for fathers in the weeks following the birth of a child, for new fathers in the workplace. Your editors want to include information about public support for the practice of paternity leave.

The magazine is developing a one-question online survey that will be emailed to their readership.

### [1] Suppose the editors want to make the headline "____% of readers support paternity leave". Which of the following survey questions is best to collect the information needed for this statement? **Choose one.**

- **Do you support paid time off for new parents?**
  - Yes
  - No

- **Do you think that new fathers should get paid time off when a child is born?**
  - Yes
  - No

- **Rate your agreement with the following statement: New parents should receive paid leave following the birth of a child.**
  - Strongly Agree
  - Agree
  - Neutral
  - Disagree
  - Strongly Disagree

- **Did you take paternity leave at the time your last child was born?**
  - Yes
  - No

### [2] Which of the following statements back up your choice in question [1]? *Note that a statement may be true but does not support your choice. Choose only the statements that support your choice. Choose all that apply. You must choose at least one.***

- It forces a person to make a definitive choice.
- It is important to know the degree to which a person agrees or disagrees.
- It references both new mothers and new fathers.
- It only references new fathers.
- It is important to know if someone has taken paternity leave.

### [3] Suppose the editors want to make the headline "Support for paternity leave on the rise". What can you do (in addition to sending out the survey question you indicated above) in order to collect the information needed to make this statement? **Choose all that apply. You must choose at least one.***

- Ask respondents to provide their gender on the survey.
- Send out the survey again after 1-2 years to the same people.
- Calculate the difference in proportions of men who agree and women who agree.
- Find out what percent of readers supported paternity leave 10 years ago.
- Find out what percent of working adults supported paternity leave 10 years ago.

---

Figure A.2: The selected response version of the *Paternity Leave* item cluster.
Miles per Gallon [S111v06]

An automotive industry lobbyist wishes to claim that cars have been getting more energy efficient over time, and thus the federal government should relax some of the environmental regulations they impose in the industry.

To back up this claim, they use Environmental Protection Agency (EPA) fuel economy data to compare the "city" miles per gallon (MPG) ratings of random samples of cars from 2012, 2013, and 2014. Using statistical software, they found the mean MPG along with 95% confidence intervals. These are reported in the following table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean MPG</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>19.41</td>
<td>(18.01, 20.81)</td>
</tr>
<tr>
<td>2013</td>
<td>20.73</td>
<td>(19.40, 22.06)</td>
</tr>
<tr>
<td>2014</td>
<td>22.03</td>
<td>(20.84, 23.22)</td>
</tr>
</tbody>
</table>

[1] The lobbyist makes the claim "There was a statistically significant improvement in fuel efficiency from 2012 to 2013." Do you agree with this statement? Why or why not?

Write your answer here.

[2] The lobbyist also makes the claim "There was a statistically significant improvement in fuel efficiency from 2012 to 2014." Do you agree with this statement? Why or why not?

Write your answer here.

[3] Is the observed 2.62 MPG improvement from 2012 to 2014 in fuel economy large enough to be of importance to an average driver? Why or why not?

Write your answer here.

[4] Recall that the lobbyist wants to use this analysis as evidence for a claim that government regulations can be relaxed. Without regard to the analysis results, do you think that he chose the right data to analyze? Why or why not?

Write your answer here.

Figure A.3: The constructed response version of the Miles per Gallon item cluster.
Miles per Gallon [S111v05]

An automotive industry lobbyist wishes to claim that cars have been getting more energy efficient over time, and thus the federal government should relax some of the environmental regulations they impose in the industry.

To back up this claim, they use Environmental Protection Agency (EPA) fuel economy data to compare the "city" miles per gallon (MPG) ratings of random samples of cars from 2012, 2013, and 2014. Using statistical software, they found the mean MPG along with 95% confidence intervals. These are reported in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Mean MPG</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>19.41</td>
<td>(18.01, 20.81)</td>
</tr>
<tr>
<td>2013</td>
<td>20.73</td>
<td>(19.40, 22.06)</td>
</tr>
<tr>
<td>2014</td>
<td>22.03</td>
<td>(20.84, 23.22)</td>
</tr>
</tbody>
</table>

[1] The lobbyist makes the claim "There was a statistically significant improvement in fuel efficiency from 2012 to 2013." Do you agree with this statement? Why or why not? Choose one.

