Title

Visualizing Item Response Data for Personalized Learning: Development of an Interaction Map Approach for Educational Assessments

Author

Ho, Eric Ming-Yin

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Visualizing Item Response Data for Personalized Learning: Development of an Interaction Map Approach for Educational Assessments

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Eric Ming-Yin Ho

2023

ABSTRACT OF THE DISSERTATION

Visualizing Item Response Data for Personalized Learning: Development of an Interaction Map Approach for Educational Assessments

by

Eric Ming-Yin Ho
Doctor of Philosophy in Education
University of California, Los Angeles, 2023
Professor Minjeong Jeon, Chair

Personalized learning, which has the potential to raise student achievement, requires understanding the competencies of students. Visualizations can help provide this understanding. Jeon et al. (2021) presented a latent space model that creates interaction maps visualizing response patterns from item response data. My dissertation proposes extending this latent space model to yield profiles that can visualize individual student competencies and developing a practical tool based on those profiles that can help promote personalized learning. Five aims will be achieved within the proposed study. First, I will investigate different formulations of the model that can yield different kinds of profiles. Second, I will introduce these profiles created from the interaction maps and the information they can convey. Third, I will investigate the validity and reliability of these profiles. Fourth, I will present empirical applications of this approach based on real-life datasets and validate this approach. Finally, I will present a web-based application that can create interaction maps and profiles from uploaded datasets.

The dissertation of Eric Ming-Yin Ho is approved.

James W. Stigler

Mark P. Hansen

Noreen M. Webb

Minjeong Jeon, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I owe my sincerest gratitude to the many people who supported me throughout my graduate school journey. This section could be a dissertation in itself because I am incredibly fortunate to have had so many colleagues, allies, and friends accompany me these past few years.

I would like to thank Minjeong Jeon, my advisor and chair of the committee, for her indefatigable support from the beginning of my program to the end during my dissertation phase. She has been one of my biggest advocates, always encouraging me through setbacks and presenting new opportunities for collaboration and research. Her guidance during the writing and presentation of my dissertation is indispensable, and I could not have completed it without her support. I am also very honored to have met her wonderful family.

I would also like to thank Noreen Webb, who was the first person from the Department of Education to reach out about my candidacy for admission. Through her, I was able to begin my collaboration with Professor Jeon. Even though I did not work as closely with her as I did with MJ, Reenie has supported me throughout the years, helping me with my 299 project and with the dissertation.

I am also grateful for Mark Hansen, who provided useful resources for my dissertation and offered me the opportunity to tutor the Educational Leadership Program doctoral students, and for James Stigler, who graciously provided the CourseKata dataset, without which this dissertation may not have been possible. I am also thankful to Karen Givvin and Ji Son for their substantive contributions and suggestions.

Even though they were not on my committee, many other faculty members have guided me throughout the years. I am especially grateful to Michael Seltzer for being the first faculty member to welcome me at orientation and for his feedback on my 299 project; the inimitable late Mike Rose, for turning the chaos of my thoughts into sharp, clear prose; and Ananda Marin and Nicole Mancevice for introducing me to the new

world of qualitative research.

Special thanks are in order for my fellow members of the Jeon Lab, Jinwen Luo and Minho Lee, with whom I have worked closely on these projects and greatly enjoyed lunches and dinners. I am humbled to have met so many other wonderful peers and colleagues: Eric Setoguchi, Mariana Barragan Torres, Meredith Langi, Jessica Schnittka, Emi Fujita-Conrads, Alex Kwako, Maria Fernandez-Paz, Renee White Eyes, Nadia Sabat Bass, Zhaopeng Ding, Teanna Feng, Shujin Zhong, Yon Soo Suh, and many others.

I would like to thank my family for their support and my dear friends, Arthur Wang, Austin Lau, and Jimmy Tran, for being there for me during this journey. Last but most certainly not least, I am eternally grateful for the patience and support of my wife, Elena Meng, whom I met at the beginning of my graduate school career and who will be there for me at its end, and beyond.

VITA

EDUCATION

2018 - 2019 Master of Arts in Education
University of California, Los Angeles

2013 - 2017 Bachelor of Arts in Data Science
Columbia University in the City of New York

WORK

2020 - 2023 Statistics Consultant
Educational Leadership Program
University of California, Los Angeles

2019 - 2022 Teaching Assistant, Associate, Fellow
Professors Glory Tobiason, Jose-Felipe Martinez, Minjeong Jeon, and Kevin Eagan
University of California, Los Angeles

2019 - 2021 Graduate Student Researcher
Professors Minjeong Jeon and Glory Tobiason
University of California, Los Angeles

2017 - 2018 Strategic Data Analyst
Alliance College-Ready Public Schools

PUBLICATIONS

Luo, J., Jeon, M., Lee, M., Ho, E., Pfammatter, A.F., Shetty, V., & Spring, B. (2022). Relationships between changing communication networks and changing perceptions of psychological safety in a team science setting: Analysis with actor-oriented social network models. *PLOS ONE, 17*(8), Article e0273899. https://doi.org/10.1371/journal.pone.0273899

Ho, E., Jeon, M., Lee, M., Luo, J., Pfammatter, A.F., Shetty, V., & Spring, B. (2021). Fostering interdisciplinary collaboration: A longitudinal social network analysis of the NIH mHealth Training Institutes. *Journal of Clinical and Translational Science, 5*(1), Article e191. https://doi.org/10.1017/cts.2021.859

CHAPTER 1

Introduction

Education agencies are prioritizing personalized learning – the shaping of instruction to meet students' needs – to support student learning and raise student outcomes (Data Quality Campaign, 2019) given promising research that addressing the diverse needs of students helps raise student achievement (Broderick et al., 2005). Despite the promise of personalized learning, one challenge is that educators require resources to help them properly leverage assessment data to foster personalized learning since these assessments can help stakeholders understand students' strengths and weaknesses (Data Quality Campaign, 2019; U.S. Department of Education, Office of Educational Technology, 2017). The U.S. Department of Education suggests that one way to address this challenge is "encouraging the development of data assessment tools that are more intuitive and include visualizations that clearly indicate what the data mean for instruction" (U.S. Department of Education, Office of Educational Technology, 2017, p. 58).

An example of such a visualization is an interaction map (also called a latent space) created by the latent space item response model (LSIRM) introduced by Jeon et al. (2021) based on item response data. This interaction map visualizes items and respondents on a two-dimensional geometric space, like a scatterplot in which items and respondents are represented by points. The positions of the items and respondents relative to one another are based on whether certain respondents are more likely to answer certain items correctly, even after accounting for item easiness and respondent ability. I discuss the usefulness of the interaction map and the latent space model in the following sections because the map and model form the basis of my dissertation.

1.1 Importance of Visualization as a Tool

Although based on the latent space model which is further based on the Rasch model (Rasch, 1961), the interaction maps are quite different from traditional item response theory (IRT) models used to analyze item response data since in contrast to model-based approaches which largely yield numerical summaries, these maps can be seen as one kind of exploratory data visualization in the tradition of exploratory data analysis (EDA). Exploratory data visualization has a long history over the last half century of research methods literature, as noted across the Exploratory Data Analysis (EDA) literature (Behrens, 1997; Tukey, 1962) in which EDA functions to "address the broad question of 'what is going on here?' [with] an emphasis on graphic representations of data...[in which] the goal of EDA is to discover patterns in data" (Behrens, 1997, p. 132). Indeed, in research on the recent impact of the important advances in statistics and data analysis over the last half century, Gelman and Vehtari (2021) note

> Following Tukey (1962), the proponents of exploratory data analysis have emphasized the limitations of asymptotic theory and the corresponding benefits of open-ended exploration and communication (Cleveland, 1985) along with a general view of data science as going beyond statistical theory (Chambers, 1993; Donoho, 2017). This fits into a view of statistical modeling that is focused more on discovery than on the testing of fixed hypotheses, and as such has been influential not just in the development of specific graphical methods but also in moving the field of statistics away from theorem-proving and toward a more open and, we would say, healthier perspective on the role of learning from data in science. An example in medical statistics is the much-cited article by Bland and Altman (1986) that recommended graphical methods for data comparison in place of correlations and regressions (p. 4).

Thus, while the interaction map approach differs from traditional model-based IRT approaches that have seen extensive use in analyzing assessment data, as an example

of exploratory data analysis and visualization, the interaction map approach is a part of a rich methodological tradition that affords many benefits. For instance, the interaction maps allow for open-ended exploration of response patterns in assessment data with much fewer restrictive assumptions.

Additionally, educators have found exploratory data visualization useful. As previously discussed, Bowers and Krumm (2021) described a partnership between researchers and education leaders in a school to pattern and visualize multiple strands of summative assessments for 476 students in Algebra I using hierarchical cluster analysis (HCA) heatmaps visualizing more than 4,000 individual data points. Reflecting on the collaborative work with the data scientists, a school leader noted

> I think that this [HCA heatmap discussion] just opened up a huge frame of conversation for us to have with course-level teams and provide them data that I think they'll be able to dig deep on and start to revise a lot of these courses...It was a totally different way of visualizing data that I think we haven't seen before...It was just a really, really interesting way to think about data. Because we think about it in simpler terms here and so it's nice to see the larger possibilities with what we can do with the data that we have (Bowers & Krumm, 2021, p. 9).

Thus, I argue that the interaction maps, while quite different from existing approaches, are part of a well-studied methodological tradition of exploratory data visualization which has seen increasing acceptance and usefulness among educational practitioners. These practitioners, such as the ones interviewed by Bowers and Krumm (2021), may then be more receptive to a practical tool based on the interaction maps and extensions of these maps as a way to help them implement personalized learning.

1.2 A Promising Approach

My dissertation is based on the interaction maps and latent space model, given their potential usefulness and appeal to practitioners and the limitations of existing approaches. However, I believe that the method presented in Jeon et al. (2021) has even greater potential and promise for practical use in education. Therefore, the purpose of this study is to introduce and investigate practical extensions and interpretations of the model presented by Jeon et al. (2021); this study is, in a way, the practical "sequel" to the methodological Jeon et al. (2021) study. These extensions can be combined into one such data assessment tool, as described by the U.S. Department of Education, to address this research problem of helping educators identify students' strengths and weaknesses so they can personalize learning.

While the model and map can generally be used for any assessment, they can be particularly useful for educational assessment data. The relative distances between the respondents and the items, as seen on the interaction map, can help educators understand whether certain students are more or less likely to answer certain items on an educational assessment correctly, even after accounting for their abilities and the difficulty of the items. For example, a larger distance from a respondent to an item means that the respondent is less likely to answer that item correctly. These distances could be interpreted as a respondent's degree of strength (or weakness) with regards to that item. Hence, functions of these distances can be used to create unique strength or weakness profiles for each student. These profiles can help an educator understand which items a student is more or less likely to answer correctly after accounting for their abilities and item difficulties.

Additionally, as suggested in Jeon et al. (2021), items on the interaction map are naturally clustered according to the shared content or standards they might assess. Therefore, it is possible for educators to not only see which specific items students may struggle with but also which specific groups of items. These groups of items may corre-

spond to specific domains or subdomains of the content being assessed and can provide additional useful information to educators. In this regard, the interaction map may provide useful diagnostic information regarding the items. For example, if items that supposedly assess the same concept are far from one another on the interaction map, that may suggest that those items are not all testing the same concept. Then the test designers or educators may further investigate these items to understand why. The student profiles, based on this unique feature of the interaction map, can then reflect not only student competencies on individual items but also show groups of items corresponding to various domains. Therefore, the map and the profiles may not only interest educators who are looking to target students who need more support but also interest test developers or designers.

Thus, my goal in this dissertation is to make this model useful for educational practitioners, including educators and test developers. This includes creating optimal presentations of the interaction maps and the individual student profiles so that educators can know for a given student which items are difficult for them. I will propose different ways of formulating the model, map, and profiles that have substantive interpretations for educators and to investigate the reliability and validity of the proposed measures used in those formulations. I also hope to present a prototype of a user-friendly application that implements this approach for practitioners.

1.3 Scope of the Dissertation

The general goal of this dissertation is to take the approach described in Jeon et al. (2021) and make it useful for educators who wish to personalize learning. But to do so, there are a few specific tasks that must be achieved. Specifically, I aim to

1. Develop various ways of presenting the interaction maps and strength/weakness profiles of individual students. These ways present improved visualizations and provide additional information over those presented in Jeon et al. (2021). These

profiles could be used by educators to understand individual student competencies.

2. Investigate various functional forms of the distance term in the model in terms of giving substantive interpretations that are beneficial to instructors and students. The distance term can be formulated in different ways in the model to yield different kinds of student profiles. For instance, educators may rather view measures of student strength versus measures of weakness.

3. Evaluate the reliability of the measures of the proposed approach. The student profiles are based on functions of the distances from the student to the items. The reliabilities of these distances, under different conditions, can be evaluated using test-retest reliability. This is important since users may wish to know under which conditions these profiles are reliable before using them.

4. Evaluate and demonstrate the validity of the proposed approach through an application of the tool to empirical datasets. Specifically, the replicated student profiles should make sense given the observed item response data. For example, if a student were to get an item incorrect, the replicated student profiles should reflect that particular weakness. I hope to demonstrate the usefulness of this approach on real-life datasets.

5. Develop a user-friendly Shiny application. I will present a prototype of an application which would allow users to upload their own data and present the interaction map and student profiles, along with other information such as the parameter estimates. This is the tool that I hope educators may find useful.

1.4 Significance

The significance of this study is that this model can be turned into a practical tool to help educators provide personalized instruction to support student learning. The

interaction map is an intuitive visualization which does not require advanced statistical knowledge to interpret. An educator can look at the map and understand which students are struggling with certain items or groups of items. For a finer-grained analysis, the educator can look at a given student's profile (essentially a bar graph) to see which items the student struggled with. This tool/model enjoys numerous advantages over traditional, existing approaches to identifying students' strengths and weaknesses. For instance, the latent space model requires less stringent assumptions than the traditional Rasch model (Jeon et al., 2021) and no a priori assumptions about the Q-matrix unlike diagnostic classification models. The promise is that educators can take advantage of this practical tool based on a psychometrically-sound and non-restrictive latent variable model.

Test developers can also benefit from this interaction map approach. Test developers may have designed items to assess certain concepts. The interaction map can show clusters of items which would suggest that those items assess similar concepts. In that way, the interaction map can validate the test design and show whether the items are working according to the test design. For example, if items that are supposed to assess a given concept are far apart in the interaction map, that could suggest that the items might not actually be assessing the same concept. These patterns can also be shown on the respondent profiles, and test developers could investigate why respondents respond differently to items.

Additionally, the interaction map allows educators to focus on certain subpopulations of students, such as traditionally disadvantaged subpopulations of students. The respondents on the interaction map can be colored by demographic background variables, allowing educators to identify any such clusters and to assess whether certain demographics of students are having trouble with certain items or groups of items. This is particularly relevant given the adverse effects of the COVID-19 pandemic on students of color and low-income students (Ceres, 2020; Dorn et al., 2020).

With the transition to online learning, online assessment data has become much

more readily available; the interaction map is well-suited to turn this data into student profiles for educators to reach actionable insights regarding the performance of such students during the pandemic and beyond. Using the web-based application that will be created for this dissertation, educators can upload online assessment data to this application to better understand their students' performance. Furthermore, online assessment data can accumulate information from multiple assessments over time. This approach could potentially be used to show learning progression for different students by updating the interaction map as new assessment data comes in. I demonstrate this utility with applications of this method to online assessment data from a class that uses an online textbook.

CHAPTER 2

Literature Review

In the literature review for my dissertation, I will first discuss the idea and promise of personalized learning. I will then focus on the limitations of existing methods, including IRT models, that use assessment data to further personalized learning. Finally, I will explain the latent space item response model in Jeon et al. (2021) which underlies the interaction map approach.

2.1 Personalized Learning and Data Use in Schools

While there is not a clear universal definition of personalized learning, Pane et al. (2017) suggest that

> Personalized learning prioritizes a clear understanding of the needs and goals of each individual student and the tailoring of instruction to address those needs and goals. These needs and goals, and progress toward meeting them, are highly visible and easily accessible to teachers as well as students and their families, are frequently discussed among these parties, and are updated accordingly (p. 6).

In this study, I focus on the first aspect of personalized learning, namely the "understanding of the needs" of each individual student through the analysis of their performance on individual items. This usage of assessment to drive personalized instruction can raise student achievement (Black & Wiliam, 2010). More recently, McCarthy et al. (2020) described and evaluated a successful strengths-based blended personalized

learning model implemented at a school district in which teachers measured student strengths with assessments and tailored instruction according to those strengths and needs. In recent years, there has been a concerted push to use big data, learning analytics, and other advanced technology to create personalized learning environments for students using large quantities of student information, including their standardized test scores and other learning outcomes (Roberts-Mahoney et al., 2016).

However, missing from much of the research literature is a discussion of strategies for involving data in evidence-based decision-making for school improvement (Bowers et al., 2014). Thus data use in schools, or how data can be used for school improvement, has become an increasingly important area of research (Bowers & Krumm, 2021). Evidence seems to suggest that schools rely on very unsophisticated and coarse analyses. For instance, Selwyn et al. (2021) noted that "data analytics" in the schools they studied involved "simple frequency counts, colour-coding and modest cross-tabulations" (p. 84). They concluded that

> Any discussion of school datafication, therefore, needs to be set against the observation that schools do not appear to be particularly motivated to respond to (or even look for) novel insights and unexpected patterns and correlations in their data. In short, school data is not a place for surprises, counterintuition and 'outside-the-box' thinking. (p. 86)

It seems reasonable to assume, then, that educators are not fully leveraging their educational or assessment data to gain more insightful information. These current practices, compared to an item-level analysis, do not provide teachers with detailed information that could help teachers make more informed decisions regarding the needs of their students nor could they allow teachers to diagnose potentially problematic items. In contrast, Bowers and Krumm (2021) demonstrated the use of a hierarchical cluster analysis heatmap to visualize the number of attempts each student made on various summative assessments; in the study, school leaders praised the tool, intrigued by the

new insights provided by the heatmap. Like Bowers and Krumm, my contribution is to provide an additional tool which, while built upon the more sophisticated latent space model, is also easy and intuitive to use for educators who may want to uncover more useful information and visualizations from assessment data than can be ascertained from simple frequency counts or numerical summaries.

2.2 Review of Existing Approaches

There are existing approaches that have been used to identify student competencies and could also potentially be used to promote personalized learning. For example, recommendation systems have become an integral part of personalized learning since they can recommend the proper learning materials to students based on current information; a measurement model like the cognitive diagnosis model is used for these recommendation systems (Tang et al., 2019). However, these existing approaches, whether or not they involve IRT, have various limitations.

2.2.1 Measures of Central Tendency

Hoover and Abrams (2013) investigated how teachers use summative student assessment data to shape instruction. They found that most teachers tend to use central tendency statistics (such as the mean or median) to understand student assessment data rather than disaggregating data by content standards or subgroups. Thus, "Despite ample opportunities to analyze and use assessment data to inform instruction, the results suggest that teachers are not taking full advantage of the formative potential of summative assessment data" (Hoover & Abrams, 2013, p. 229). This study shows that teachers could use more advanced technology (such as the tool I propose in this dissertation) and leverage item analysis more fully to disaggregate and better understand assessment data. While item analysis through the IRT framework could yield more useful insights, these methods also have their limitations.

2.2.2 Multidimensional IRT Models

In multidimensional IRT (MIRT) models , the probability of answering an item correctly depends on multiple ability dimensions. That is because different items may assess different competencies. These models are useful since like the interaction map approach, they provide finer-grained information about the various competencies of students instead of reducing those estimates into a single parameter like in the unidimensional model: "Correspondingly, the MIRT measurement model will provide individual ability profiles as test results rather than single scores" (Hartig & Höhler, 2009, p. 58). A summary of various MIRT models is presented by Hartig and Höhler (2009).

MIRT models have been used to foster personalized learning. For example, Park et al. (2019) used a MIRT model to monitor students' various competencies to help provide personalized instruction within online learning environments. However, in the MIRT model, "the number of dimensions that should be included in the model is not a trivial problem and needs to be considered" (Hartig & Höhler, 2009, p. 60). It is also not always immediately clear which items correspond to which dimensions. In contrast, no a priori knowledge is required in the interaction map approach. The interaction map approach is a unique, exploratory method which naturally displays the competencies of students and the similarities of items based on response patterns.

2.2.3 Cognitive Diagnosis Models

Cognitive diagnosis models (CDMs; also referred to as diagnostic classification models or DCMs) are another existing approach that could also provide unique profiles of student strengths and weaknesses. Unlike traditional multidimensional IRT models which specify continuous latent traits, CDMs classify students into various discrete classes, based on the attributes they are theorized to have mastered. However, CDMs also have various limitations. For example, such models do not take heterogeneity among students within classes or class profiles into account, which may result in coarse con-

clusions about the strengths and weaknesses of students. Further, in practice, applications of CDMs are restricted by the number of attributes because the number of latent classes exponentially grows with the number of attributes. For example, with seven attributes, one needs to deal with 128 latent class profiles, which may be too many, posing computational and interpretive issues for a small classroom assessment. The latent space approach avoids these issues since this approach does not create latent class profiles. Regardless, natural clusters (analogous to the latent class profiles) of students and items can be visualized on the interaction map. As described later, the model can create varying strength/weakness profiles that can capture the heterogeneity of students, even within the same clusters. CDMs do not account for such heterogeneity for respondents with the same attribute profile.

Additionally, standard CDMs require a pre-specification of the item-attribute matrix specifying which attribute(s) each item measures. Misspecification of the item-attribute matrix can bias conclusions (de la Torre, 2009; DeCarlo, 2011). CDMs based on empirically driven item-attribute matrices have been proposed (Chen et al., 2018; Chen et al., 2015; Chiu, 2013), but those are not yet available for practical use. The latent space approach does not require item group (or attribute) information for model estimation. Items will naturally be grouped together on the interaction map if supported by the data.

2.2.4 The Saltus Model

The Saltus model (Wilson, 1989) can show differences in performance on dichotomous items due to student growth or differences in developmental stages. The model assumes that respondents can be classified into different stages. There are sets of items that correspond to each stage as well. The model states that respondents within a given stage answer all items in a manner consistent with other respondents in that stage. For example, respondents within a given developmental stage (e.g. pre-operational stage) may have the same advantage in answering a certain set of items (perhaps those that

respondents in the pre-operational stage are expected to answer correctly). An application of the Saltus model is presented in Draney and Wilson (2007). Similarly, the Saltus model could be used to promote personalized learning in identifying students at varying cognitive stages and allowing teachers to target particular groups of students.

This model enjoys certain advantages over CDMs. For example, while CDMs do not take heterogeneity within classes into account, the Saltus model allows respondents' proficiency to vary within the same class. Furthermore, the Saltus model is appropriate for modeling respondent-item interactions, since a respondent's underlying latent class membership would affect their ability to answer certain items correctly.

Regardless, there are certain limitations to the Saltus model. Firstly, the number of latent person classes must be pre-specified. Secondly, the item-group memberships must also be known a priori. In the interaction map approach, the number of latent person classes and the item-group memberships do not need to be known a priori. The interaction map allows for heterogeneity within person classes without requiring pre-specification of the number of person latent classes. Items will also be clustered together naturally on the interaction map. The interaction map approach also provides a useful visualization, unlike the Saltus model.

2.2.5 Wright (Construct) Maps

One option for modeling items and respondents in a shared space is a Wright (or construct) map (Wilson, 2003). The Wright map displays items and respondents along a shared continuum of a single construct. On one side, the respondents are ordered by their ability levels. On the other, the items are ordered by their easiness levels. The map thus displays two mirrored vertical histograms. If a respondent were located above a given item, that respondent is more likely to answer that item correctly and vice versa. The Wright map is used by Wilson and Lehrer (2021) in their discussion of learning progression. The authors use the Wright map to validate the construct map they developed. In the Wright map, they categorize the items by the ordered levels of the construct to

show that students of higher abilities are able to correctly answer items of higher difficulty or levels of the construct. This is one advantage of the Wright map over the latent space approach in that the Wright map can display the main effects (respondent ability and item easiness) in a more intuitive manner than the latent space approach.