- Yes, because the mean in 2012, 19.41, is contained in the 2013 interval, (19.40, 22.06).
- Yes, because the mean increased 20.73-19.41=1.32 MPG which is a significant amount.
- No, because the mean increased only 1.32 MPG which is not a significant amount.
- No, because the upper bound of the confidence interval in 2012 (20.81) is higher than the lower bound of the 2013 interval (19.40).
- No, because the lobbyist should not use the confidence intervals to determine significance. A t-test for the difference in means should be used.

[2] The lobbyist also makes the claim "There was a statistically significant improvement in fuel efficiency from 2012 to 2014." Do you agree with this statement? Why or why not? Choose one.

- Yes, because the mean increased 22.03-19.41=2.62 MPG which is a significant amount.
- Yes, because the upper bound of the confidence interval in 2012 (20.81) is below than the lower bound of the 2014 interval (20.84).
- No, because the mean in 2012, 19.41, is not contained in the 2014 interval, (20.84, 23.22).
- No, because the mean increased only 2.62 MPG which is not a significant amount.
- No, because the lobbyist should not use the confidence intervals to determine significance. A t-test for the difference in means should be used.

Figure A.4: The selected response version of the Miles per Gallon item cluster.
Chocolate Happiness [S112v11]

A psychologist conducts an experiment to see whether consuming chocolate increases happiness. For this experiment, 340 undergraduate students from her university volunteered to participate and were administered a 10-item happiness scale. Possible scores are 1-10 with a score of 10 meaning the "most happy." They were then given a chocolate bar and instructed to eat it. Twenty minutes later, they took the 10-item scale again. For each participant, the pre-chocolate score was subtracted from the post-chocolate score. The mean difference in scores was found to be 3.5. The 95% confidence interval was (2.2, 4.8).

[1] Consider the following incomplete concluding sentences from a report on this study. Choose the best option for filling in each of the blanks.

had higher happiness scores on average after eating a chocolate bar. This finding is


Write your answer here.


Write your answer here.

[4] Is the difference in happiness scores large enough to be of practical significance?

- Yes.
- No.
- I can't tell from the information provided.


Write your answer here.

Figure A.5: The constructed response version of the Chocolate item cluster.
A psychologist conducts an experiment to see whether consuming chocolate increases happiness. For this experiment, 340 undergraduate students from her university volunteered to participate and were administered a 10-item happiness scale. Possible scores are 1-10 with a score of 10 meaning the “most happy.” They were then given a chocolate bar and instructed to eat it. Twenty minutes later, they took the 10-item scale again. For each participant, the pre-chocolate score was subtracted from the post-chocolate score. The mean difference in scores was found to be 3.5. The 95% confidence interval was (2.2, 4.8).

[1] Consider the following incomplete concluding sentences from a report on this study. Choose the best option for filling in each of the blanks.

---

had higher happiness scores on average after eating a chocolate bar. This finding is _ **Select one**_.

---

[2] Which of the following statements support your completion of the first sentence? Choose all that apply. You must choose at least one.

- The participants in the study are a representative sample.
- 340 is a large enough sample size.
- It’s the most specific I can be.
- The effect can be generalized.
- There were no younger children nor adults included in the study.
- The sample was not selected randomly.

[3] Which of the following statements support your completion of the second sentence? Choose all that apply. You must choose at least one.

- We do not know the scores for everyone, just the mean.
- 3.5 is not a significant increase.
- 340 is a large enough sample size.
- 340 is not a large enough sample size.
- The confidence interval does not contain 0.
- The results may not apply to everyone.

[4] Is the difference in happiness scores large enough to be of _practical significance_?

- Yes.
- No.
- I can’t tell from the information provided.

---

Figure A.6: The selected response version of the _Chocolate_ item cluster.
Appendix B

Reliability simulation

There were two factors in the simulation study, summarized in Tables B.1 and B.2, one with two levels (the variance of the person distributions) and one with three levels (the form design), resulting in six model estimations per repetition. One-hundred repetitions were run and the EAP reliability indexes, variances of, and covariances between each of the three substantive dimensions were saved. The simulation study was conducted using the TAM (Kiefer, Robitzsch, & Wu, 2018) package in R.

Within each repetition, two full response data sets were simulated, one using a multivariate normal distribution with the empirical variance/covariance matrix (reported in Section 2.3.3) and one using a multivariate normal distribution with a wider variances for each of the three substantive dimensions. For these, the variances of the testlet dimensions were unchanged and the covariances among the substantive dimensions were fixed so that the correlations among them was the same as in the empirical results (covariances for all testlet dimensions were fixed to zero).