The main limitation of the Wright map is that it only orders items and respondents along a single continuum based on their main effects. While that may be useful in certain situations, this map may not be able to differentiate items assessing many different constructs or levels of a construct (unless those different levels also differed in terms of their overall easiness parameters). For example, Wilson and Lehrer (2021) could only use a single construct ("Modeling Variability") in validating their learning progression model. In contrast, the latent space approach groups items according to the shared constructs they assess along as many dimensions as necessary. These dimensions can have additional substantive meaning for easier interpretation. For example, as shown in Jeon et al. (2021), using the DRV dataset and rotating the latent space, the two axes can represent the different types of inference. The latent space approach offers additional flexibility, especially for visualizing multidimensional items. In the context of learning progression, this means that the latent space approach allows one to understand how a learner may or may not be making progress on multiple constructs, instead of just one.

2.3 Review of Latent Space Models

Latent space models (Hoff et al., 2002) are a model-based means to graphically visualize network data. These models can reflect network transitivity and reciprocity by plotting nodes/actors on a social or latent space (a space of unobserved latent characteristics) and modeling the probability of a relational tie as a function of observed covariates and as a function of the positions of the actors on this space (such as Euclidean distance). In the case when Euclidean distance is used, the smaller the distance (or the closer the actors), the more likely a relational tie exists between the individuals. The reasoning is

that individuals who share the same characteristics (observed or unobserved) are more likely to have a relational tie. These individuals are likely to be closer in this social space. Furthermore, this model captures transitivity naturally since if ties exist between individuals i and j and between individuals j and k, then individuals i and j and j and k should be close to each other on this social space. On the social space, that also means individuals i and k should be close to each other as well and likely to have a tie, as we would expect based on transitivity.

Recently, latent space models have been used in the context of item response theory (IRT) to capture dependencies and unobserved heterogeneity among respondents and items on assessments (Jeon et al., 2021; Jin & Jeon, 2019). The idea is that assessment data is a form of network data with interactions between and among respondents and items. In the case of Jin and Jeon's doubly latent space joint model (2019), one can imagine that when two given respondents answer an item correctly, an interaction exists between those respondents, suggesting some sort of similarity between them. In the simplified latent space item response model proposed by Jeon et al. (2021) that captures the same idea, when a respondent answers an item correctly, an interaction exists between the respondent and the item.

2.3.1 Latent Space Item Response Model

The simplified model, as described in Jeon et al. (2021), can be written as

$$\text{logit}(P(y_{ji} = 1|\theta_j, \beta_i, \gamma, \boldsymbol{z}_j, \boldsymbol{w}_i)) = \theta_j + \beta_i - \gamma||\boldsymbol{z}_j - \boldsymbol{w}_i|| \tag{2.1}$$

where y_{ij} is the binary response of respondent j to item i (i.e. the i-th row and j-th column entry in the sociomatrix), θ_j is the respondent effect (i.e. respondent ability; this is denoted as α_j in Jeon et al. (2021)), β_i is the item effect (i.e. item easiness), $\left|\left|\boldsymbol{z}_j - \boldsymbol{w}_i\right|\right|$ is the Euclidean distance between the n-dimensional latent space positions of the respondent $\boldsymbol{z}_j$ and item $\boldsymbol{w}_i$, and γ is the weight or influence of the distance term.

This model states that the log-odds of respondent i responding correctly to item j depends not only on the respondent's ability θ_j and the item easiness β_i as in the Rasch model but also on some function of the respondent position $\boldsymbol{z}_j$ and item position $\boldsymbol{w}_i$. Although Jeon et al. (2021) presented a few possibilities for functions of these positions, this application uses the distance effect (specifically, the Euclidean distance can be used) for interpretability. These positions can be represented on some n-dimensional latent space or interaction map; for interpretability, these positions can be visualized on a 2-dimensional space, so $\boldsymbol{z}_j$ and $\boldsymbol{w}_i$ will be two-dimensional vectors. This distance term captures the unobserved dependence not accounted for by the respondent and item effects; γ can quantify this amount of unobserved dependence and is the "weight" of the distance term.

If certain items were closer to certain respondents (or their distances were smaller), that would suggest that those respondents were more likely to answer those items correctly even after accounting for respondent and item effects. This would be visual confirmation of an unobserved dependence structure in the data. Thus, while the positions themselves may not be interesting, on a latent space, one can see the distances between items and respondents and understand the extent to which unobserved characteristics operate in the data.

One can think of the respondent and item effects θ_j and β_i as main effects in this model and the distance term as an interaction term. The respondent and item parameters are still interesting even in the presence of unobserved dependencies; one can still understand them as respondent abilities and item easiness that can capture heterogeneity. Indeed, the respondent ability parameters from the latent space item response model correlate strongly with those estimated from the Rasch model (Jeon et al., 2021) and with total scores on assessments. However, the main effects need to be interpreted in consideration with the interaction term. Under the latent space item response model, while respondents may have similar abilities, their probabilities of providing correct responses to the same item may still differ. The new model offers more fine-grained information

in that regard, and it is this information (i.e. the distances) that can be used to create the unique student profiles to diagnose strengths and weaknesses.

This model is also simpler than the one proposed by Jin and Jeon (2019) and relaxes many of the stringent assumptions made by traditional IRT models. For example, a classic IRT model, the Rasch model (Rasch, 1961), assumes that the probability of a correct response to a binary item is only a function of the respondent ability and item easiness. Some of the assumptions are that the item responses are independent of one another given the respondent abilities and item easiness (i.e. the local independence assumption) and that respondents with the same level of ability have the same success probability. However, these assumptions are usually violated in practice; for example, local dependence may exist among items in a testlet, and unobserved heterogeneity among respondents (such as cultural background) may influence response probabilities. The model can account for and visualize these dependencies and heterogeneity (i.e. the unobserved latent characteristics Hoff et al. spoke of) without a priori knowledge by plotting items and respondents on a latent space. These unique patterns of respondent-item interactions can become apparent when one notices clusters of respondents or items in the latent space. This plot in the latent space is referred to as an interaction map in this study because the map visualizes the unique respondent-item interactions. An example of the map, created from the DRV dataset presented in Jeon et al. (2021), is shown in Figure 2.1, with smaller dots representing respondents and the larger bubbles representing items. Both are color-coded according to their values of θ and β, respectively (lighter colors represent higher-ability respondents and easier items), but the items are also sized according to the item main effects (i.e. the larger the bubble, the easier the item).

Bayesian methods, specifically the use of Monte Carlo Markov Chain methods (MCMC), are used to estimate the desired parameters in the model – a discussion is presented in Jeon et al. (2021).

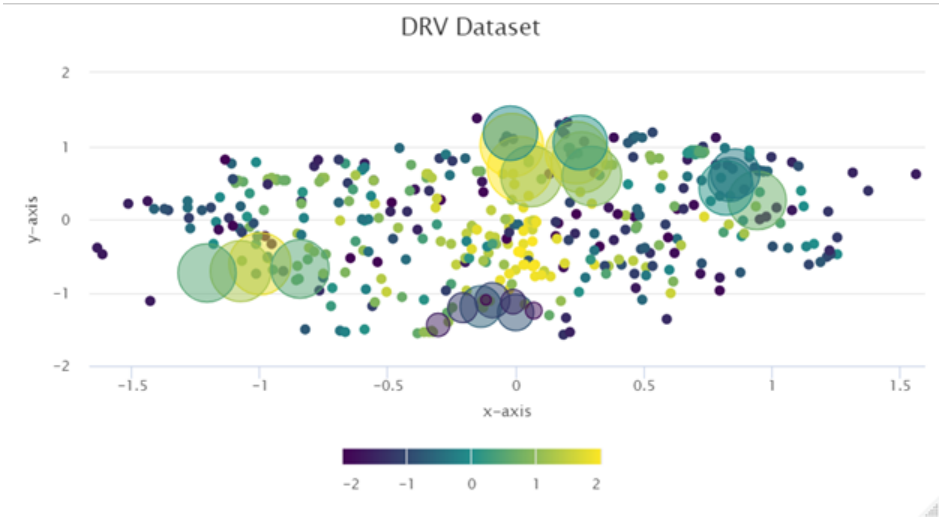


Figure 2.1: An example of an interaction map created from the DRV dataset used in Jeon et al. (2021)

CHAPTER 3

Interaction Map and Profiles

The interaction map as presented in Jeon et al. (2021) has great potential for practical use by practitioners to foster personalized learning through the generation of student profiles. The interaction map and profiles presented in this chapter can be very useful to educators for two reasons. First, these maps and profiles can be generated from data most closely related to their daily work in the classroom. In fact, teachers generally prioritize the use of such data from formative and interim classroom assessments to inform their teaching (Jennings & Jennings, 2020; Reeves et al., 2021; Wilkerson et al., 2021). Hence, educators may be more inclined to use the maps, profiles, and the web-based application which can help them make sense of data they are interested in working with. Second, the map and profiles can provide the kinds of visualizations that educators are interested in. Based on feedback from the National Science Foundation (NSF) Education Data Analytics Collaborative Workshop, Kang and Bowers (2021) reported that participants found that item analysis visualizations and non-standardized test data visualizations were most useful. The map and profiles provide unique item-level analyses that educators may find useful.

The interaction map as shown in Jeon et al. (2021) does not have these useful features. Therefore, the goal of this chapter is to improve the presentation of the interaction maps and introduce the profiles to be more instructive and informative to instructors and test developers. I first present modifications to the interaction map which allow for interactivity and the incorporation of item and respondent effects. I demonstrate clustering techniques that can help educators identify groups of students for personalized instruction. Secondly, I introduce student profiles which can be created from the maps

to provide educators more in-depth analyses of student performances. These too can also be modified depending on what educators are most interested in viewing. Finally, I show how the map and profiles may differ when the visualizations are presented in different numbers of dimensions.

I demonstrate these examples using data from the Competence Profile Test of Deductive Reasoning—Verbal (DRV) assessment (Spiel & Glück, 2008; Spiel et al., 2001). The DRV assessment measures the deductive reasoning of children in various developmental stages through items based on three design factors each with their own levels: 1. Type of Inference [Modus Ponens (MP), Modus Tollens (MT), Negation of Antecedent (NA), and Affirmation of Consequence (AC)], 2. Content of Conditional [Concrete (CO), Abstract (AB), and Counterfactual (CF)], and 3. Precedent of Antecedent [No Negation (UN) and Negation (N)]. Each item on the test involves some combination of these three design factors. This dataset, which contains the responses of 418 students to 24 binary items, was analyzed using the latent space item response model in Jin and Jeon (2019) and Jeon et al. (2021).

3.1 Interaction Maps

Figure 8a in Jeon et al. (2021) displays the interaction map for the DRV dataset. Figure 3.1 is a close reproduction of that figure, with one exception. While the respondents are represented as points in the original figure, Figure 3.1 displays respondents (in blue) and items (in red) by their identifiers. This can help educators identify on the map which particular students are struggling with certain items.

In the following sections, I present a few other options to improve the interpretability and usefulness of the interaction maps.

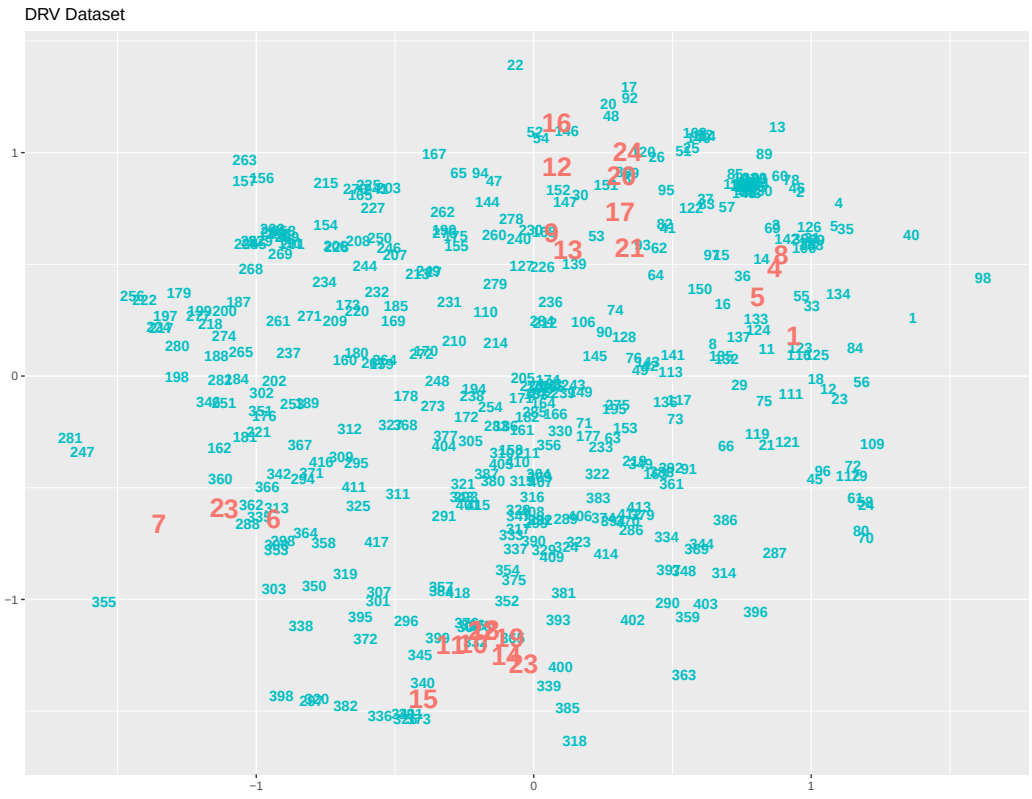


Figure 3.1: Interaction map of the DRV data

3.1.1 Removing axes labels

The relative distances between respondents and items are the most useful information to be gleaned from the interaction map. The axes labels are extraneous information; while they can be used to help users estimate the actual distances between respondents and items, that information can be presented in the profiles as described in the following section. In fact, the axes labels can be confusing (K. Givvin & J. Son, personal communication, December 17, 2021). Therefore, one small improvement is removing the axis labels as shown in Figure 3.2

There are two reasons why the axes labels may be retained. First, the scale of the axes are dependent on the estimate of γ, which can quantify the amount of unobserved respondent-item interactions (and hence the degree to which the latent space item response model is needed for the data). Retaining the axes labels may allow users

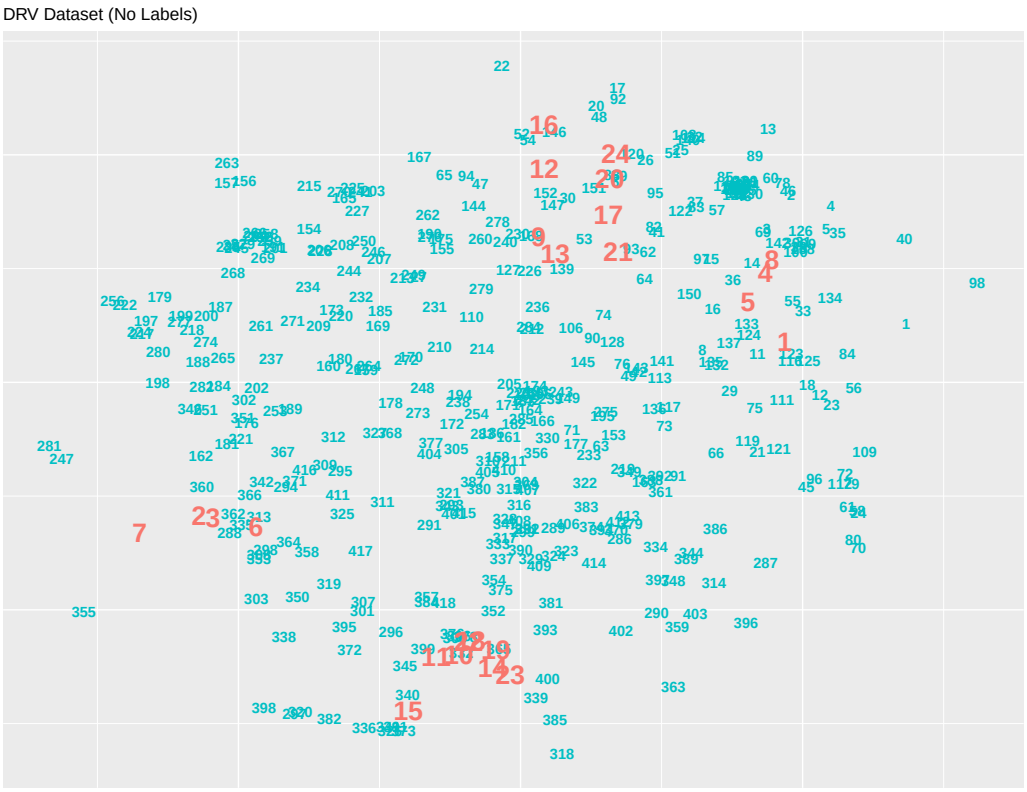


Figure 3.2: Interaction map of the DRV data with no axes labels

to understand the amount of dependencies in the data. Regardless, users can look at the estimate of γ itself for this information without needing to look at the axes labels. Second, Jeon et al. (2021) describe how substantive interpretations can be attached to the latent space dimensions and axes using a rotation. They demonstrate this by rotating the latent space of the DRV data and inferring that the two axes differentiate the items based on the "content of the conditionals" design factor. However, this may only be useful when there is an idea of the test design factors. In this situation, it is known that the DRV has three design factors and that the axes could represent them. When these design factors are not known a priori, it may not be possible to attach substantive meaning to the axes even after rotation. Given these complications, the axes labels can probably be removed for the sake of interpretability.

3.1.2 Incorporating item and respondent effects

While the interaction map can be useful to visualize students and items, the map currently presented in Jeon et al. (2021) does not include information regarding the respondent and item parameter estimates (θ_j and β_i respectively). Therefore, one way to improve the presentation is to include information about the item and respondent effects in the visualization.

A bubble chart is one such option. Sedrakyan et al. (2019) note that bubble charts "can be effective for the purpose of having a quick glance and getting a high-level overview" (p. 31) of group or class-level performance. Specifically, in the bubble chart, each item can be sized according to the estimated item easiness parameters. Thus the educator can see which items, on average, were more difficult than others.

Figure 3.3 is an example of such a visualization, created by the highcharter package in R (Kunst, 2022). Each respondent is represented by a point, and each item is represented by a bubble. The respondents and items are colored by their ability or easiness parameter estimates. For example, darker points represent respondents of lower overall ability while dark bubbles represent items with lower easiness parameters (or more difficult items). Furthermore, the item bubbles are also sized according to the item easiness parameters, so larger bubbles signify easier items and vice versa. This visualization provides information beyond what is provided in the original interaction map presented by Jeon et al. (2021). While the distances between the respondents and items are the novel contribution from the latent space item response model, it may still interest users to see the traditional item and respondent effects in the interaction map. While the interaction map answers more specific questions such as "Which students have trouble with certain items?", the bubble chart also answers more general questions such as "In general, what items pose the most difficulty for students?" and "Which students in general need more support?"

For example, Jeon et al. (2021) are able to distinguish four item groups based on different combinations of levels of two design factors. However, educators and test de-

signers may wish to know which particular item groups, overall, were most difficult for students. This option could provide that information. The darker colored bubbles in the bottom of Figure 3.3 represent items with lower easiness parameters. Those are items 10, 11, 14, 15, 18, 19, 22, and 23. According to Jeon et al. (2021), those items correspond to item group 2, which contains more advanced items that involve the abstract and counterfactual levels of the content of conditional design factor. From this interaction map, in addition to the item and respondent groupings, the educator would further understand that those particular items are more difficult for the students in general and could tailor instruction to help students handle those kinds of items. The educator might also notice the concentration of yellow dots in the center, representing students of high overall ability who are equally likely to answer items from the four item groups correctly; the educator may not need to focus as much on those students. A test designer can also use this to validate their test design (e.g., the items assessing higher levels of deductive reasoning are more difficult, as they should be). In conclusion, this visualization provides additional item-level analyses that Kang and Bowers (2021) claim teachers would find useful.

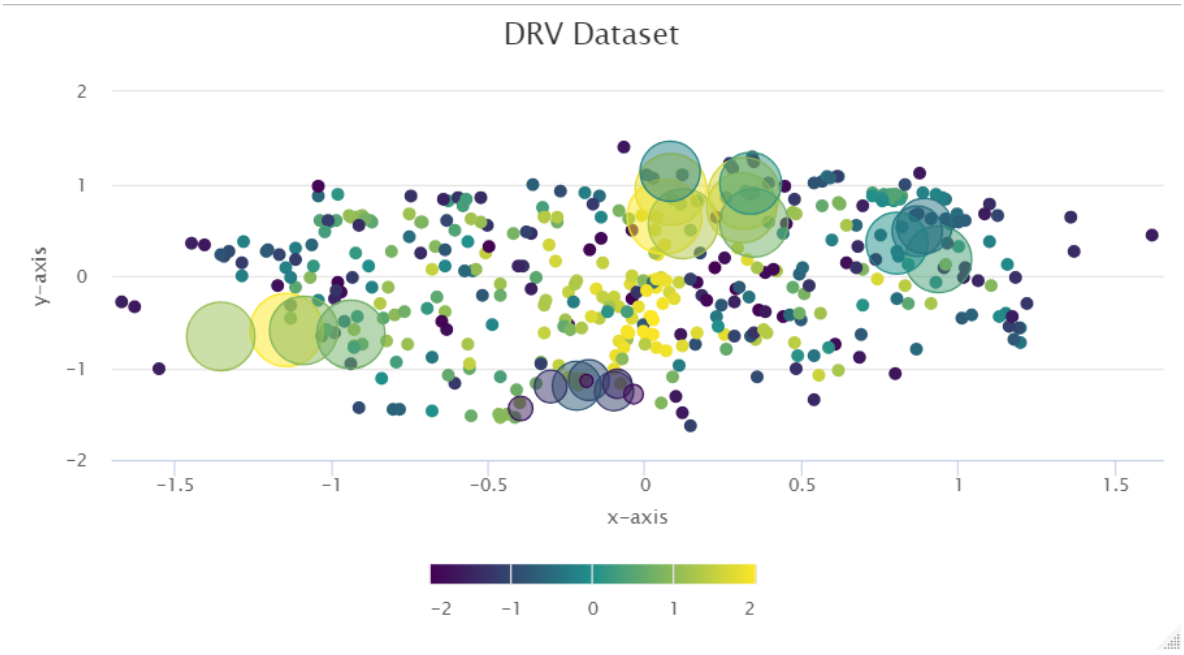


Figure 3.3: Interaction map of the DRV data with respondent and item effects

More importantly, as noted later, respondents of very high abilities and very low abilities tend to be equidistant from the item groups. That is because those students are equally likely (or unlikely) to answer all items correctly. Therefore, in the latent space for the DRV data, students who answer all items correctly and those who answer them all incorrectly tend to be clustered in the center of the latent space. Showing the main effects of the respondents, as denoted by their color, can help users distinguish between the two groups of students.

3.1.3 Interactivity

Interactivity can improve presentations of data by allowing users to better understand more specific parts of visualizations. Interactivity is one metric that can define the success of visualizations (Halim & Muhammad, 2017).

The highcharter package allows for interactivity of the maps. Hovering over the respondents or items can create a tooltip that displays important information such as the respondent/item number and the exact respondent or item main effects, as can be seen in Figure 3.4. That way, it is possible to visualize the main effects of the respondents and items while retaining the identifiers. Additionally, this interactivity can make viewing information easier. For example, as can been seen from Figure 3.2, item group 2 is clustered very closely together, making it very difficult to see which item numbers belong to that cluster. By hovering over the bubbles, users can read the tooltips for each item in that cluster, thus allowing them to see the items that are within that cluster. The same can be done for respondents, since there are clusters of respondents in Figure 3.2, and it may be difficult to distinguish the identifiers of the students in those clusters.

This interactivity would require a dynamic visualization rather than a static image. Therefore, a web-based application would be necessary to incorporate this kind of interactivity into the interaction map. I present an example of such an application in a later chapter. An example of such a visualization is shown in Figure 3.5. This plot combines aspects from the original latent space plot and the interactivity demonstrated

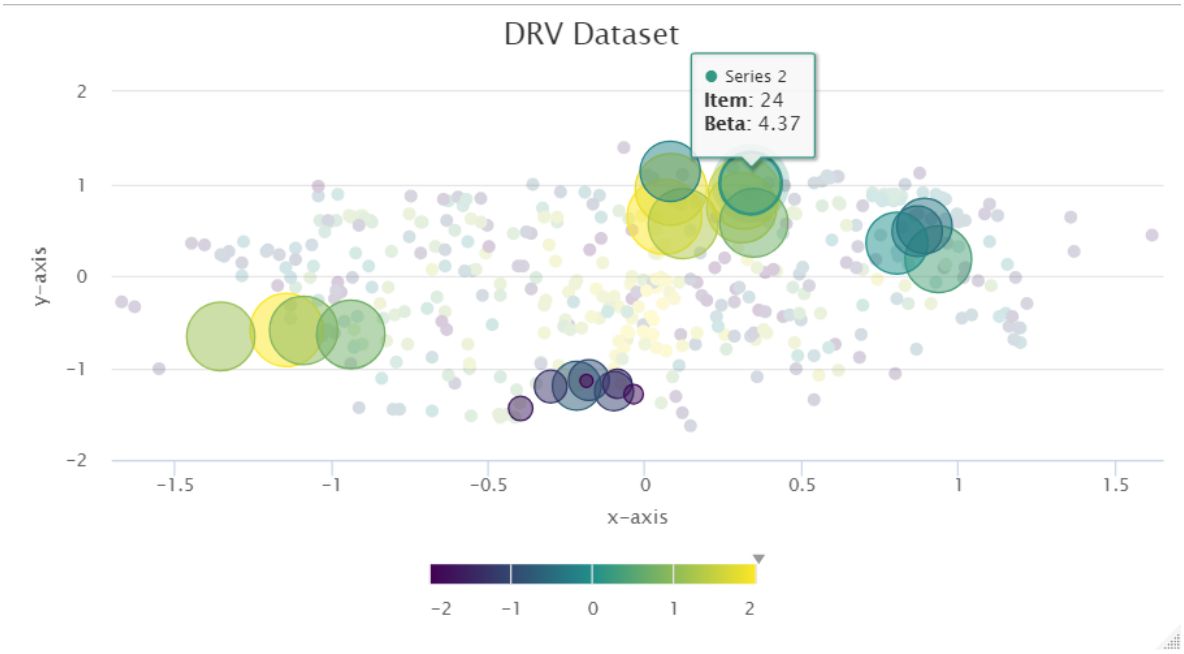


Figure 3.4: Interaction map of the DRV data with tooltip

in this section. That way, a user can see the main effects of the respondents or items while also quickly identifying the specific items or respondents in the latent space.

3.1.4 Clustering

While the item clusters may be obvious on the interaction maps, the respondent clusters are not. Educators may be interested in seeing distinct clusters of respondents based on the distances so they can categorize their students and deliver targeted interventions to those groups. In Jin and Jeon (2019), spectral clustering was used to group respondents. In this section, I present some other options.

3.1.4.1 Hierarchical Cluster Analysis Heatmaps

Hierarchical cluster analysis and heatmaps can be a useful way to not only provide another way to visualize these distances but also to cluster both items and respondents based on shared patterns. Hierarchical cluster analysis (HCA) is a statistical

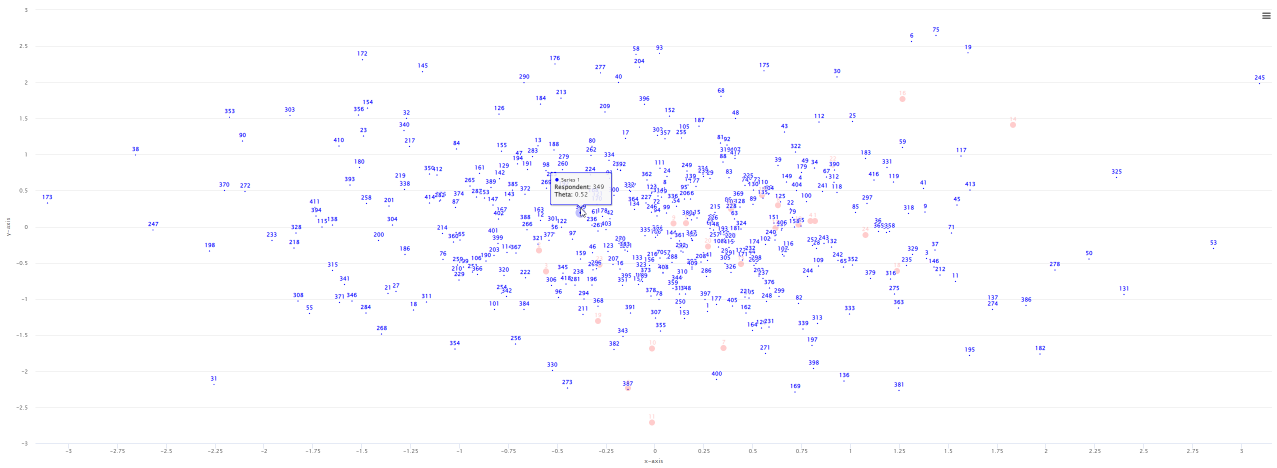


Figure 3.5: Interaction map of the DRV data with respondent and item effects in the tooltip

method that uses distance measures to reorder multivariate data such that similar data patterns are clustered. HCA first uses a distance metric (such as Euclidean distance) to determine which data points are closest to one another. Then, an agglomeration method is used to cluster the data points using this distance metric. The resulting clusters and patterns can be visualized using heatmaps using the ComplexHeatmap package (Gu et al., 2016). An example is presented in Bowers (2010) using longitudinal data from the grading histories of K-12 students. Students can be clustered according to similarities in the patterns of grades they received throughout their educational journeys.

Figure 3.6 is the heatmap created from the DRV dataset. Each row represents an item (denoted by the number on the right) and each column represents a given respondent (denoted by very small numbers on the bottom). Each cell is color-coded according to the distance from a respondent to an item, with warmer, redder colors denoting larger distances (i.e. weakness on the item) and darker colors for smaller distances (i.e. strength). Additionally, the dendrograms or cluster trees on the left and the top of the heatmap visualize the similarities between respondents and items. Longer lines indicate greater dissimilarity whereas smaller, closer lines denote respondents or items that are more similar to one another. Following the advice provided in Bowers (2010) for educa-

tional data, the distance metric is the uncentered correlation between observations and the agglomeration method is average linkage used to create the dendrograms.

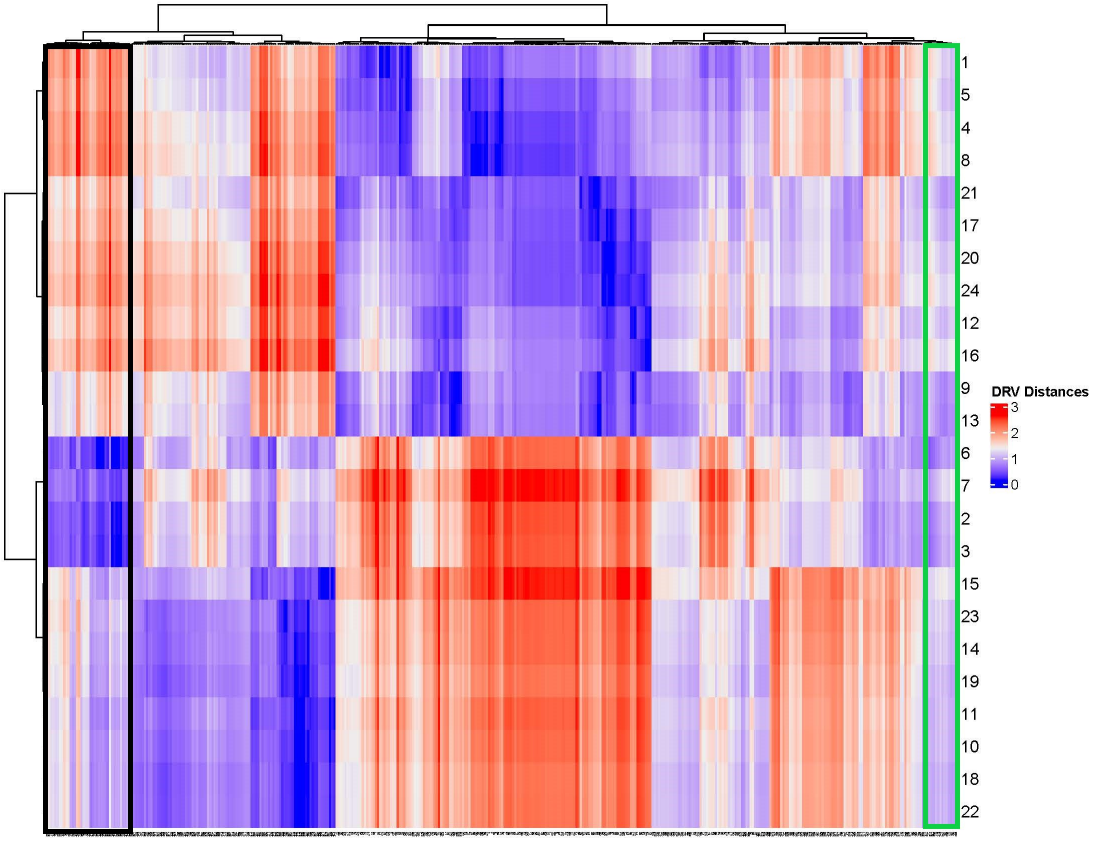


Figure 3.6: Heatmap of the DRV data

Figure 3.6 validates the four item groups found in the original interaction map. The dendrogram (on the left) shows the similarities among the items based on the distance measures. The rows are also ordered so that more similar items (such as 6, 7, 2, and 3) are next to one another. These dendrograms can visualize the similarities of these items and item groups.

More importantly, the respondents can be clustered too. The dendrograms on the top of the heatmap visualize the similarities among the respondents. Like the items, the respondents are ordered so that respondents with similar patterns of distances to items are adjacent to one another. A user can look at the ordering of the respondents or the dendrograms to cluster these respondents. For example, the black rectangle in Figure 3.6

denotes the respondents who are very close to items 2, 3, 6, and 7 on the interaction map. The cells corresponding to those respondents and those items are deep blue to reflect that fact. Those same respondents are also farther away from other items, as shown by the lighter cells. On the other hand, the green rectangle denotes respondents who are in the center of the interaction map and thus are equidistant from the four item groups. The cells have roughly the same colors across the items. This heatmap allows users to identify clusters of respondents based on their distances either through the ordering of the respondents, the dendrograms, or the blocks of colors on the heatmap itself.

Another advantage of the heatmap is that annotations can be included on the heatmap to denote categorical variables. For example, Bowers (2010) annotated heatmaps to include information about student gender and dropout status. In this case, a user could include annotations to denote students who receive free or reduced price lunch and see whether there are systematic patterns in the heatmap according to those annotations.

It may also be useful to cluster not just on the distances but also on the overall ability parameter, or θ. Figure 3.7 shows the heatmap with the addition of the θ term. The heatmap looks similar to Figure 3.6 but includes some additional information about respondent clusters. In the interaction map, respondents 243 and 407 are in the center, roughly equidistant to the item clusters. However, in the heatmap, respondent 243 is in the cluster denoted by the green rectangle. These respondents have very low θ (as seen from the deep blue cells in the "theta" column). Respondent 407, on the other hand, is in the cluster denoted by the black rectangle which contains respondents with the highest θ in the sample. Despite these differences in their overall ability, the patterns of their distances are similar since they are in the same location in the interaction map. This heatmap helps us distinguish these two different groups of respondents based on their θ who would otherwise be indistinguishable on the interaction map. Substantively, while these students are equally likely to answer all items correctly, one group represents students who essentially answered all items incorrectly (i.e. equally likely to answer all

items incorrectly) while the other represents students who answered them all correctly (i.e. equally likely to answer all items correctly).

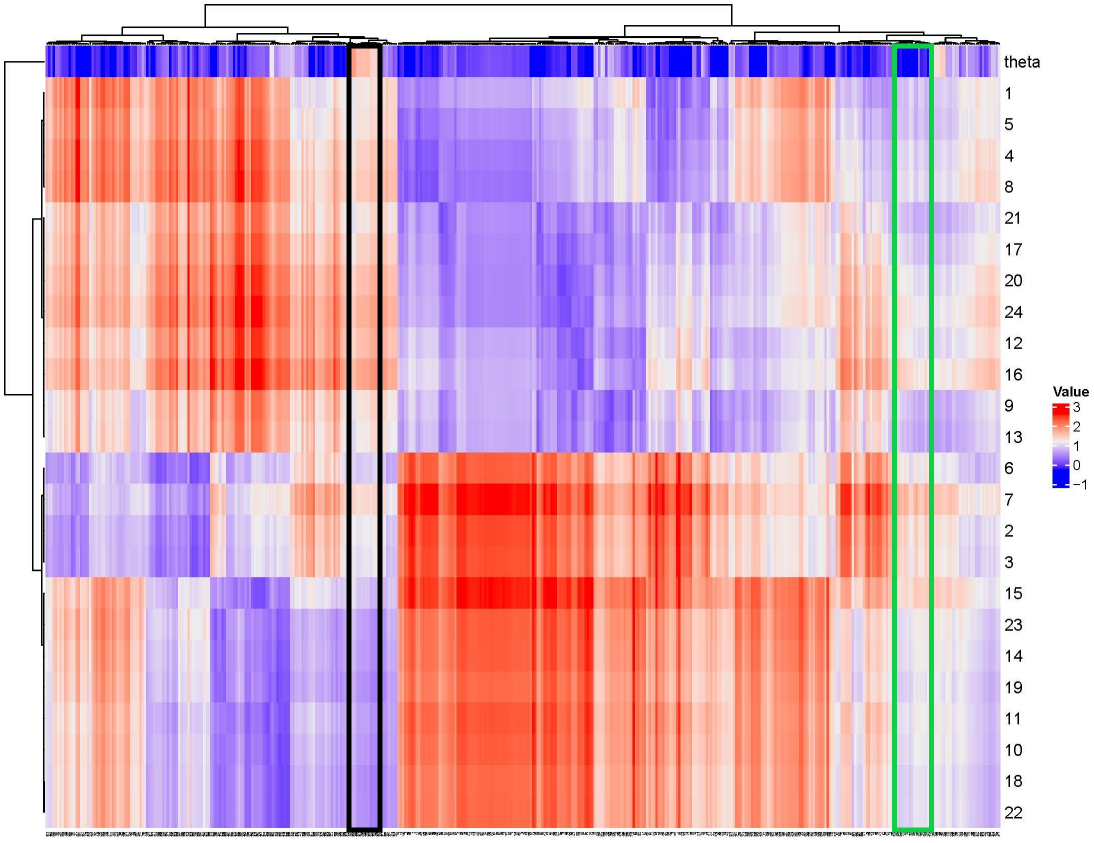


Figure 3.7: Heatmap of the DRV data with θ

Since the four item groups are evident from the heatmaps, the items can be grouped together. Following Jeon et al. (2021), the item groups can be denoted by I1 (items 2, 3, 6, and 7), I2 (items 10, 11, 14, 15, 18, 19, 22, and 23), I3 (items 9, 12, 13, 16, 17, 20, 21, and 24), and I4 (items 1, 4, 5, and 8). First, the centroids of those four groups can be calculated by taking the average of the coordinates of the items that comprise each group. Then, the distances among the respondents and the centroids can be calculated and used to create the heatmap shown in Figure 3.8.

The clustering of the items and the respondents can be made clearer by looking at the dendrograms themselves. Recall that the dendrograms can show the items and respondents as ”leaves” on a tree. The height of the branches (as denoted by the y-axis)

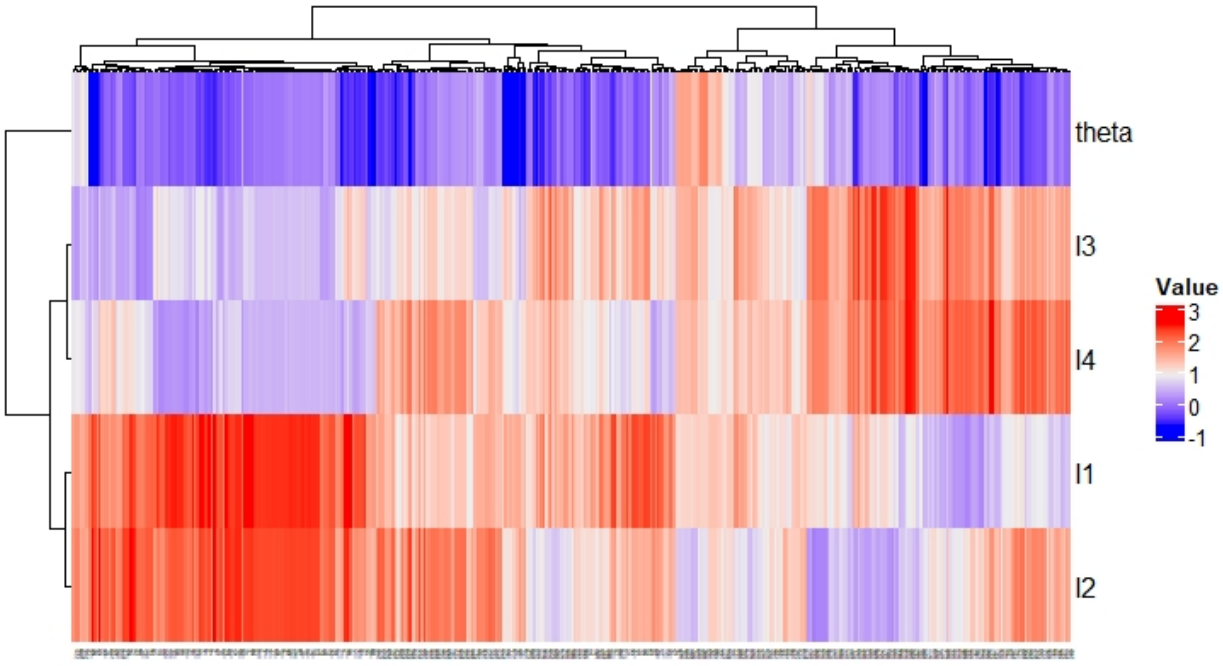
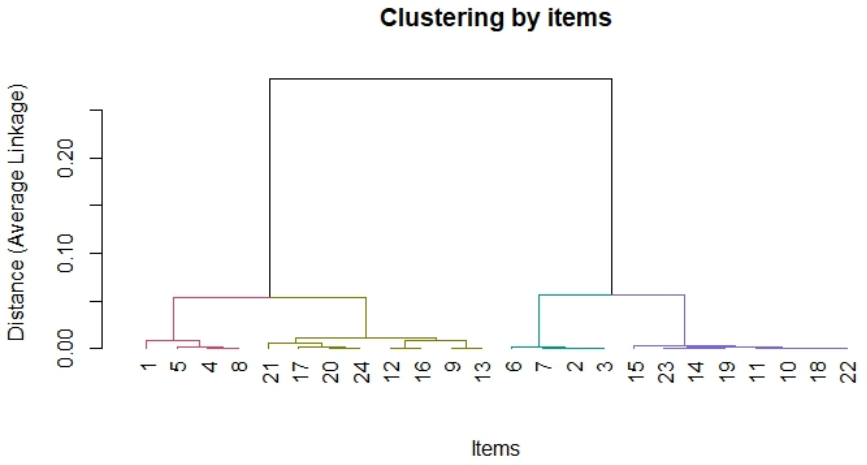


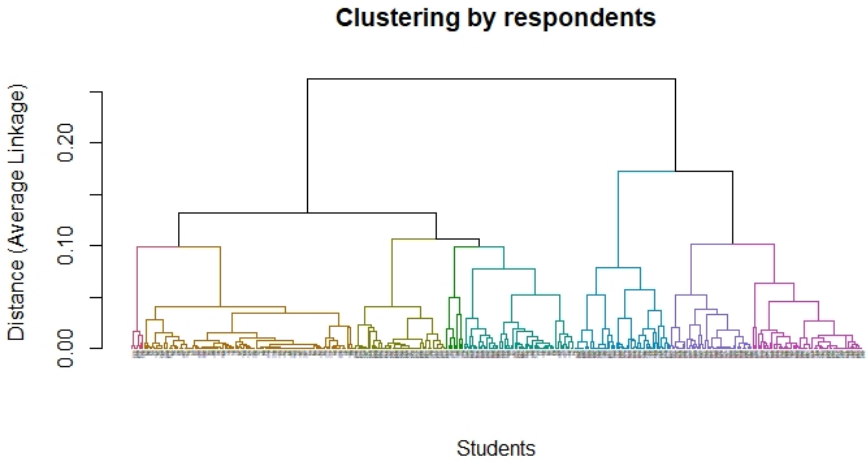
Figure 3.8: Heatmap of the DRV data with θ and item groups

denotes the degree of dissimilarity among the items or respondents. Therefore, smaller trees (with much shorter vertical distances or branches) mark respondents or items that are more similar to one another.

Figure 3.9 shows the dendrograms created for items and for respondents. To help identify the clusters, the trees of the dendrogram can be colored using k-means clustering. For example, the four item groups can be colored as seen in Figure 3.9a. Item group I1 is teal, item group I2 is purple, item group I3 is yellow, and item group I4 is red. The respondents can also be clustered based on their item response patterns and θ as well in Figure 3.9b. For example, blue denotes the respondents who have high θ and are relatively equidistant to the four item groups. While the respondents in the red and brown groups are likely to answer items from item group I3 correctly, the former have higher θ than do the latter.



(a) Clustering by items



(b) Clustering by respondents

Figure 3.9: Dendrograms and clustering items and respondents by k-means

3.1.4.2 Latent Profile Analysis

One strength of hierarchical clustering is that the degrees of similarities can be visualized with the heatmaps or dendrograms. However, cluster-based methods can be subjective since there are a variety of cluster solutions (Pastor et al., 2007). For example, different numbers of clusters can be used to group the respondents in Figure 3.9b.

Therefore, another solution is to use a model-based approach called latent profile analysis (LPA). LPA can group respondents into latent profiles based on similarities in their patterns of observed, continuous indicators. Different models can be specified based on the variance-covariance matrix of indicator variables. For example, one model specification may specify that variances of the variables are identical and the covariances are zero across profiles. Another model specification may allow the variances or the covariances to vary across profiles. It is possible to use fit indices such as AIC or BIC to compare different model specifications and understand which number of profiles best fits the data.

In this context, the idea is to cluster respondents based on their distances to the items and their ability levels θ. Different numbers of profiles and model specifications can be tested. These model specifications may specify whether we believe that the variances or covariances of the distances to items vary across profiles. The tidyLPA package in R can be used to conduct the LPA (Rosenberg et al., 2018).

The interaction maps and hierarchical clustering suggest the existence of four main item groups. Therefore, to make the LPA more tractable, the centroids of the items belonging to each of the four clusters will be calculated (as was done in the HCA). There will be five indicator variables total: the distance of each respondent to each of the four item group centroids and the respondent's θ. Based on the AIC and BIC fit indices, the model with five profiles and with varying means, equal variances, and varying covariances fit the DRV data the best.

Figure 3.12 shows the profiles from the best-fitting model. Each class is repre-

sented as a colored box (for five colors total). Each box attempts to capture most of the observations belonging to a given class. The centroid, represented as a dot in the middle, shows the center of these observations. The y-axis represents the standardized value for each of the variables in the x-axis, so the values would represent distances (the distances can be negative in this plot because all variables were scaled) from the respondents to the centroids of item clusters I1, I2, I3, and I4. The y-axis values would represent overall ability level for the theta variable. Finally, lines connect the centroids to show how each class differs based on the five variables. The characteristics of each of the five classes can be summarized as follows:

- Class 1: Students close to I1 and I3, far from I2 and I4
- Class 2: Students close to I3 and I4, far from I1 and I2
- Class 3: Students close to I1 and I2, far from I3 and I4
- Class 4: Students roughly equidistant to the four item groups
- Class 5: Students close to I2 and I4, far from I1 and I3

A barplot, as seen in Figure 3.10, can visualize the differences between the classes. The x-axis denotes the class, and the filled bars represent each variable. The height of each bar represents the average value of the variable for a given class. One can see that for example, Class 4 has the highest average value of θ.