These two generated data sets were used in three ways: (1) as a “full” data set as though each respondent took all of the items across all forms with no missing item responses/scores, (2) as “parent” forms as though each respondent took all of the items on the parent form they were assigned (Form G, H, or I), and (3) as “child” forms in which the same form design that was used in the field testing were used (Form G1, G2, H1, H2, I1, or I2). For

<table>
<thead>
<tr>
<th>Generating values</th>
<th>LDC var (SD)</th>
<th>MRC var (SD)</th>
<th>FoI var (SD)</th>
<th>LDC/MRC cov (corr)</th>
<th>LDC/FoI cov (corr)</th>
<th>MRC/FoI cov (corr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical</td>
<td>0.292</td>
<td>0.522</td>
<td>0.087</td>
<td>0.290</td>
<td>0.109</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>(0.540)</td>
<td>(0.722)</td>
<td>(0.295)</td>
<td>(0.742)</td>
<td>(0.684)</td>
<td>(0.718)</td>
</tr>
<tr>
<td>Wider</td>
<td>2.250</td>
<td>2.250</td>
<td>2.250</td>
<td>1.670</td>
<td>1.539</td>
<td>1.616</td>
</tr>
<tr>
<td></td>
<td>(1.500)</td>
<td>(1.500)</td>
<td>(1.500)</td>
<td>(0.742)</td>
<td>(0.684)</td>
<td>(0.718)</td>
</tr>
</tbody>
</table>
Table B.2: Details on simulation form designs.

<table>
<thead>
<tr>
<th>Missingness pattern (form design)</th>
<th>LDC items</th>
<th></th>
<th></th>
<th></th>
<th>MRC items</th>
<th></th>
<th></th>
<th></th>
<th>FoI items</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Max</td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Max</td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Max</td>
<td></td>
</tr>
<tr>
<td>No missing</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td></td>
<td>13</td>
<td>13</td>
<td>13</td>
<td></td>
<td>19</td>
<td>19</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Parent forms</td>
<td>3</td>
<td>8.8</td>
<td>14</td>
<td></td>
<td>3</td>
<td>5.6</td>
<td>14</td>
<td></td>
<td>8</td>
<td>12.7</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Child forms</td>
<td>0</td>
<td>7.2</td>
<td>12</td>
<td></td>
<td>0</td>
<td>4.1</td>
<td>11</td>
<td></td>
<td>0</td>
<td>9.8</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

Table B.3: Simulation study results, reliability estimates.

<table>
<thead>
<tr>
<th>Generating variance/covariances</th>
<th>Missing/form design</th>
<th>LDC</th>
<th>MRC</th>
<th>FoI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical</td>
<td>Child forms</td>
<td>0.438</td>
<td>0.399</td>
<td>0.352</td>
</tr>
<tr>
<td>Empirical</td>
<td>Parent forms</td>
<td>0.496</td>
<td>0.497</td>
<td>0.400</td>
</tr>
<tr>
<td>Empirical</td>
<td>No missing</td>
<td>0.649</td>
<td>0.705</td>
<td>0.502</td>
</tr>
<tr>
<td>Wider</td>
<td>Child forms</td>
<td>0.789</td>
<td>0.738</td>
<td>0.837</td>
</tr>
<tr>
<td>Wider</td>
<td>Parent forms</td>
<td>0.818</td>
<td>0.800</td>
<td>0.866</td>
</tr>
<tr>
<td>Wider</td>
<td>No missing</td>
<td>0.894</td>
<td>0.904</td>
<td>0.906</td>
</tr>
</tbody>
</table>

the parent and child forms, the same simulated responses were used, but items that did not appear on the assigned form for a given respondent were recoded as missing. The same form assignment was used as was used in the field test data.

The average reliabilities for each of the three substantive dimensions are provided in Table B.3. If we use a common rule of thumb that reliabilities of 0.8 and above are acceptable, we see in the first three rows of the table that on average, the narrow empirical variances of the person distributions do not produce acceptable reliability indices. With wider person distributions, however, the form design used in the field testing yields reliabilities all at least above 0.7. A modest increase in the number of items per form (the Parent forms), would, on average, yield acceptable reliability.