Note that the students in Class 4 tend to have higher values of θ compared to those in the other classes, corroborating our prior observations that the higher ability students tend to be equidistant from the items. This LPA can identify specific students within each class. This would allow instructors to better identify students in specific clusters and therefore understand how to best address their needs. For example, the LPA would specify which specific students belong to Class 2. Since I1 and I2 items have the ”Negation of Antecedent” (NA) and ”Affirmation of Consequent” (AC) levels of the ”Type of inference” design factor while I3 and I4 items have the ”Modus Ponens” (MP)

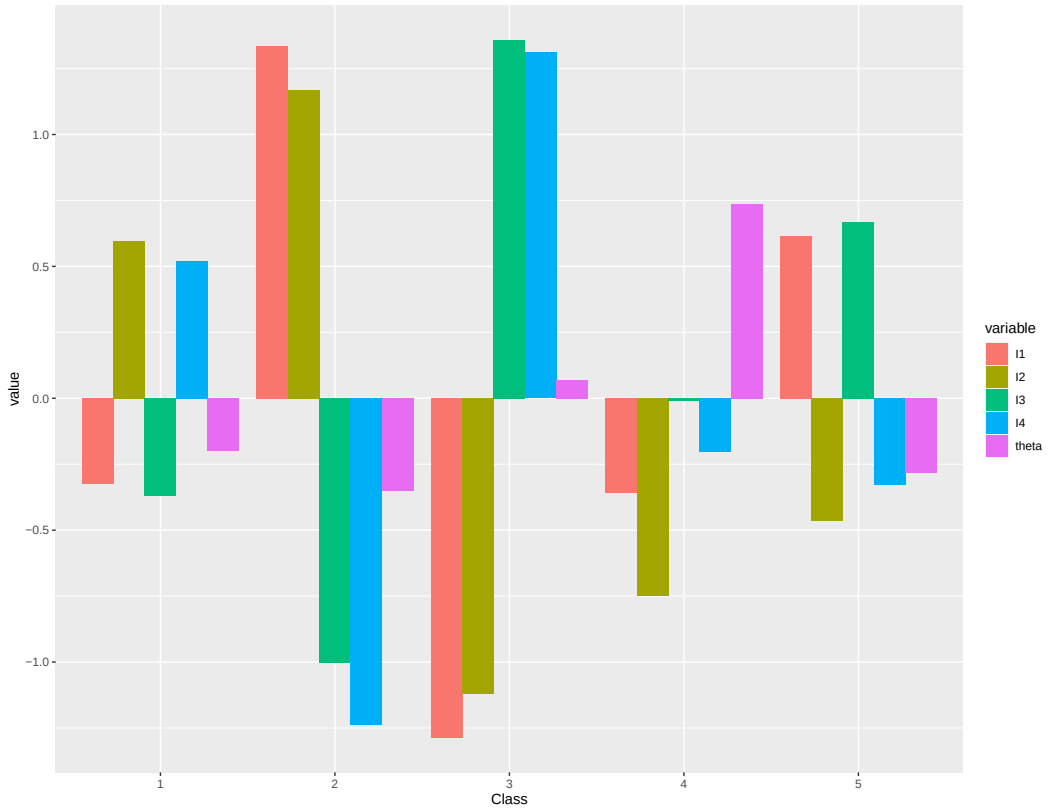


Figure 3.10: Barplot of the characteristics of each profiles/class from the best-fitting model

and "Modus Tollens" (MT) levels, the instructor would understand that those students in Class 1 would need more help or tutoring with understanding items that have the more complex type of inference levels of NA and AC. While it is harder to visualize which students are more similar to others in terms of their response patterns (as can be seen in the HCA), the LPA approach offers greater clarity regarding cluster assignments. In hindsight, this respondent grouping makes sense, since each pair of item groups shares similar design factors.

For an easier visualization, the barplot in 3.11 shows the differences among the classes. Each class is represented as an individual bar plot (a facet). The y-axis also reflects the raw distances and values, since negative distances shown in Figure 3.10 may be confusing.

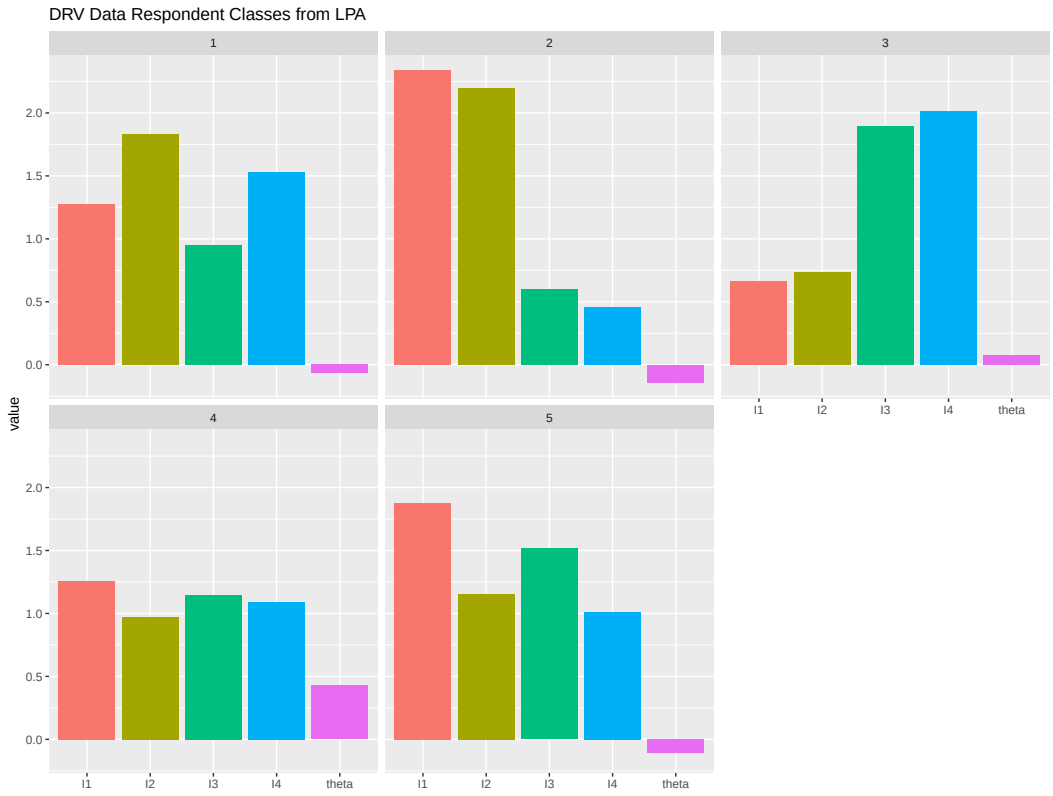


Figure 3.11: Barplot of the characteristics of each profiles/class from the best-fitting model, separated by facets

For example, respondent 407 belongs to Class 4. That means respondent 407 is roughly equidistant to the four item groups and has very high ability. In fact, respondent 407 ranks near the top percentile in terms of θ. On the other hand, respondent 122 is in Class 2, which means that they are close to I3 and I4 (i.e. stronger on MT/MP items) and farther from I1 and I2 (i.e. weaker on NA/AC items).

While this model-based approach has less subjectivity, the trade-off is that the computation time can be considerable. Trying to compare models using the person-item distances instead of the person-item group distances is very computationally expensive and time-consuming due to the large number of parameters in the variance-covariance matrices that need to be estimated. Because of the large number of parameters to be estimated, more complex models may require larger samples. Therefore, clustering us-

ing LPA may not be computationally feasible in certain situations, such as when the sample size is small or when very complex models need to be fit. Additionally, this model-based method is not perfect and may give classifications for certain respondents that may not make that much sense. For example, respondent 22 is in Class 1, which suggests that they are close to I1 and I3. However, they are actually close to I1 and I2. Model-based methods can provide useful summaries, but summaries do not necessarily account properly for all cases.

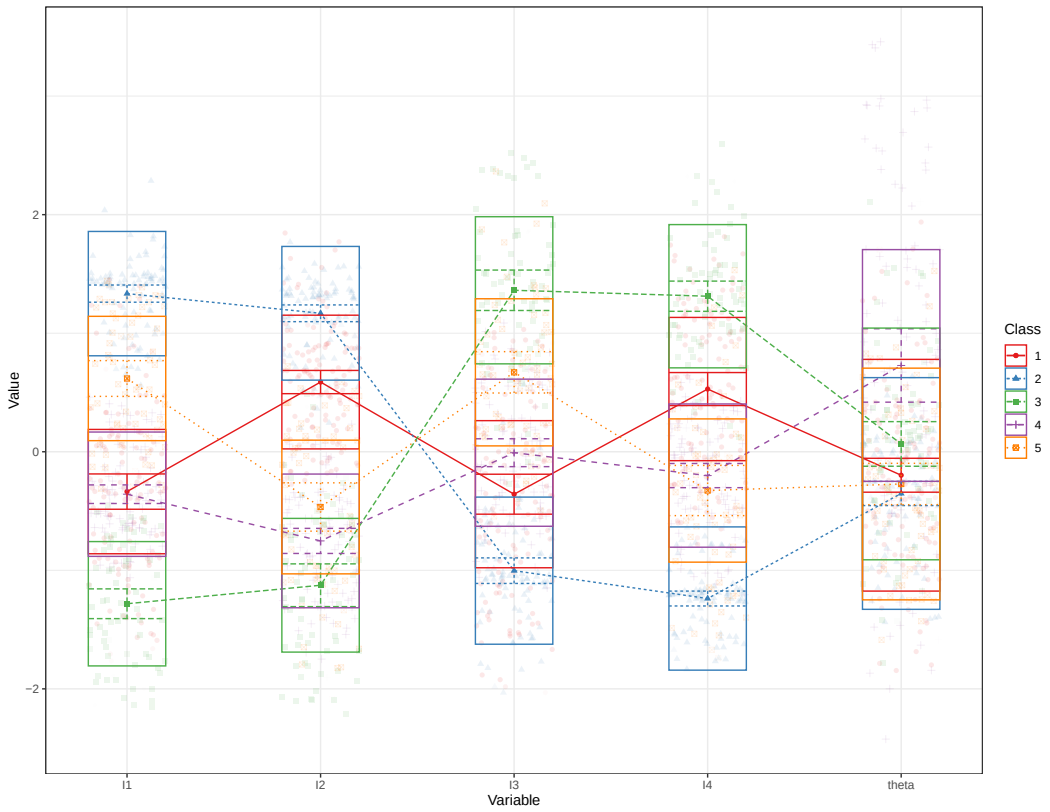


Figure 3.12: Characteristics of each profiles/class from the best-fitting model

3.2 Profiles

It can be difficult to make individual inferences regarding certain students if there are many students and items that clutter the interaction map. Furthermore, educators may also be interested in understanding individual student performance asides

from understanding general trends among all students. Thus in this section, I demonstrate individual profiles that can help visualize the performances of individual students. These profiles can be created from the distances between an individual student's position on the interaction map and the positions of all the items. As discussed previously, these distances can represent specific respondent-item interactions – the greater the distance, the less likely the student is able to answer the item correctly and vice versa. Hence, one can interpret these distances as measures of student weaknesses or strengths (depending on the model formulation, as discussed in the previous section) that go beyond the conventional measures of respondent abilities provided by standard IRT models.

These profiles can be visualized as bar charts. Sedrakyan et al. (2019) argue that "Bar-charts are visualization techniques (e.g., Fig. 2) to show a relationship between a part and a whole or compare categories, thus allowing to compare the planned and actual achievements during a learning process (e.g., online tests to measure learning goal outcomes)" (p. 28). The bar charts are appropriate because in the profiles, we compare the students' performance on various items and also what we would expect to see if a given student were not more likely to answer certain items correctly. Educators are also likely to be familiar with bar charts since in a review of education dashboards across K-12 and higher education for both teachers and learners, the data visualization most often used was a bar chart (Schwendimann et al., 2017). These profiles can also show respondent and item effects. In the following profiles, the respondent ability parameter estimates (along with their percentile rank in the sample) are shown in the title of the profile while the individual bars, representing items, are colored by the item easiness parameters with lighter colors denoting easier items.

3.2.1 Choices of axes

In the profile for each student, the x-axis can represent items that the given student answered. The y-axis can be some function of the distance from the item to the student. In the following sections, I present different options and their advantages and

disadvantages.

3.2.1.1 Raw distances

One option is the identity function or plotting just the distances on the y-axis. While this is the simplest function, it is difficult to compare different students using this method. For instance, a student who is equidistant from all items might have large distances to all items compared to a student who is very close to one item but very far from the others. Based on the y-axis values, one might say that the former student is very weak on all the items compared to the latter student, although it may be more informative to say that the former student is equally likely to answer all items correct, rather than that the student is "weak" on all items.

For example, the profiles of respondents 22 and 407 are shown in Figure 3.13. Respondent 22 is an example of a respondent close to one item group (specifically, items 9, 12, 13, 16, 17, 20, 21, 24). On the other hand, respondent 407 is in the center of the interaction map. Based on the profiles, one may claim that respondent 22 is stronger than respondent 407 on the aforementioned items. The more accurate interpretation is that respondent 22 is better able to answer those items correctly relative to the other items. Respondent 407 has no such propensity and would seem to be equally likely to answer all items correctly. In terms of comparison between the two students, it may not be strictly accurate to say that respondent 407 is weaker than respondent 22 on those items. Rather, one can see that respondent 407 has higher ability (based on the theta value) than respondent 22, so even if the distances from respondent 407 to those items are greater, respondent 407 should be able to answer those items correctly due to the greater value of theta. The mismatch between the maximum of the y-axis makes this interpretation difficult; since these are raw distances, the maximum distance can vary across respondents.

In general, working with the raw distances can be misleading. For instance, a respondent can have a very large distance to an item can still have a higher correct

response probability to an item than a respondent with a small distance. This may be possible because the former respondent has a higher estimate of the ability parameter which makes up for the larger distance term. Comparing the raw distances may only make sense when the comparison is between respondents of similar ability estimates.

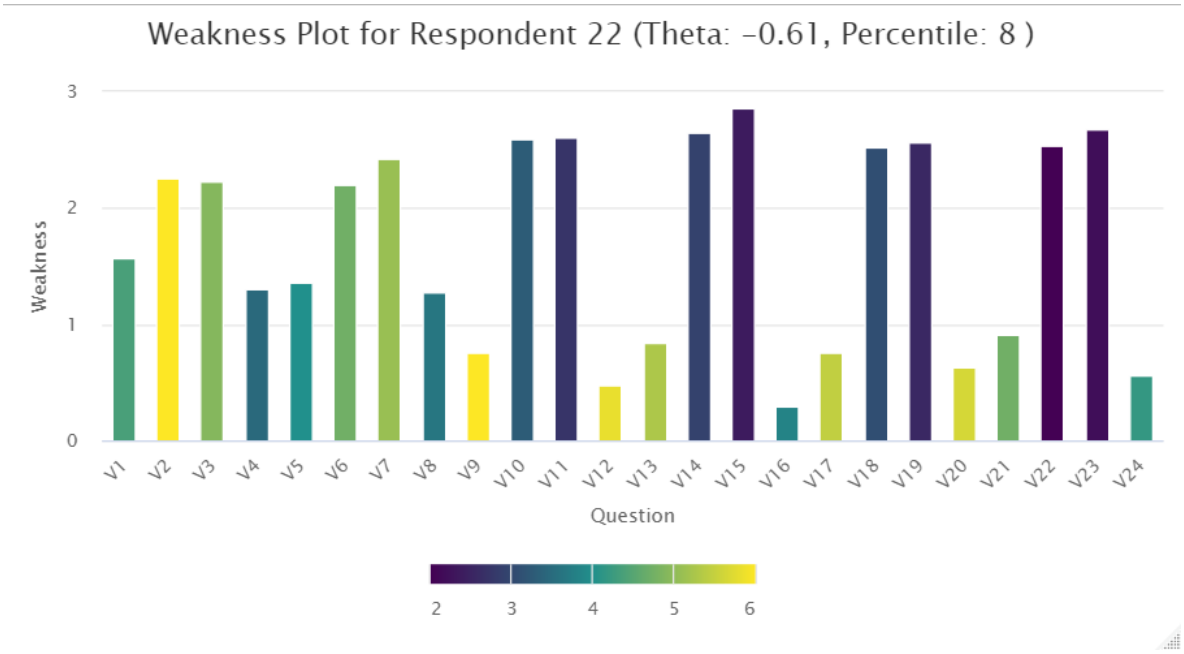
Another choice of the y-axis or function of the distance term could improve the interpretability of the profiles.

3.2.1.2 Proportion of total distance

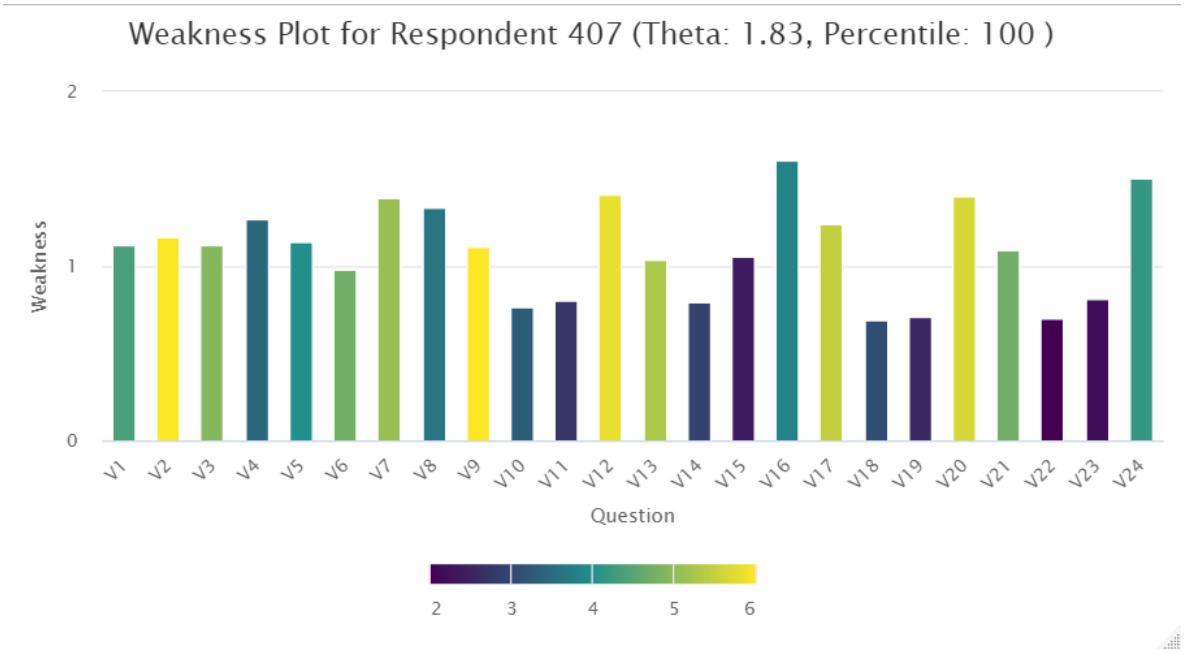
Another option is to let the y-axis reflect the proportion of the total distance to all items that is comprised of the distance to a given item. For a given respondent, the sum of the distances to all items is calculated. Then, for each item, the distance to that item is divided by the total distance. The quotient is the height of the bar for that item in the profile.

For a student who is equally likely to answer all items correctly, the student should be roughly equidistant to all items. Therefore, the height of the bars should be roughly equal to one divided by the number of items. In the profile, a line can be drawn to denote this height so that a user would know that it is the expected height of the bars for a student who is equally likely to answer all items correctly. For students who are closer to certain items, or are more likely to answer those items correctly, the heights of those bars should be closer to zero. For students farther from certain items, the heights of those bars should be closer to (but not exceed) one.

The advantage of this function of the distance term is that it can facilitate comparisons between different respondents since the scale of the y-axis is the same for every respondent (from zero to one). One can therefore say that a given student has more trouble with a given item relative to all the other items they answered because a greater proportion of the total distance is attributed to a given item. These kinds of comparisons can be made even if the respondent abilities are not the same.



(a) Profile of respondent 22



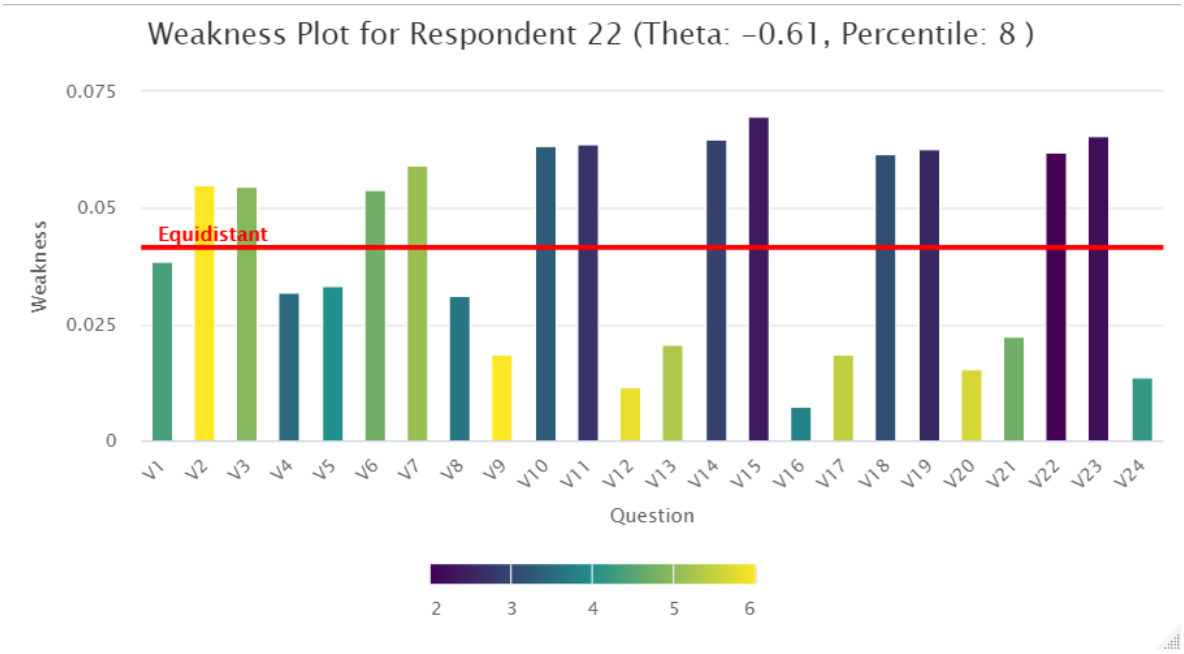
(b) Profile of respondent 407

Figure 3.13: Profile of two respondents with raw distances in the y-axis

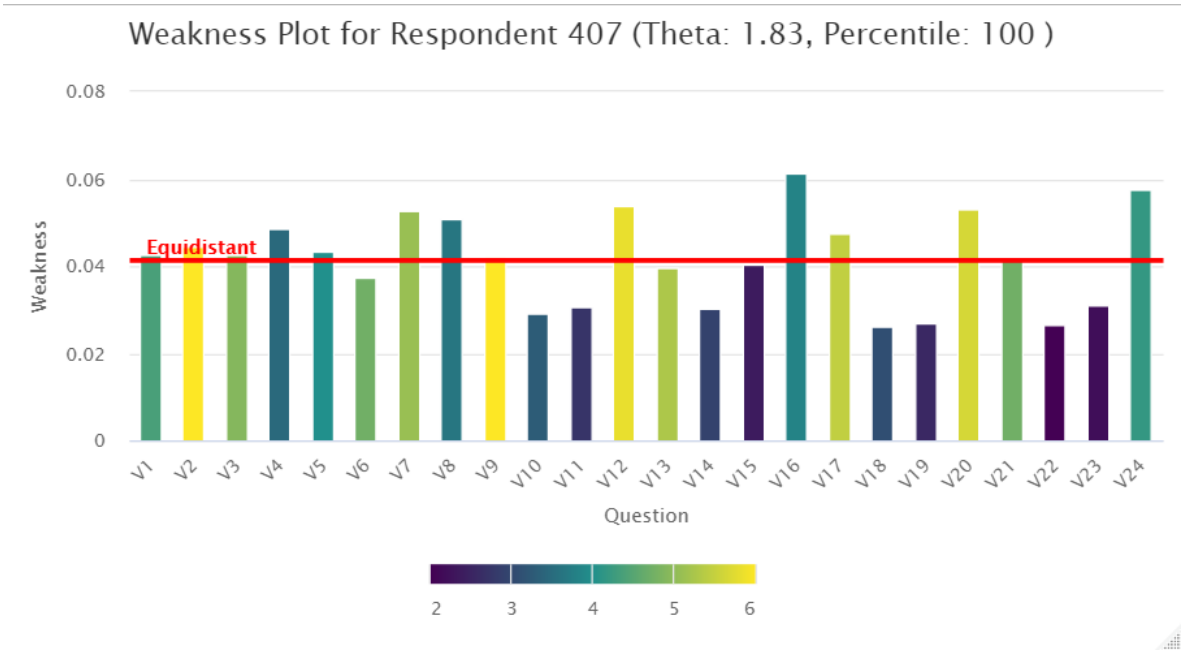
For example, the profiles of respondents 22 and 407 are shown in Figure 3.14. Note that the scales of the y-axes are similar and cannot exceed one for any respondent. Respondent 22 has higher bars for items such as 10 and 11, suggesting that a greater proportion of the total distance is attributed to those items. That means respondent 22 is farther from those items and would need more support on them. In contrast, the heights of the bars for respondent 407 are more uniform and closer to the equidistant line, suggesting that this respondent is equally likely to answer all items correctly.

One disadvantage is that it is possible for two respondents to have the same height of the bar even if their distances to a given item are different. For example, respondents A and B may have distances of 2 and 3 to an item respectively but total distances of 6 and 9. In that situation, even though the raw distances to the items are different, the heights of the bars for those respondents to those items would be the same ($\frac{2}{6} = \frac{3}{9} = \frac{1}{3}$). This particular choice of the y-axis would obscure the differences in raw distances to a given item across different respondents. Another problem is that two respondents with the same distance to an item may actually have different heights of the bar. For instance, two respondents may have distances of 3 to an item, but respective total distances of 5 and 9. The heights of the bars would not be the same. In this case, the heights of the bars would obscure the fact that these respondents actually have the same distances to that item.

It may not be that problematic for the height of the bars to be different when the raw distances are the same or for the height of the bars to be the same when the raw distances are different. After all, the main advantage of this function of the distance term is that it allows users to understand, for a given respondent, which items pose the greatest difficulties based on the proportion of the total distance. For example, respondents may have different raw distances to an item but have the same heights of the bars; that may be okay since that item posed the same relative level of difficulty for those respondents. For example, assuming that respondents A and B have equivalent abilities, it is true that respondent A may have less trouble (in absolute terms) with the



(a) Profile of respondent 22



(b) Profile of respondent 407

Figure 3.14: Profile of two respondents with proportion of the total distance in the y-axis

item compared to respondent B because her distance term (2) acts as a smaller penalty to the correct response probability. Regardless, both respondent A and respondent B, in relative terms, find that item equally difficult in comparison to all the other items they have answered. The educator might group these respondents together, in contrast to other respondents who found that item disproportionately more difficult relative to the other items they have answered.

3.2.1.3 Logistic function

The main disadvantage of using the proportion of the total distance as the y-axis is that there is no one-to-one mapping from the distances to the heights of the bars; there are infinitely many distances that can map to the same height. One possible way to solve this issue is to use the logistic function which can map distances in the domain of real numbers to a range between zero and one. That way, the y-axis remains bounded from zero to one for all respondents yet the logistic function is a monotonic function that mirrors differences in the raw distance. For example, a respondents A and B would have different heights with the logistic function, and it would be apparent that the distance for respondent B is larger.

3.2.2 Incorporating respondent and item information

The profiles presented above rely only on the distance terms. Since these distance terms capture unobserved interactions between respondents and items, those profiles can be useful in understanding the differences in the dependencies across items for a given respondent. However, it is important to remember that the probability of a correct response (more specifically, the log-odds) is dependent not only on the distance term but also on the overall respondent and item main effects (see Equation 2.1). The log-odds are therefore dependent on three terms: θ_j, β_i, and $\gamma||\boldsymbol{z}_j - \boldsymbol{w}_i||$ (i.e. the distance/gamma term). That means that even a respondent with a very large distance/gamma term may still have a higher probability of answering an item correctly compared to another re-

spondent with a smaller distance/gamma term but much smaller θ or overall ability. Therefore, to better be able to compare different respondents, it is desirable to incorporate respondent and item information into the profiles to provide additional context to the distance/gamma term.

Although as shown earlier the respondent ranking can be shown in the title of the profile and the bars can be colored by the item easiness parameters, since the respondent and item parameters are numeric, it should make sense to represent them on the bar graph as well. These parameters can be included as part of the bar graphs. Two options are outlined below.

3.2.2.1 Proportion of total

One option is to create a 100-percent stacked bar graph where the y-axis represents percentage of the log-odds attributed to the three terms. To calculate this proportion, the magnitude of each of the three terms is divided by the total log-odds of a correct response for that item. In this bar graph, the proportion of the log-odds attributed to each term is visualized as three bars stacked on top of one another.

This profile makes it easier to evaluate the contributions of each term for a given respondent. For example, most of a given respondent's correct response probability can be attributed to their overall ability θ or to the distance/gamma term. While this profile may not be as useful in seeing which students have higher correct responses probabilities, this profile is more useful in identifying students-item pairs disproportionately affected by dependence. For example, some students may answer a given item correctly not so much because of their overall ability but because the item itself is largely too easy (high β) or that the student has unique mastery of that particular item or domain (high distance/gamma term). This profile allows the user to evaluate the magnitude of these influences. In practice, this is useful because this profile can begin to pinpoint the reasons behind a respondent's success (or lack thereof) in answering an item correctly.

Figure 3.15 shows profiles for the two respondents. It is clear that for respon-

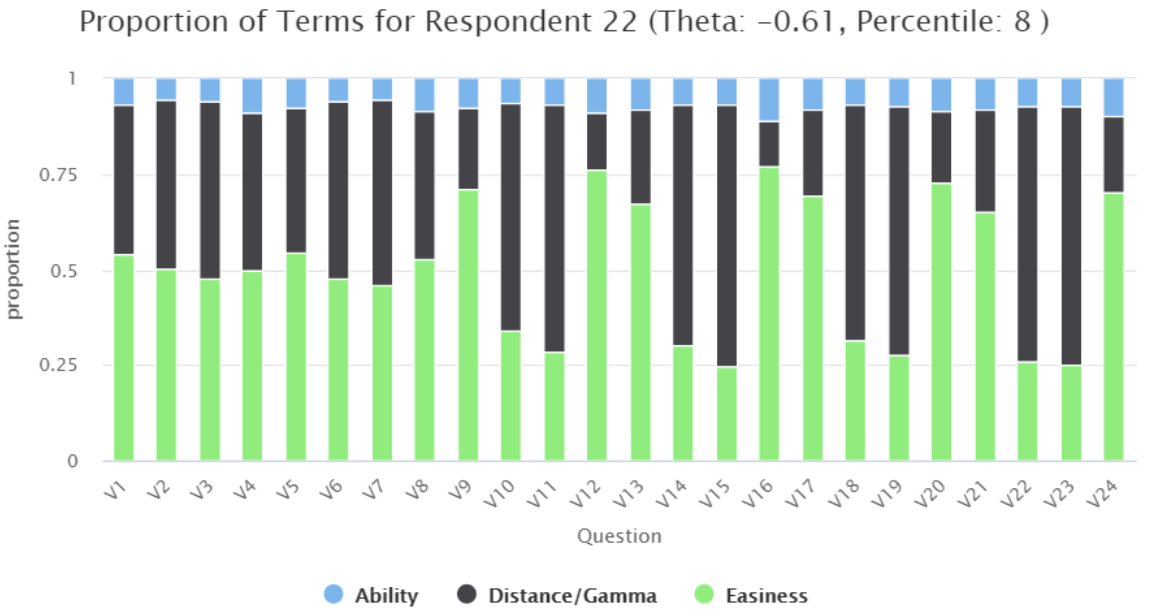
dent 22, the ability parameters contribute little to the log-odds of a correct response, in contrast to the distance/gamma term. Therefore, even for items that the respondent did answer correctly, the correct response probability is mainly driven by the item easiness parameter (i.e. the items were already very easy to answer correctly anyway) and the distance/gamma term (i.e. domain-specific knowledge). In contrast, the contributions from the ability parameters are greater for respondent 407. A tooltip can show the exact percentage contribution. For example, the distance/gamma term comprises about 29% of the combined magnitude of the three terms for item 5.

3.2.2.2 Probability of correct response

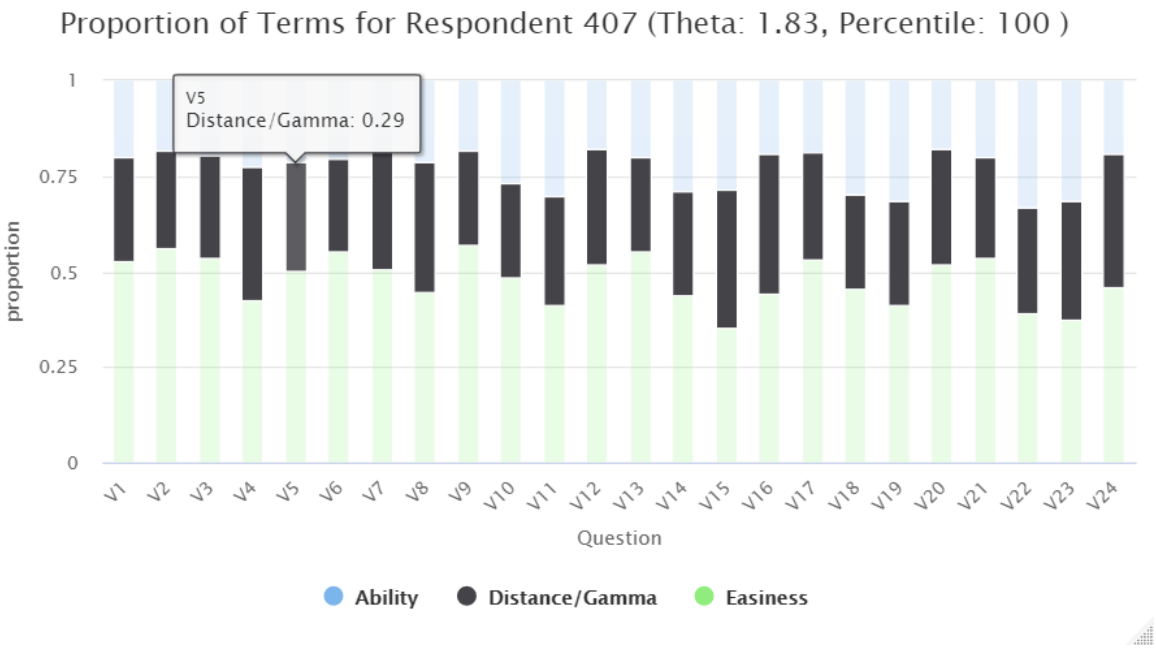
The disadvantage of the previous profiles is that they do not allow for easy comparisons between respondents. Another option is to create a side-by-side bar graph with the y-axis as the probability of a correct response. For each item, the total height of the bar represents the estimated correct response probability for that item. That bar is decomposed by the student's ability (θ), the item easiness (β), and the distance/gamma term. The lengths of those smaller bars are created by multiplying the proportions created in the previous profile by the correct response probability. Note that for a given respondent profile, the height corresponding to θ should be the same since the student overall ability is uniform across all items.

Figure 3.16 shows that by hovering over each bar, the user can see the total estimated correct response probability and the portion (in terms of probability) contributed by the given term for each item.

Figure 3.17 demonstrates these profiles for respondents 22 and 407. For respondent 22, we can see that the heights of the bars, which denote the correct response probabilities, are much lower for items 10, 11, 14, 15, 18, 19, 22, and 23. Respondent 22 struggled particularly with these items, as can also be seen on the interaction map, so an educator may try to work with this student on the domains represented by those items. Each item also has stacked bars, each representing one term. We can see that this



(a) Profile of respondent 22



(b) Profile of respondent 407

Figure 3.15: Profile of two respondents with proportion of terms

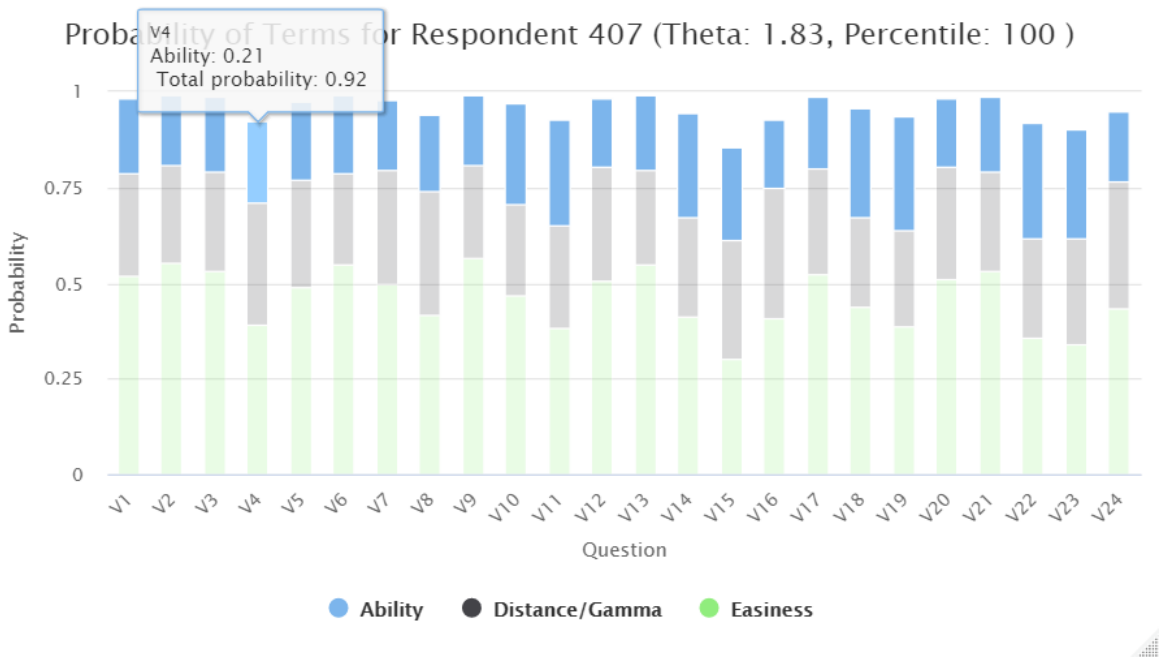
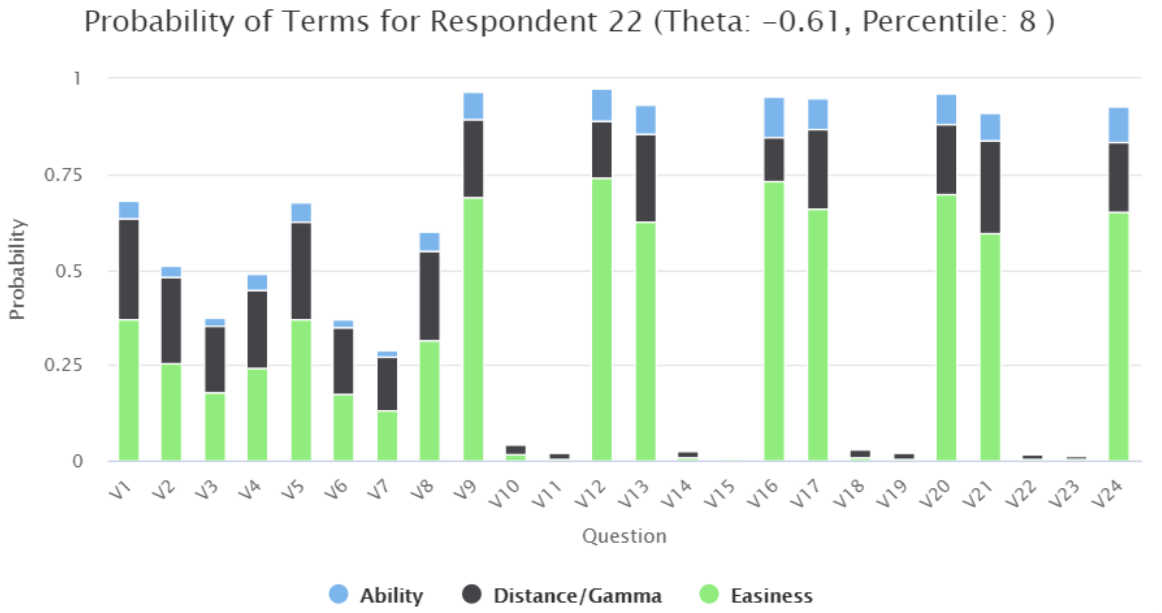


Figure 3.16: Respondent 407 with tooltip showing the total correct response probability and the portion contributed by the term

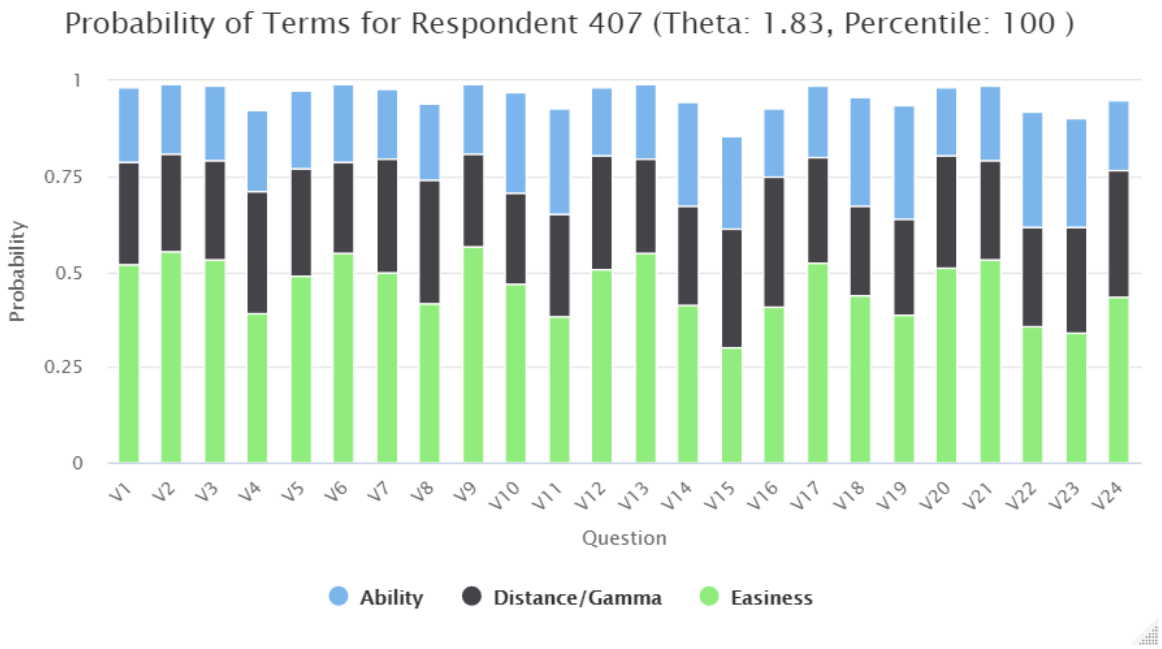
respondent's overall ability is quite low given the small height of the blue bar (in fact, the θ estimate is negative). We can also tell that this respondent has higher correct response probabilities for items which already have high easiness parameters (or longer green bars). In contrast, the distance/gamma terms (represented by black bars) are greater in magnitude for more difficult items.

For respondent 407, we can see that their ability estimates (i.e. the blue bars) are greater, confirming that this respondent has higher overall ability. Additionally, we can see that in general, the distance/gamma terms are relatively equal across the items compared to those from respondent 22. This confirms our earlier observations from Figure 3.14. It seems that respondent 407's relative success on these items are driven mostly by their higher overall ability and smaller distance/gamma terms.

There are a few advantages to this approach. The user can compare the heights of the bars across different respondent profiles and understand which respondent has a higher probability of answering a given item correctly. Furthermore, the stacked bar



(a) Profile of respondent 22



(b) Profile of respondent 407

Figure 3.17: Profile of two respondents with estimated correct response probabilities

graph allows users to see the relative contributions of each term to the correct response probability. That way, a user can understand whether most of the correct response probability can be attributed to the overall respondent ability, the item easiness, or the dependence. For instance, a respondent may only have answered an item correctly not because their overall ability was high but because they have high mastery of the domain represented by that item (but not in other domains). This allows educators to focus on building students' overall understanding.

3.3 Different Dimensions

The interaction map is presented in two-dimensional space in Jeon et al. (2021). In the previous sections, the map and profiles were based on this two-dimensional space. However, what happens when the map is presented in different numbers of dimensions? In other words, what happens when z_j and w_i are n-dimensional vectors, for values of n other than $n = 2$? For visualization purposes, I will focus on the cases when $n = 1$ and $n = 3$.

3.3.1 One Dimension

Figure 3.18 shows the interaction map of the DRV data when $n = 1$. This jitter plot shows the one-dimensional positions of respondents and items along the y-axis and separated along the x-axis. It is the vertical positions of the points that matter, not the horizontal positions since jittering was introduced to avoid overlap. The interaction map shows that the four item groups have been condensed into two, with I1 and I2 combined into one group and I3 and I4 combined into another. The substantive interpretation is that whereas the two-dimensional plot differentiated the items by both Type of Inference and the Content of the Conditional, the one-dimensional plot only differentiated by the Type of Inference. With the fewer number of dimensions, the map creates the groups based on the most differentiating design factor. This could be useful for test designers

who wish to see whether the most differentiating design factor is working as intended. Unfortunately, this plot is not very informative for looking at the relative distances of respondents to items. Much information is lost by removing the additional dimension.

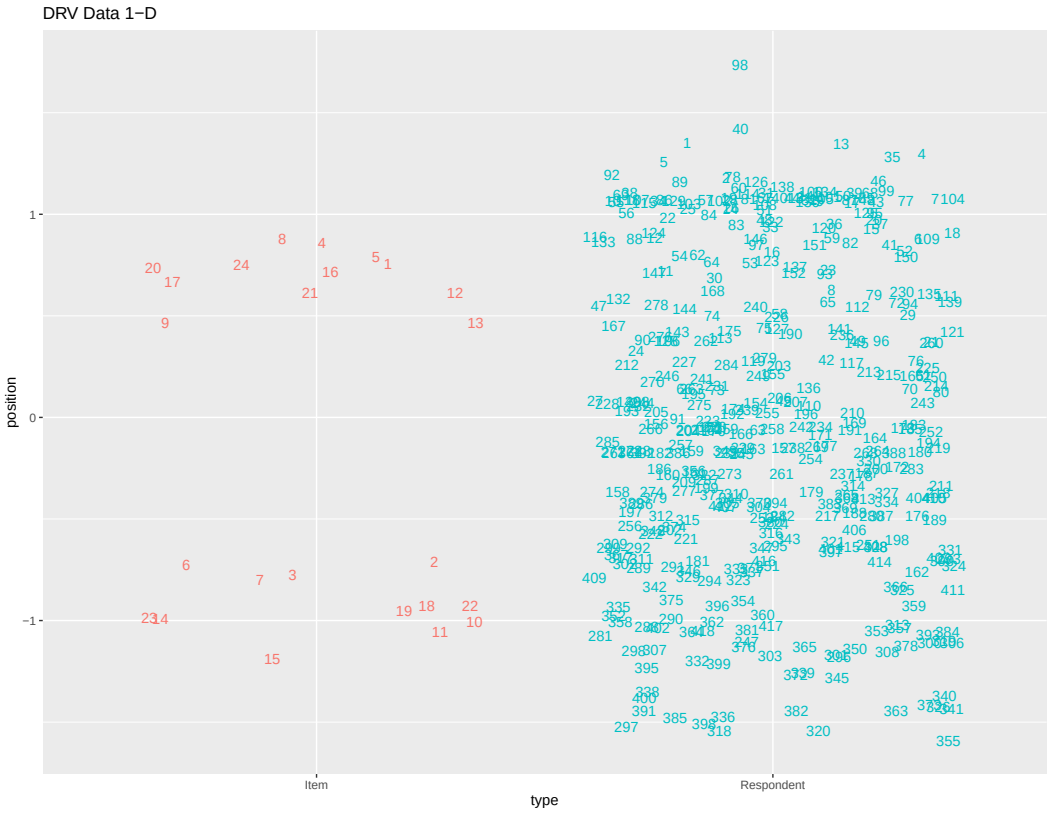


Figure 3.18: Interaction map of the DRV data in one-dimension

3.3.1.1 Comparison with the Wright Map

Notice that this representation resembles a Wright map since the items and respondents are ordered on a single continuum. Figure 3.19 shows the Wright map created from the one parameter logistic (1-PL) Rasch model fitted to the DRV data. In the Wright map, the respondents and items are plotted on the same logit scale. The left-hand side is a histogram of the respondents, with respondents of higher abilities located higher on the continuum and vice versa. On the right-hand side, the items are located based on their item difficulties, with more difficult items located higher on the continuum. The

idea is that respondents located above a given item are more than 50% likely to answer that item correctly, since the respondents' ability parameter is greater than the item's difficulty parameter.

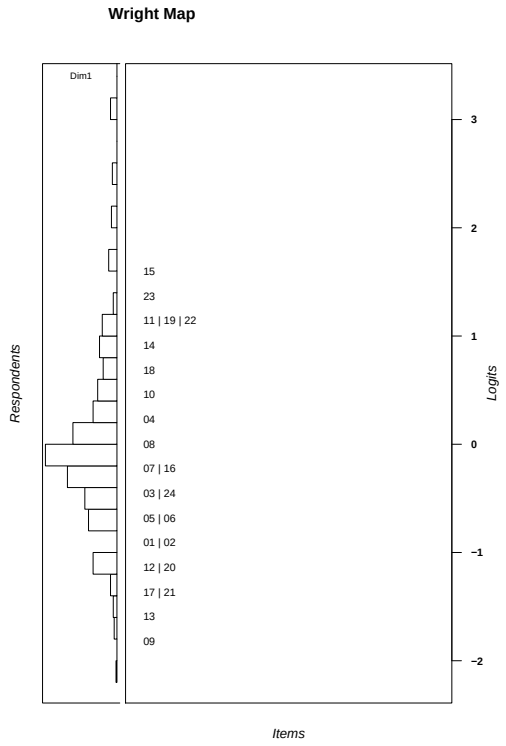


Figure 3.19: Wright map created from the 1-PL Rasch model fitted to the DRV data

In comparison with Figure 3.18, one salient difference is that the item groupings are not as apparent in the Wright map. Furthermore, some of the items appear to be "flipped"; for instance, item 9 is at the bottom of the Wright map whereas item 9 is near the top in Figure 3.18. However, there do appear to be some similarities. For example, items 9, 13, 17, 21, 20, 12, 1, and 5 are clustered together in the bottom of the Wright map while those items are grouped in the top of Figure 3.18. Items 15, 23, 11, 19, 22, 14, 18, and 10 are grouped in the top of the Wright map while they are in the bottom of the interaction map. Despite the fact that the groupings are more apparent in the interaction map, some of the items are close to one another across the interaction map and Wright map.

To further investigate these differences, Figure 3.20 shows the interaction map when the latent space item response model contains only the distance term, not the respondent and item main effects. The idea is that in such a model, the distance term would encapsulate not only the dependencies but also the main effects. Therefore, this new interaction should more closely resemble the Wright map. Indeed, Figure 3.20 shows that the item groupings are mostly intact. Additionally, we see that items 2, 3, 6, and 7 are in between the two item clusters for both maps.

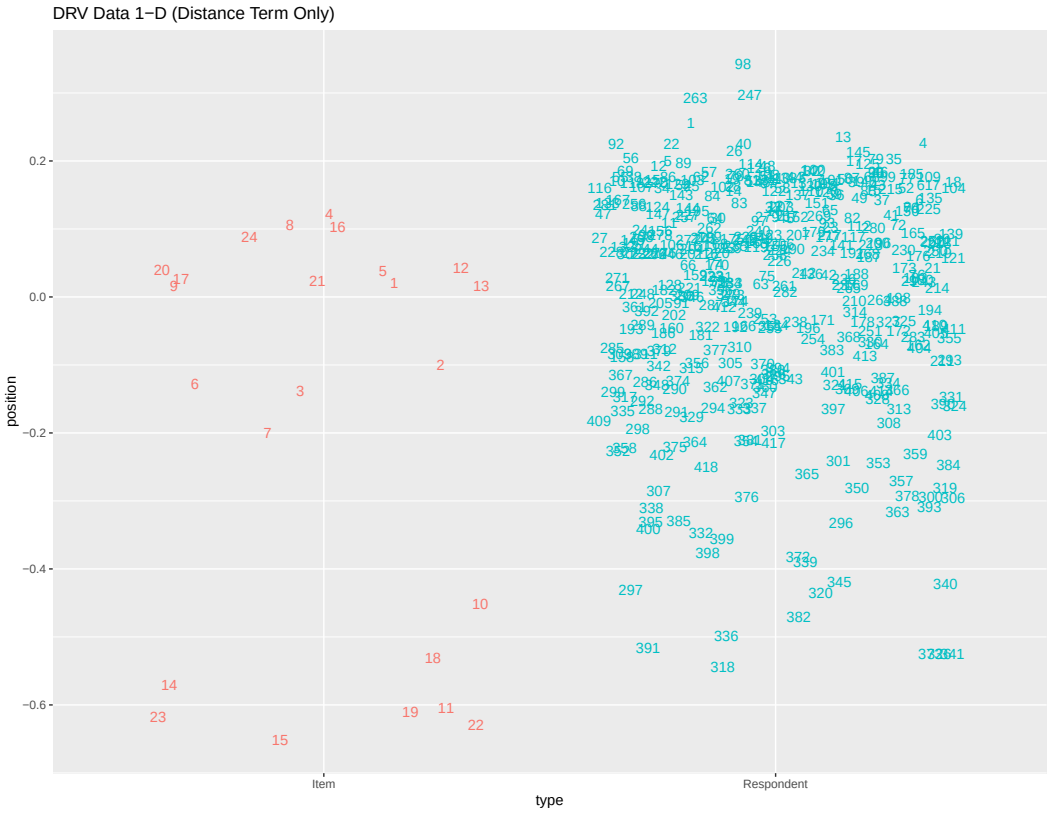


Figure 3.20: Interaction map of the DRV data in one-dimension when the model contains only the distance term

It is possible that these item groupings emerge in the Wright map simply because there are considerable differences between the two item groups in terms of item difficulty. Both the latent space model and Rasch model provide very similar estimates of the item main effects. To test this, the Wright map and interaction map were created for simulated data which involves the item responses of 200 respondents and 20 items.

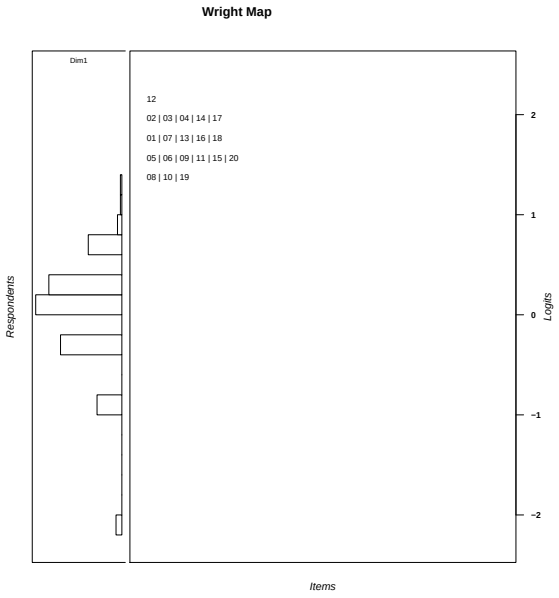
The respondent abilities and item easiness do not vary across respondents or items. However, a dependency term was introduced such that each group of 50 respondents was much more likely to answer a given set of five items correctly. Therefore, there are four different item groups.

Figure 3.21 shows the interaction map and the Wright map. The Wright map in 3.21a is not able to distinguish among the four item groups because the items were simulated to have the same item difficulties. However, the interaction map in 3.21b is able to do so and also show the groups of respondents likely to answer those items correctly.

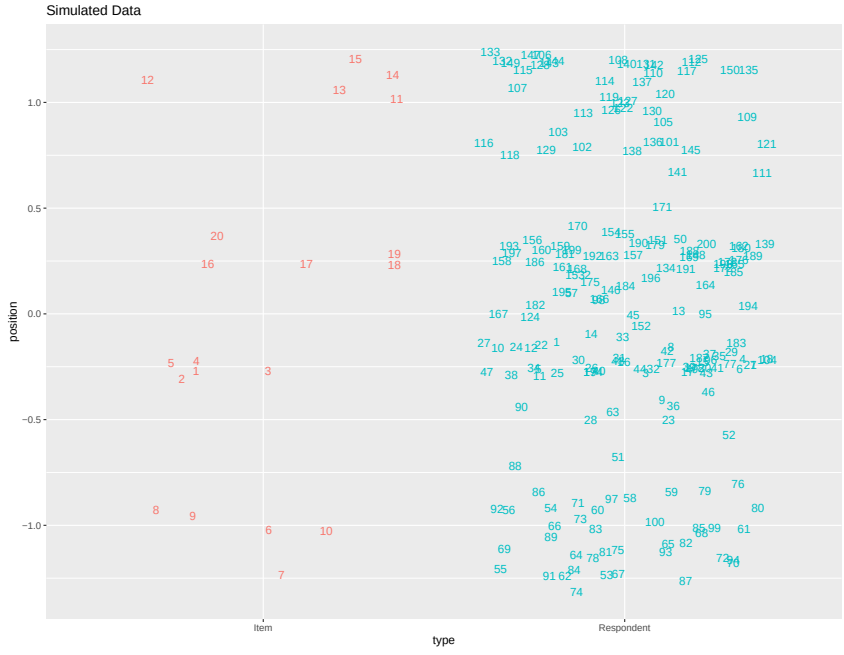
In conclusion, the Wright map may resemble the one-dimensional interaction map, but there are a few differences. The interaction map, in general, is better able to create respondent and item groups that may emerge from dependencies even after accounting for respondent and item main effects. On the other hand, the Wright map is much better at displaying overall respondent and item main effects, since those are not shown in the interaction map. In fact, the numbers on the continuum of the interaction map are not directly interpretable. Therefore, if one suspected that respondents may answer items differently regardless of their own overall abilities, one could use the one-dimensional interaction map. Otherwise, absent any meaningful dependencies or results from the interaction map, one can look at the Wright map for rankings of the respondent and item main effects.

3.3.2 Three Dimensions

Figure 3.22 shows the interaction map when $n = 3$, created using the plotly package in R (Sievert, 2020). The respondents are colored in shades of green with higher ability students colored with lighter green, and the items are colored in shades of red with easier items colored with darker red. The advantage of the three-dimensional plot is that it is better at differentiating respondents and items. For example, in the two-dimensional plot, while respondents of very low and very high abilities were clustered



(a) Wright map created from the simulated data



(b) Interaction map created from the simulated data

Figure 3.21: Wright map and interaction map created from simulated data

in the center together, in the three-dimensional plot, those two groups of respondents are farther from one another. More importantly, instead of the four item groups, it is possible to see six item groups, which are circled.

- Black: items 10, 11, 14, 15
- Pink: items 2, 3, 6, 7
- Yellow: items 18, 19, 22, 23
- Purple: items 17, 20, 21, 24
- Blue: items 1, 4, 5, 8
- Red: items 9, 12, 13, 16

I2 is split into the items circled by black and yellow while I3 is split into items circled by purple and red. It seems that for I2, which consists of the NA and AC items, the items can be further divided into the CO and CF levels of the content of conditional factor. For I3, which consists of MT and MP items, the items can be further divided into the AB and CF levels of the content of conditional factor. This three-dimensional plot is thus better able to discern finer-grained differences among the items, which may be very useful to test designers. Unfortunately, for practical use by educators, these plots may be more challenging to use since the 3-D plot requires interactivity (i.e. dragging and moving the plot on various axes) to properly view the entire space.

3.3.3 Comparisons of Profiles Among Different Dimensions

To demonstrate the differences between the interaction maps of different dimensions with regards to the respondents, the profiles are created for respondent 22.

In the one-dimensional plot, there are two item clusters. I1 refers to the items in the bottom cluster, or the items that pertain to the NA or AC levels of the Type of Inference design factor. I2 refers to the items in the top cluster, or the items that pertain

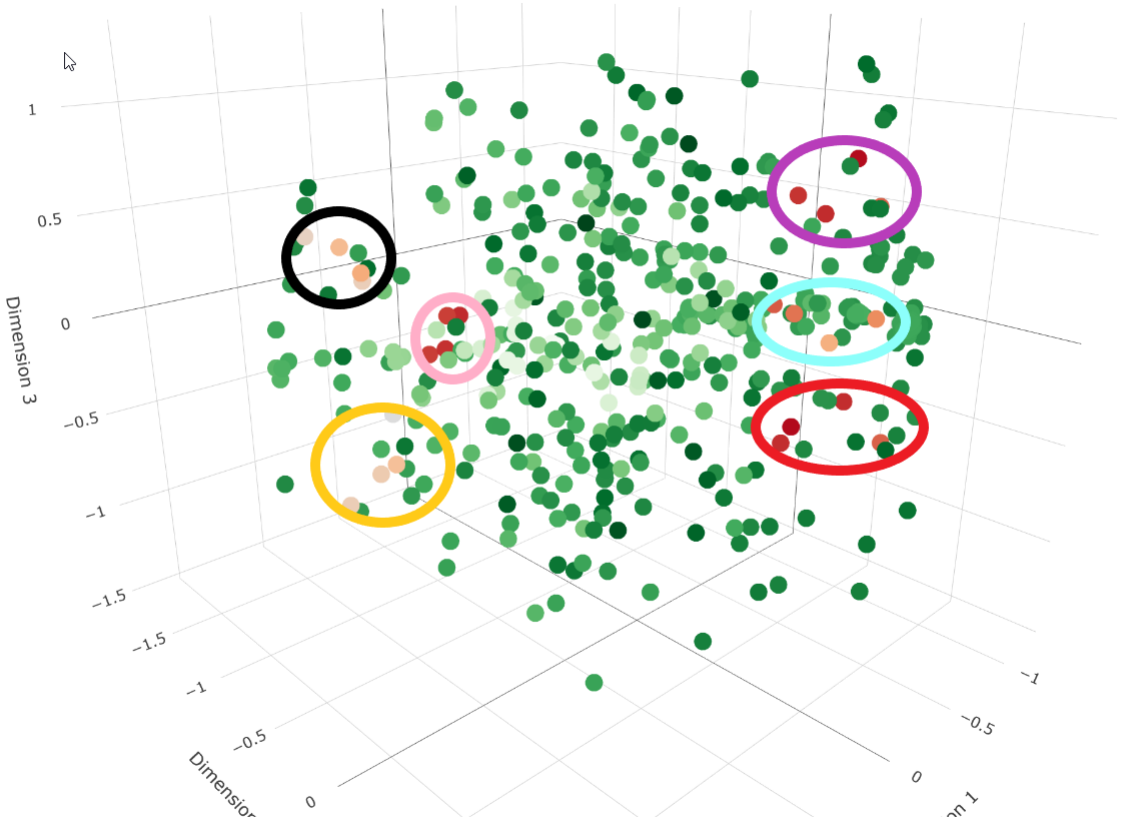


Figure 3.22: Interaction map of the DRV data in three-dimensions

to the MP or MT levels of the design factor. In Figure 3.23, the weakness profile for respondent 22 is shown. The heights of the bars represent the proportion of the total distance comprised of the distance from the respondent to the centroid of each item cluster. The bar for I1 is higher, suggesting that respondent 22 is farther from I1 than from I2. Respondent 22 tends to be weaker on the NA or AC items.

However, the one-dimensional plot provides limited information regarding only one out of the three design factors. In the two-dimensional plot, there are four item clusters. The item clusters correspond to those described in Jeon et al. (2021). Respondent 22 is weaker on items in clusters I1 and I2. I1 and I2 in this case correspond to the items in cluster I1 in the one-dimensional case. I1 and I2 correspond to items with the CO and AB/CF levels of the Content of the Conditional design factor, respectively. I3 and

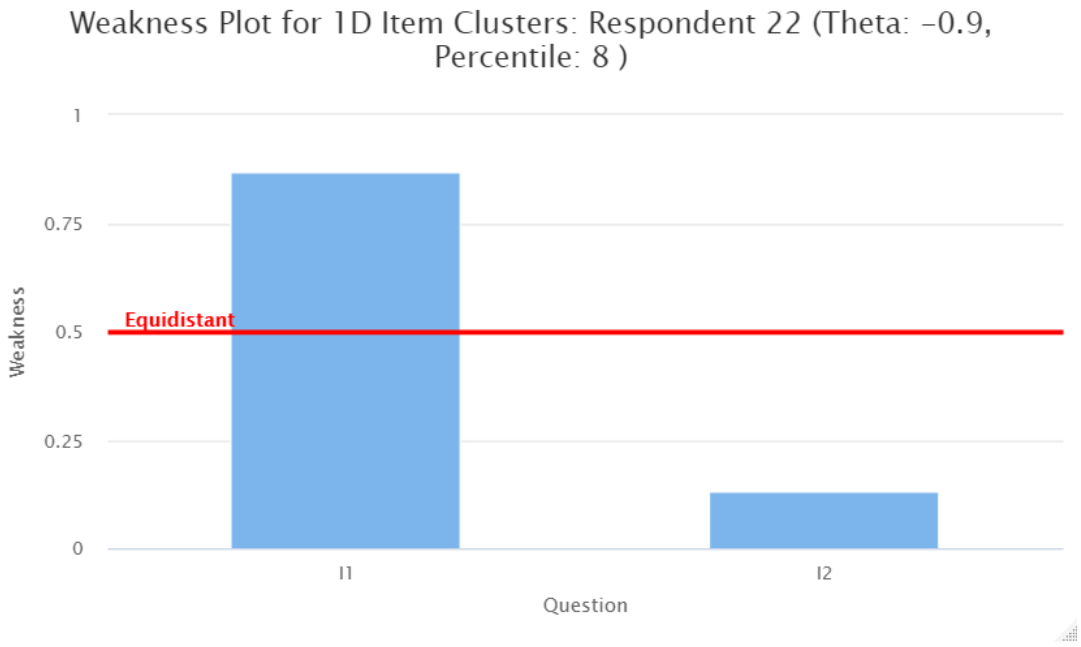


Figure 3.23: Profile of respondent 22 in the one-dimensional latent space

I4 correspond to the items with the AB/CF and CO levels, respectively. It seems that respondent 22 is not so weak on the items with a combination of the AB/CF levels of the Content of the Conditional factor and the MT/MP levels of the Type of Inference design factor.

Finally, in the three-dimensional plot, there are six item clusters. Respondent 22 is particularly weak on item clusters I1, I2, and I3. Those clusters are composed of items from the I1 and I2 clusters in the two-dimensional case. In this case I2 and I3 now correspond to the AB and CF levels, respectively. We can see that respondent 22 is stronger on item clusters I4, I5, and I6, which all contain items with the MT/MP levels. However, the respondent has relatively more difficulty when those MT/MP items have the CO level of the Content of the Conditional factor (I5).

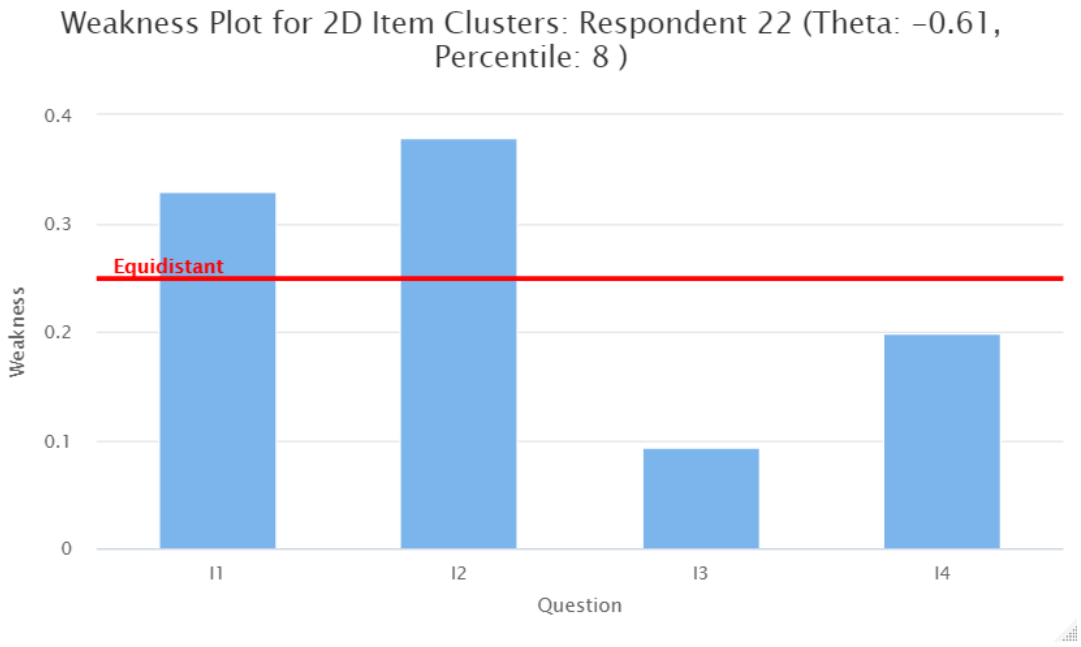


Figure 3.24: Profile of respondent 22 in the two-dimensional latent space

3.4 Summary and Recommendations for Use

There is much potential in the interaction map to yield useful insights. I presented various options to customize the map and profiles. Here I summarize considerations and provide recommendations for practitioners to choose the optimal combination of options for their needs.

3.4.1 Interaction Map

The following are some guidelines for choosing the appropriate interaction map and clustering method.

The original interaction map, as presented in Jeon et al. (2021) and in Section 3.1.1, is a good place to start since it clearly displays the distance term, which is the key contribution of the LSIRM. Educators who are most interested in individual-level performances with regards to certain items should utilize this map. The numbering of the respondents and items makes it easy to find certain respondents the educator is

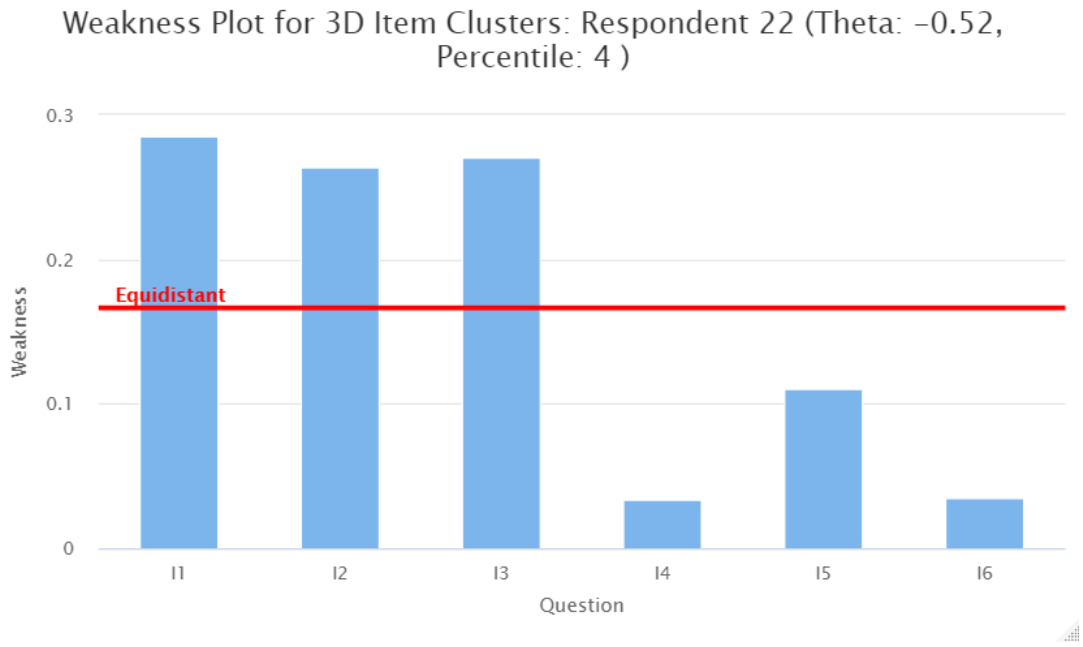


Figure 3.25: Profile of respondent 22 in the three-dimensional latent space

interested in. For example, in PDF form, the user could use CTRL+F to find a certain respondent or item.

The bubble chart, seen in 3.1.2, is perhaps best used as a complement to the original interaction map because it displays additional item and respondent effects. The chart may be better used by researchers or psychometricians who would not only want to see the distances between the respondents and items but also the main effects as estimated by the model. The main effects can be useful for item diagnostic purposes. For example, items which are far too difficult may warrant additional investigation. As seen in Figure 3.3, the items in the bottom part of the map are overall more difficult, thus allowing test developers and researchers to investigate whether this is intended. Furthermore, the bubble chart presents information most practitioners are used to receiving from other software, such as the overall respondent abilities or item difficulties.

This display is also suitable for educators who want a more holistic view of the respondents and items. For example, an educator may see that a given respondent is more likely to answer certain items correctly but may also want to know about

the respondent's overall ability level. In this case, the bubbles can provide this general information and better facilitate comparisons across items and respondents since educators can not only compare respondents based on their distances but also based on their overall ability levels.

Another very important reason to use this bubble chart is that the original interaction map does not differentiate very well between respondents who answer all items incorrectly and those who answer all items correctly. Both groups would be equidistant from the items (in the case of the DRV data, they would be in the center of the map). However, there is no way to distinguish between those two groups. One option would be to remove students who answered all items correctly or incorrectly. This may be justified because the purpose of the interaction map is to help instructors provide targeted instruction to address students' areas of improvement. Therefore, there is no need to visualize students who do not need improvement in all areas or those who need improvement in all areas. However, it can be difficult to decide the cut-off for deciding which students to include. For instance, some students may have answered all but one item correctly. Additionally, some may want to visualize those students anyway. Therefore, the bubble chart, which colors the respondents by their values of θ, solves this issue. This chart becomes necessary for visualizing respondents at both ends of the spectrum - those who answer all items correctly and those who answer them all incorrectly.

The main disadvantage of the bubble charts is that the respondent and item identifiers are not all visible. A compromise would be the chart shown in Figure 3.1.3, which allows the user to hover each bubble to see the identifiers and the associated main effect. The actual magnitudes of the main effects may not interest practitioners (who can see the relative magnitudes based on the sizes or colors of the bubbles), but they may interest researchers.

In terms of clustering, the hierarchical cluster analysis heatmaps are preferable for more exploratory analyses. The heatmaps make it easy to view individual-level distances to items and group them together based on similarities. Practitioners may use the

clustergrams shown in Figure 3.9 to identify respondents who exhibit similar response patterns and group them together for similar instruction to meet their shared needs. On the other hand, the latent profile analysis is more suitable for users with stronger statistical backgrounds or those who wish to conduct more confirmatory analyses. The LPA requires using fit indices to compare different model specifications (i.e. the number of classes) before settling on a single model. However, the clustering from this approach is more model-based and can lead to stronger, more consistent inferences compared to the more qualitative approach provided by the HCA. Additionally, the LPA yields the list of respondents belonging to each class, thereby presenting an easy list for educators to work with (instead of forcing them to manually group students as would be necessary for the HCA). Educators who wish to get a better idea of systematic patterns based on the distances may opt for the HCA first. Then, after a qualitative analysis of the heatmap, one could opt for LPA based on a hypothesized certain number of classes.

3.4.2 Profiles

The following are some guidelines for choosing which appropriate profile to use.

Firstly, the practitioner should look at the profiles based on the raw distances only and focus on the disparities (if any) in the heights of the bars. This answers the question of "Which items may a given student struggle with more than others?" For a student with roughly equal heights of the bars, the practitioner can conclude that the student is equally weak or strong on all the items. On the other hand, unequal bars suggest that the educators may need to work with that particular student on concepts related to certain items.

However, the profiles based on the raw distances and proportion of total distance should generally only be used to compare respondents with similar levels of θ. Therefore, as a second step, it is ideal to interpret the distances along with the respondent and item main effects. The practitioner may choose to look at the profile with the total correct response probability or the proportion of total distance. The former profile answers the

following question: "For respondents whose bars are roughly equal in the distance-only profile, which respondents are very strong on all items and which ones are very weak?" The distance-only profiles should be identical for respondents of very high ability and of very low ability, so the correct response probability profiles are usually for distinguishing between the two, allowing the educator to support students who are struggling on all items and concepts. The latter profile answers the question of "If my student is struggling, how much can be attributed to their overall ability or domain-specific proficiency?" For example, a student may struggle with certain math items because they either lack a strong foundation in mathematics or because they lack specific knowledge required to answer those items. The proportion of total distance profiles allow the educator to understand whether the student would benefit from supplemental instruction pertaining specifically to concepts covered by those items.

In general, it is best for practitioners to begin with the distance-only profiles since those are the easiest to interpret and understand. Those profiles present the most useful information provided by the LSIRM in understanding the disparities in performance on items. Since the correct response probability and proportion of total distance profiles are more complicated to interpret, they are best used as complements to the distance-only profiles. Educators wishing to compare respondents or delve deeper into the performances of their students may find those profiles useful.

3.4.3 Different Dimensions

Finally, the interaction maps presented in different numbers of dimensions can offer different kinds of information. In general, the lower-dimensional interaction maps provide coarser information, potentially obscuring some design factors. Higher-dimensional interaction maps can reveal finer distinctions among items and thereby provide more specific information on the kinds of items that a given respondent may struggle with. However, these maps can be harder to visualize or understand, since the patterns in the 3-D plot are not apparent without the ability to rotate the space.

CHAPTER 4

Model Formulations

The interaction maps and the profiles, seen in the previous chapter, are based on the distance term in the latent space item response model. However, different forms of the distance term in the model can offer different kinds of interaction maps and student profiles, which in turn can lead to different substantive interpretations made by educators. Other specifications are possible. For example, while the distance term presented in Jeon et al. (2021) is the Euclidean distance, Jeon et al. (2021) acknowledge that the multiplicative effect of the respondent and item positions may be an interesting alternative. Therefore, the goal of this chapter is to investigate different formulations of the model which can provide different but equally sensible and useful interpretations.

I begin with smaller modifications to the existing model and end with the most substantial deviation. I first retain the use of the Euclidean distance but change the sign of this term in the model to present alternative "strength" and "weakness" interaction maps and profiles. I then compare interaction maps created from different Minkowski distances of the distance effect. I finally investigate using the multiplicative effect instead, specifically using the cosine similarity function, which can potentially render the main effects more comparable and interpretable.

I will compare the results from the altered models to those from the original model based on the Competence Profile Test of Deductive Reasoning—Verbal assessment (DRV) dataset analyzed in Jeon et al. (2021). A comparison of the interaction maps and the parameters estimates can elucidate the different interpretations that each of the specifications may provide.

4.1 "Weakness" and "Strength" Profiles

The formulation of the latent space model presented in Jeon et al. (2021) is

$$\text{logit}(P(y_{ji} = 1|\theta_j, \beta_i, \gamma, \boldsymbol{z}_j, \boldsymbol{w}_i)) = \theta_j + \beta_i - \gamma||\boldsymbol{z}_j - \boldsymbol{w}_i|| \tag{4.1}$$

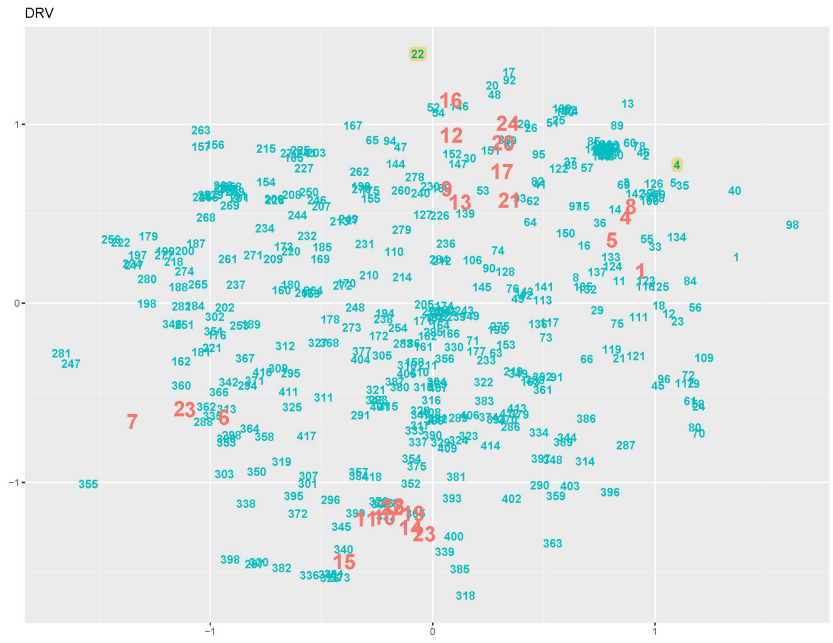
.

However, one alternative formulation ("reverse") involves an additive distance term like so:

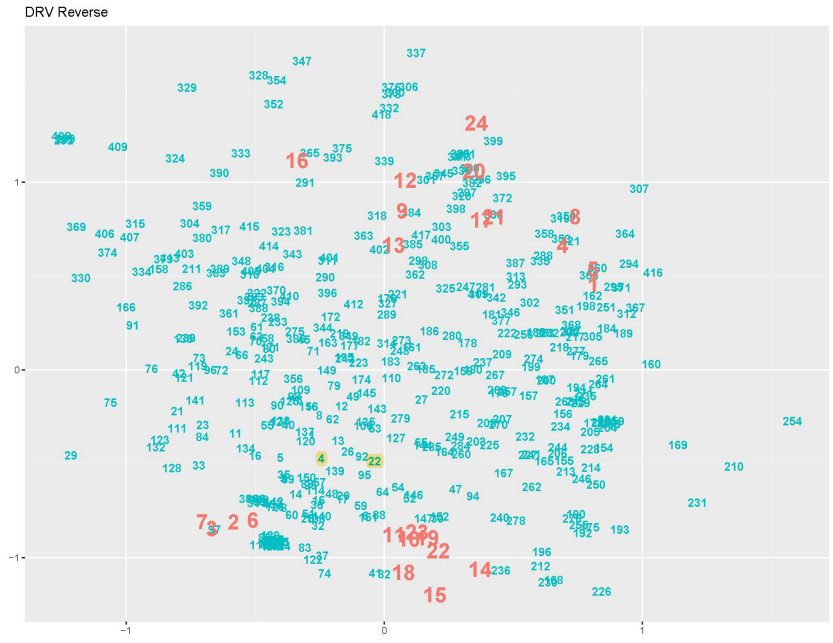
$$\text{logit}(P(y_{ji} = 1|\theta_j, \beta_i, \gamma, \boldsymbol{z}_j, \boldsymbol{w}_i)) = \theta_j + \beta_i + \gamma||\boldsymbol{z}_j - \boldsymbol{w}_i|| \tag{4.2}$$

These two formulations offer potentially different interpretations. Equation 4.1 involves the distance as a penalty term for the log-odds of answering the item correctly; as the distance increases, the probability that the respondent answers the item correctly decreases. Equation 4.2, in contrast, suggests that as the distance increases, so too does the probability of the respondent answering the item correctly. The main reason for the second formulation is that the distances can be interpreted as strengths instead of weaknesses. This provides additional flexibility for educators who may wish to see strength profiles instead of weakness profiles. To explore these differences, the interaction maps from both models are fitted to the DRV dataset.

The interaction maps from the original model and the reverse model fit for the DRV data are shown in Figure 4.1. The item clusters resemble each other regardless of the model. However, note that the respondent positions are generally mirrored. For example, respondents 4 and 22 are highlighted, and their positions seem to be mirrored across the different models; whereas they were close to the top two item clusters, now they are close to the bottom two clusters. In the interaction map for the original model, respondents 4 and 22 are closer to the top two item clusters because they are more likely to respond correctly to those items; in the other interaction map, they are closer to the top two item clusters because they are more likely to respond *incorrectly* to those items.



(a) Interaction map under Equation 4.1



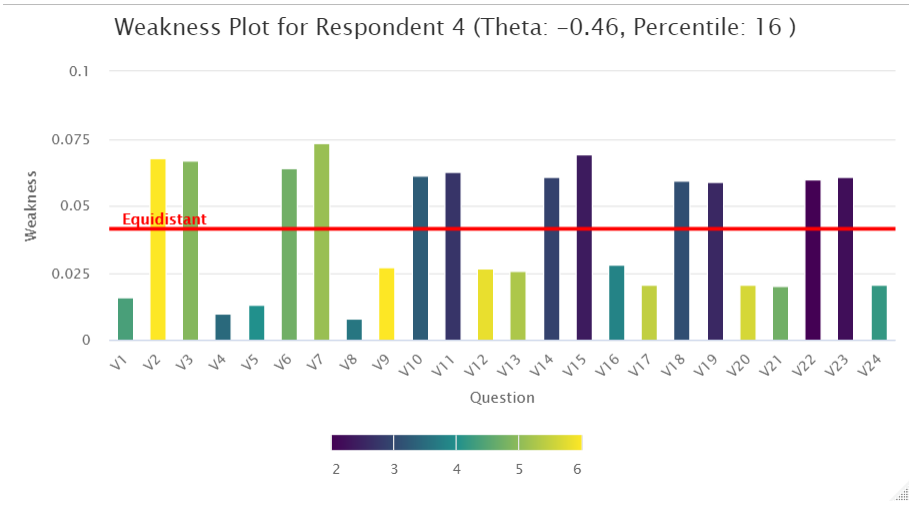
(b) Interaction map under Equation 4.2

Figure 4.1: Interaction maps from the DRV data

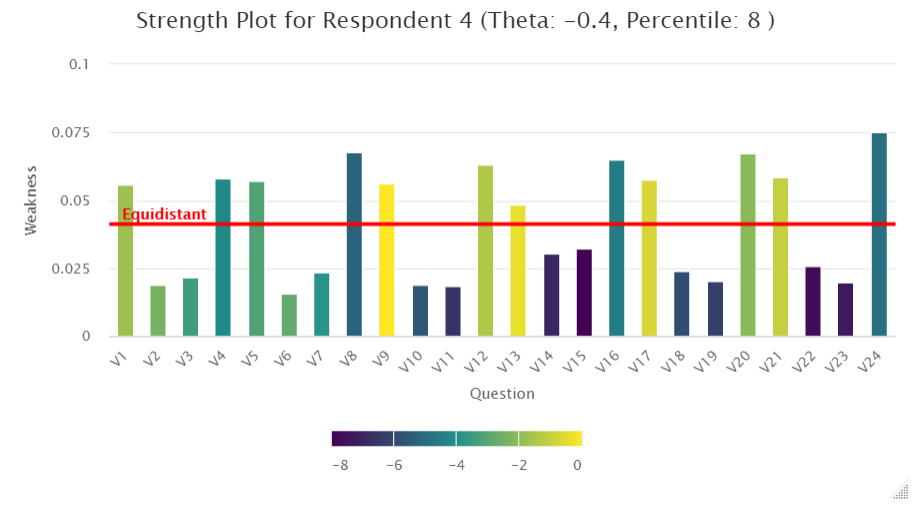
To substantiate these qualitative observations, the cosine similarities can be calculated for each respondent, based on the two positions they have for the two models. Cosine similarity can be used to investigate the similarities of positions; it can determine whether the positions, which can be represented by vectors, are in the same direction, orthogonal, or in the opposite directions. For example, two interaction maps can be created from two different models from the same dataset. A given student would have two different positions, one for each interaction map. Cosine similarity could be used to analyze those two different positions to see whether the student has similar positions or opposite positions across the two models. The cosine similarity of two vectors can take on values between -1 and 1. The cosine similarity is -1 if two vectors are pointing in opposite direction, 0 if they are orthogonal, and 1 if they point in the same direction.

Generally the similarities are close to negative one, suggesting that the position vectors point in the opposite directions. This means that in the reverse model formulation, the respondents are closer to the items they are less likely to answer correctly and vice versa. Hence this shows that the reverse model can be used to create strength profiles.

Finally, Figure 4.2 shows the profiles for respondent 4 under the two models. Note that Figure 4.2a shows that respondent 4 is weaker on items 2, 3, 6, 7, 10, 11, 14, 15, 18, 19, 22, and 23 based on the height of the bars. Figure 4.2b shows similar information but as a strength profile. We can see that respondent 4 is strong on items 1, 4, 5, 8, 9, 12, 13, 16, 17, 20, 21, and 24. This shows us that the two models can provide different yet equally important and corroborating information. An instructor could look at Figure 4.2a, see the high bars for items 2 and 3 and understand that the student is much weaker on those items. Or, the instructor could look at Figure 4.2b, see the high bars for items 1 and 24 and understand that the student is strong on those items.



(a) Respondent profile under Equation 4.1



(b) Respondent profile under Equation 4.2

Figure 4.2: Profiles for respondent 4 in the DRV data.

4.2 Different Minkowski Distances

The model in 4.1 uses the Euclidean distance. The Euclidean distance is a special case of the Minkowski distance. The Minkowski distance of order p between two points $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is

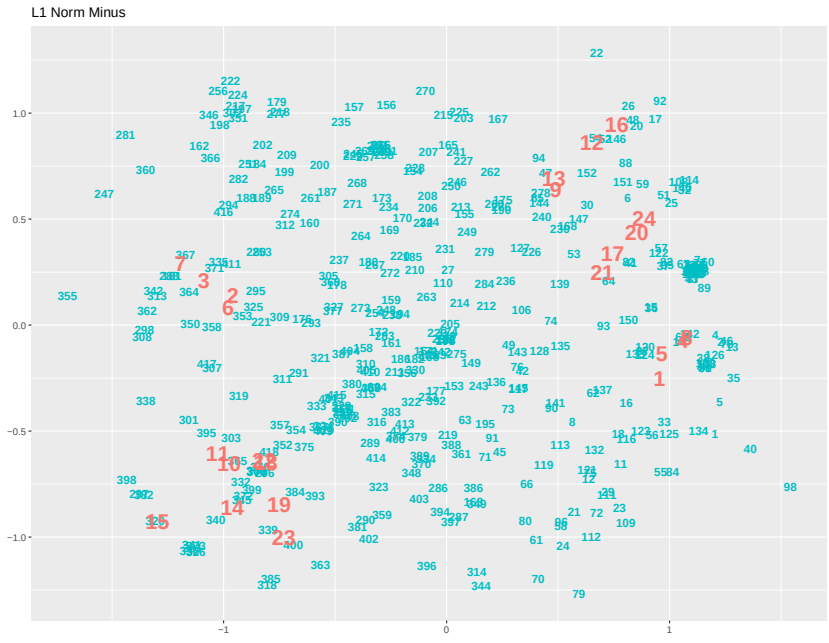
$$D(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{4.3}$$

The Euclidean distance is the special case when $p = 2$. The Minkowski distance of a vector can be defined as a metric for $p \geq 1$. The question is, does the choice of p affect the results from the model?

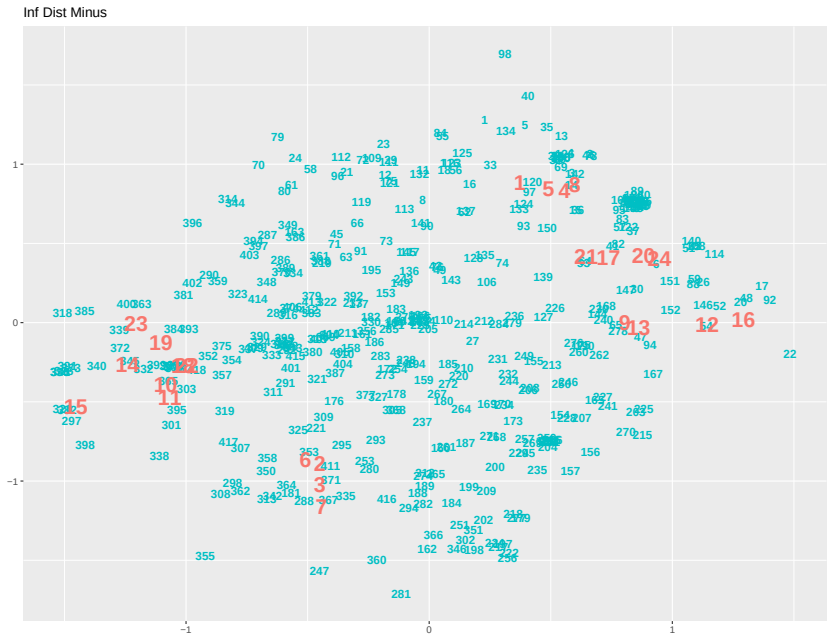
Preliminary investigations suggest that the choice of p does not significantly change the interpretation of the interaction maps. Figure 4.3 shows the interaction maps created when the Euclidean distance is replaced by the Manhattan distance ($p = 1$) and the Chebyshev distance ($p = \infty$). The interaction maps are similar to those created from the model with the Euclidean norm. Although the positions of the item clusters are different, the positions of the respondents relative to the four item groups remain relatively unchanged.

Since the interaction maps can be rotated in infinitely many ways, it is ultimately important to assess whether the respondent-to-item distances are largely preserved. We assume that the Euclidean distance will still be used to generate the profiles from the models that do not necessarily use the Euclidean distance for the distance term. That is because the Euclidean distance, or the straight-line distance, is the easiest to interpret on the generated interaction maps.

The bipartite matrix of such distances was calculated for each of the two different Minkowski distances. In other words, for each Minkowski distance, the distance of each item from each respondent was calculated. The correlation of these respondent-item distances was calculated. The lowest correlation was 0.85 with a median of 0.9984, suggesting excellent consistency of the distances across both Minkowski distances. This



(a) Interaction map when $p = 1$ (Manhattan distance)



(b) Interaction map when $p = \infty$ (Chebyshev distance)

Figure 4.3: Interaction maps when $p = 1$ and $p = \infty$

confirms our qualitative finding that the interaction maps (and therefore, the profiles) do not change much regardless of the choice of Minkowski distance. Furthermore, the estimates of the respondent θ and the item β are consistent across these models, with correlations all greater than 0.95

For practitioners, it is enough to know that the conclusions they would draw from the interaction maps and profiles would not change for different values of p in the distance term. Future research might focus on understanding why these results are robust to such changes in the model.

4.3 Cosine Similarity

Another option is using cosine similarity itself as the distance measure. This model can be specified as

$$\text{logit}(P(y_{ji} = 1|\theta_j, \beta_i, \gamma, \boldsymbol{z}_j, \boldsymbol{w}_i)) = \theta_j + \beta_i + \gamma \cos\theta \tag{4.4}$$

where $\cos\theta$ is the cosine similarity between $\boldsymbol{z}_j$ and $\boldsymbol{w}_i$.

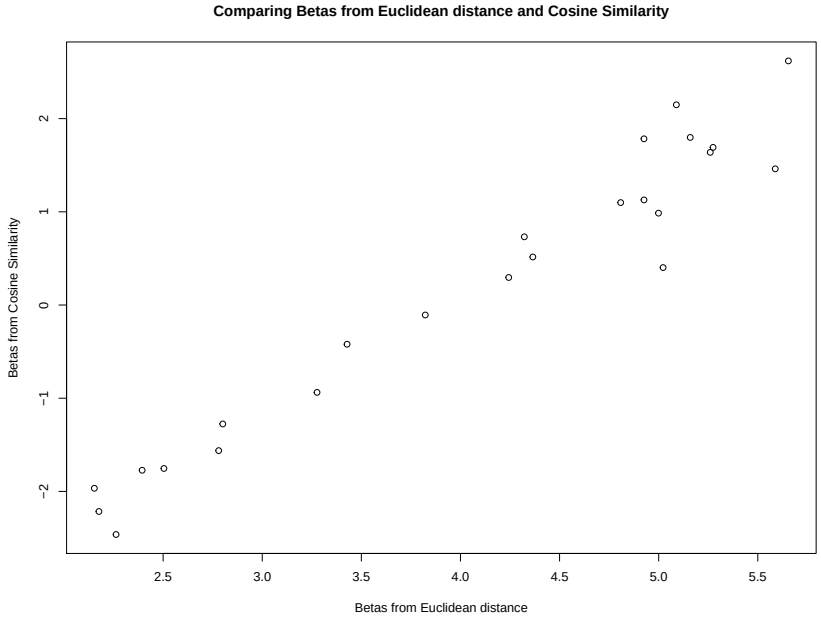
Unlike the Euclidean distance or any L_p norm, which must be non-negative, cosine similarity ranges from -1 to 1. This provides additional flexibility to the model because while the Euclidean distance can only be subtractive (original model) or additive (reverse model) to the log-odds, cosine similarity can act as both. For example, if the model includes an additive distance term, then a respondent who is likely to answer an item correctly, even accounting for item easiness and respondent ability, should have a positive value for the cosine similarity. Likewise, a respondent not likely to answer the item correctly would have a negative value. This is not possible for the Euclidean distance. Additionally, it can be easier to compare distances across different respondents using cosine similarity because the values are constrained between -1 and 1, whereas the Euclidean distance ranges from zero to positive infinity.

Since the Euclidean distances are constrained to be non-negative, the model

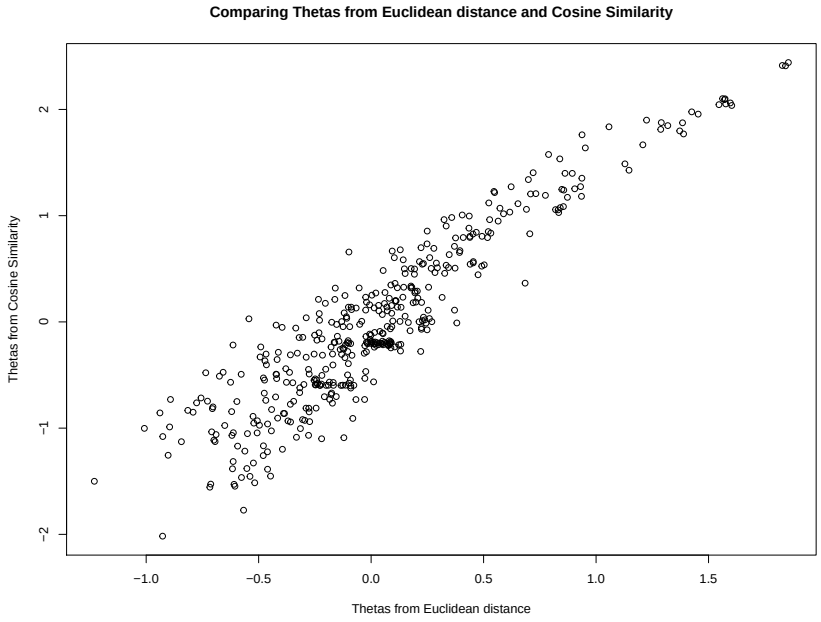
tends to "compensate" by inflating estimates of the item easiness parameters β. Figure 4.4 compares estimates of θ and β using the Euclidean distance and the cosine similarity as distance terms. In general, there is a positive correlation between the two estimates, suggesting that the ordering of respondents and items by their parameter estimates is not affected by the choice of model. However, the scales of the axes in Figure 4.4a differ. Specifically, the x-axis, which denotes the β estimates from the Euclidean distance model (Model 4.1), suggests that the estimates are larger than those from the cosine similarity model. Hypothetically, one can interpret the β from the Euclidean distance model as the "maximum" item easiness that would be achieved were it not for the penalty imposed by the distance term. While this is a possible substantive interpretation for the inflated β estimates, it is not the easiest to understand. In contrast, the β estimates from Model 4.4 are not inflated and can be interpreted similarly as the estimates that would be attained from the original Rasch model. In fact, the θ and β are roughly on the same scale under Model 4.4, allowing them to be compared on the same scale.

This fact is made more apparent by comparing the respondent profiles in Figures 3.15 and 4.5. The proportion accounted for by β is quite larger in Figure 3.15 due to the inflated estimates of β. In contrast, the proportion accounted for by β is smaller in Figure 4.5. This can aid in the interpretability of the profiles, since educators viewing the profiles in Figure 3.15 may be confused as to why the item easiness seems to have such a large influence on the correct response probability, both for the weaker student 22 and for the stronger student 407. Figure 4.5 makes it clear that a student's probability of answering an item correctly is not completely driven by item easiness but by the other two factors as well.

Additionally, the interaction map in Figure 4.6 created under Model 4.4 retains the item groupings and patterns found in Figure 4.1a. Although Jeon et al. (2021) express reservations regarding the use of cosine similarity in the model, it seems that respondents who are likely to answer items correctly are still grouped closely to those items and vice versa. It does not seem that there is information loss in that regard.

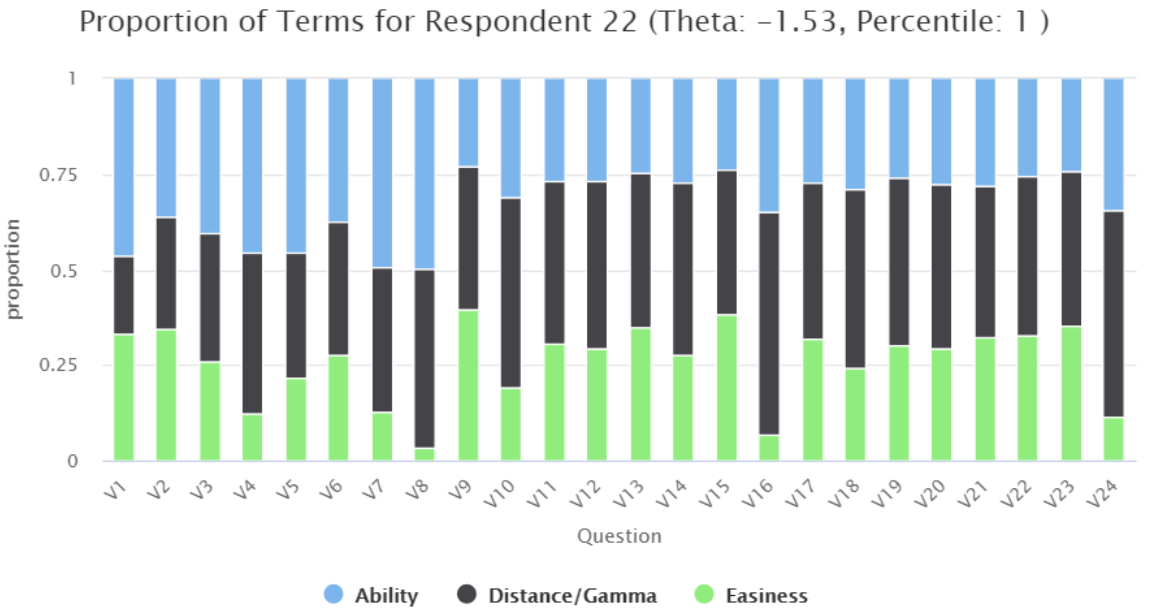


(a) Comparison of β using the Euclidean distance and cosine similarity

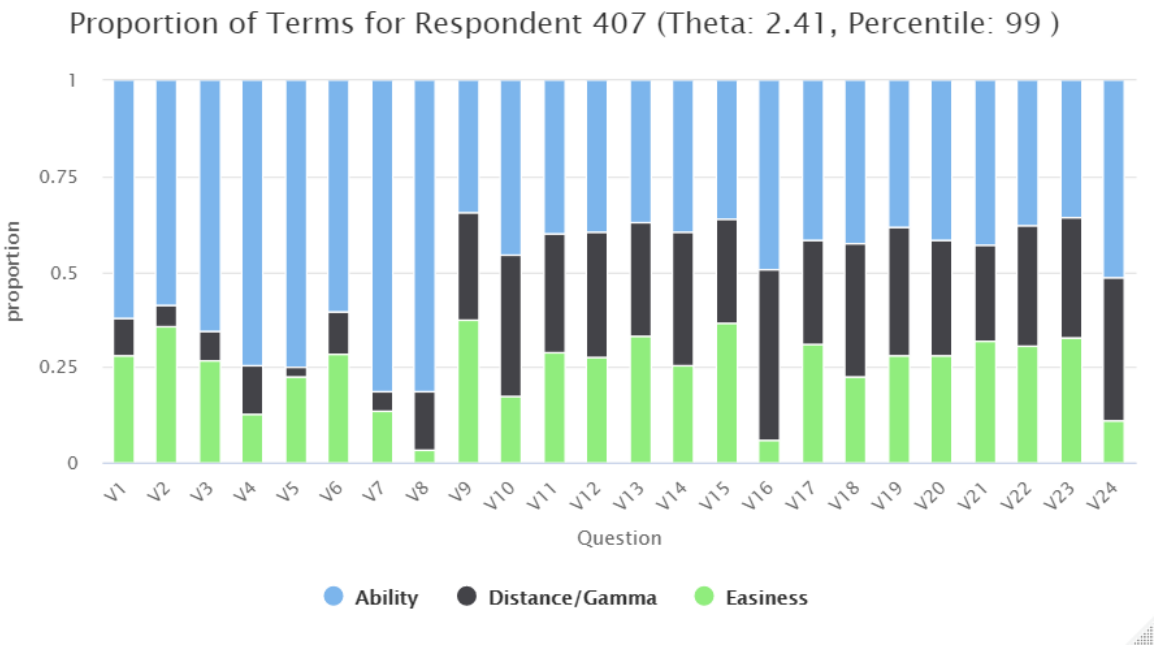


(b) Comparison of θ using the Euclidean distance and cosine similarity

Figure 4.4: Comparison of θ and β using the Euclidean distance and cosine similarity



(a) Profile of respondent 22 under the cosine similarity distance term



(b) Profile of respondent 407 under the cosine similarity distance term

Figure 4.5: Profile of two respondents with proportions under cosine similarity

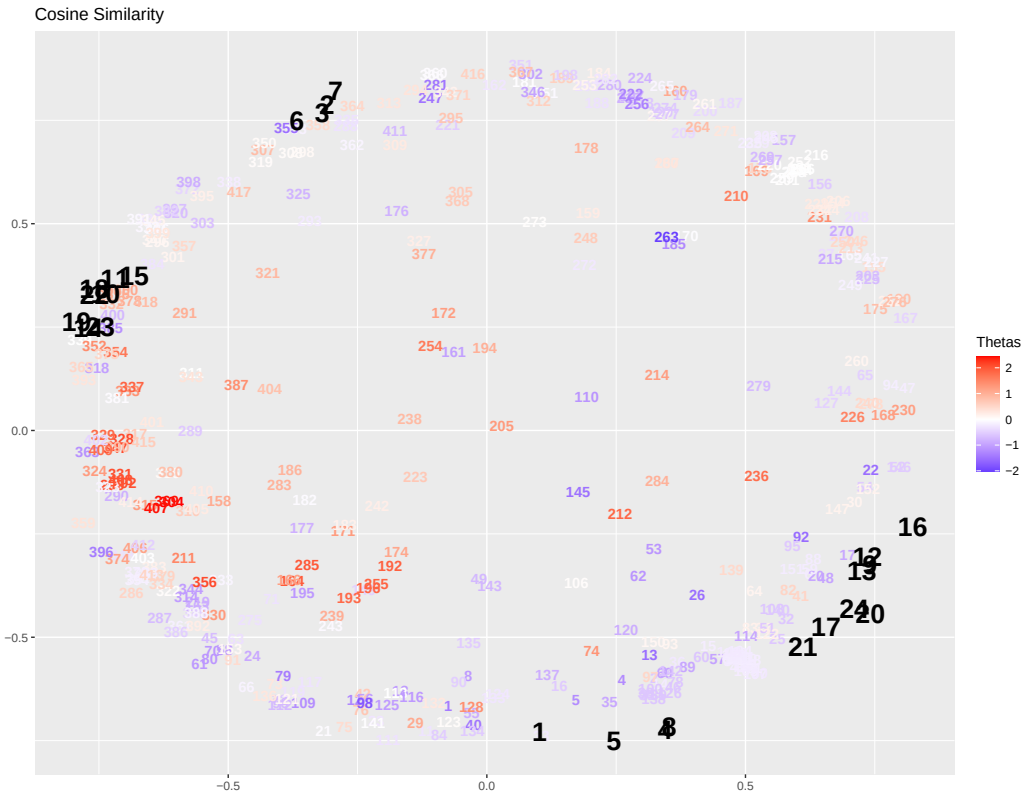


Figure 4.6: Interaction map of the DRV data using cosine similarity as the distance measure.

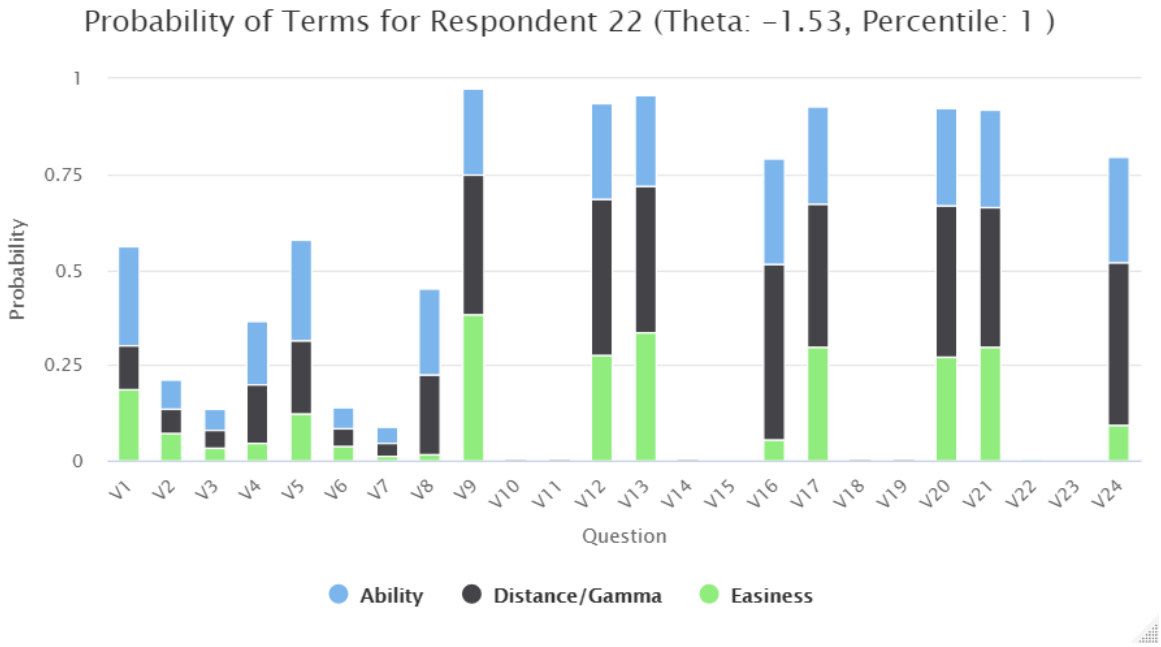
The respondents in Figure 4.6 are colored by their values of θ to show an interesting observation that respondents of very high ability are no longer clustered in the center (as seen in Figure 4.1a) but rather in the left-hand part of the map (see respondent 407, for example). One downside to this formulation is that it becomes harder to explain why such respondents would be in that part of the map, if they were equally likely to answer all items correctly. The solution to this is to examine the profiles, as seen in Figure 4.7. This figure resembles Figure 3.17 in showing that respondent 407 is very likely to answer all items correctly without the inflated estimates of β. Despite the larger imbalances among the cosine similarity distance terms for respondent 407, these imbalances do not matter so much since the respondent's high θ ensures that their log-odds of answering an item correctly would be very high anyway. In that case, the estimated values of the distance term may not matter so much. Therefore, as was discussed in

previous formulations, it is very important to interpret the distance term, Euclidean or cosine similarity, in the context of the item easiness and respondent ability parameters.

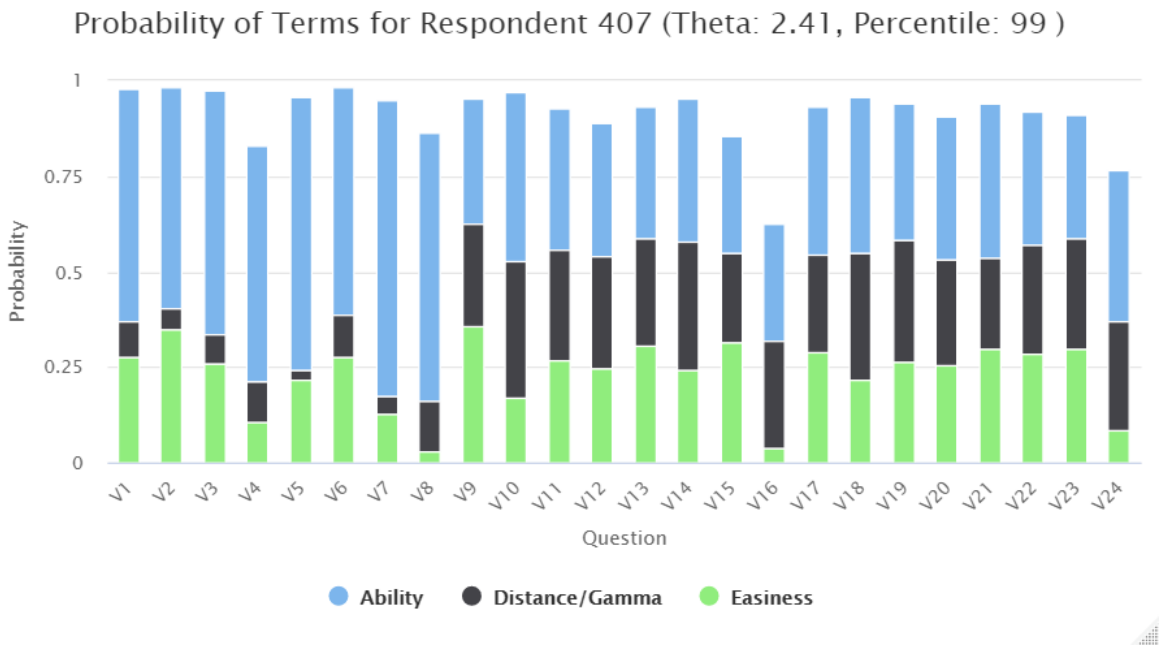
4.4 Summary

In this chapter, I first showed how the model can be slightly modified, with a different sign for the distance term, to yield strength and weakness interaction maps and profiles. This may help educators who wish to see the same information presented in different ways, whether they prefer to view the distances as measures of student strengths or weaknesses on certain items. Furthermore, I showed that the interaction maps are relatively robust to the different specifications of the distance terms. Among the different Minkowski distances presented, the interaction maps look very similar, and the respondent-item distances and main effects are consistent. Additional investigations could involve simulation studies to understand the conditions (if any) under which the maps or profiles may differ depending on the Minkowski distance used and clarify why the results may be robust to different specifications of these distances.

The use of the cosine similarity may be of more concern to practitioners. Using the Euclidean distance, or any non-negative distance effect, may lead to inflated estimates of the item easiness parameters. Preliminary investigations have shown that the use of the cosine similarity may attenuate this problem while retaining most of the respondent and item clustering seen in the interaction map created from the Euclidean distance. However, interpreting the cosine similarity distances may be more confusing since those ”distances” are no longer strictly non-negative. Therefore, practitioners may choose to use the Euclidean distance effect if it is difficult for them to interpret ”distances” that may be negative. However, the cosine similarity may be used if the educator is more concerned with interpreting and comparing respondent abilities and item easiness while also visualizing item and respondent groupings on the interaction map.



(a) Profile of respondent 22 under the cosine similarity distance term



(b) Profile of respondent 407 under the cosine similarity distance term

Figure 4.7: Profile of two respondents with raw values under cosine similarity

CHAPTER 5

Reliability and Validity

The purpose of this chapter is to investigate the reliability and validity of the profiles described in the previous chapter. While Jeon et al. (2021) investigated the main effects of the model, they did not evaluate the reliability of the distances in the latent space. Since the profiles are based on the distances estimated from the model, it is important to evaluate the reliability and validity of these distances to ensure that these profiles are also reliable and valid for use by educational practitioners. Additionally, this investigation would allow us to understand the circumstances under which the distances (and thereby the map and profiles) are reliable.

I first describe the simulated data that will simulate the dependence in the data we hope to capture and that we will use to investigate reliability and validity. I then present two definitions of reliability that will be used to gauge the reliability of the distance term under various conditions. Finally, I show preliminary evidence of the validity of the distance term in its ability to recover the dependency introduced in the simulated data.

5.1 Simulated Data to Investigate Reliability and Validity

The validity and reliability of the distances can be assessed using simulated data. In the case of reliability, this allows us to investigate how various factors such as number of items can affect the reliability of the distances and thus the profiles. The data is simulated following the procedure described in Appendix A of the supplementary materials presented in Jeon et al. (2021). Binary item response data is generated from

the following model

$$\text{logit}(P(y_{ji} = 1|\theta_j, \beta_i, \gamma, \boldsymbol{z}_j, \boldsymbol{w}_i)) = \theta_j + \beta_i + \xi_{ji}$$

where $\xi_{ji} \in \mathbb{R}$ represents the interaction effect between respondent j's ability and item i's easiness. This effect influences the correct response probability beyond the main effects of θ_j and β_i; hence, ξ_{ji} induces local dependence. Local dependence can be simulated by generating ξ_{ji} from a normal distribution with a given mean and variance. For example, ξ_{ji} can be generated from a normal distribution with mean of 2 and variance of 0.2 for certain respondent-item pairs to ensure that those respondents are more likely to answer those items correctly even after accounting for their abilities and item easiness.

5.2 Reliability

While there are multiple ways to define reliability, I present two definitions of reliability that will be used to evaluate the reliability of the distances. The first comes from Templin and Bradshaw (2013) who used this method to evaluate the reliability of CDMs. The second follows from the intraclass correlation (ICC) definition of reliability.

5.2.1 Estimating Reliability (Templin & Bradshaw, 2013)

One can treat these distances as another kind of continuous latent variable estimated from an IRT model. Therefore, the reliability of the distances can be estimated using the simulation-based resampling approach proposed by Templin and Bradshaw (2013) based on test-retest reliability. The algorithm used to calculate the reliability of the distances for a given item i is as follows:

1. Calculate the distance from each respondent e to item i for each iteration kept from the MCMC.

2. Calculate the posterior mean $\widehat{\theta_{ei}}$ and the standard deviation $\widehat{\sigma_{ei}}$ of the distances for

each respondent.

3. Randomly sample one respondent.
4. Draw two values θ_{e1} and θ_{e2} from the distribution $\mathcal{N}\left(\widehat{\theta_{ei}}, \widehat{\sigma_{ei}}\right)$.
5. Repeat steps 3 and 4 many times.
6. Calculate the Pearson correlation between the set of $(\theta_{e1}, \theta_{e2})$ for all respondents.

These steps are repeated for each item to yield reliability coefficients for all the items.

5.2.2 Estimating Reliability (ICC)

The second way to calculate reliability comes from the definition of the intraclass correlation coefficient (ICC). In this definition, the reliability ρ_θ is estimated as the ratio of the between-respondent (or true score) variance σ_b^2 to the total variance σ_t^2:

$$\hat{\rho}_\theta = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_t^2}$$

where the total variance is the sum of between-respondent variance (”between variance”) and within-respondent (or error) variance (”within variance”).

$$\hat{\sigma}_t^2 = \hat{\sigma}_b^2 + \hat{\sigma}_w^2$$

In the estimation of the latent space item response model, each iteration of the MCMC yields respondent and item positions which can be used to calculate the pairwise distances between respondents and items. If the MCMC chain has converged, we can be reasonably sure that we are sampling from the posterior distribution (after the burn-in period). Therefore, we can use these iterations to calculate the within variance $\hat{\sigma}_w^2$ and between variance $\hat{\sigma}_b^2$. The following formulas calculate the reliability of the distances for a given item.

Let N be the number of respondents and M be the number of kept iterations. Let $\hat{\theta}_{im}$ be the distance from a respondent i to a given item in the m-th iteration. Then the between variance can be calculated as

$$\hat{\sigma}_b^2 = \frac{\sum_N^{i=1} (\bar{\theta} - \hat{\theta}_i)^2}{N}$$

where

$$\hat{\theta}_i = \frac{\sum_M^{m=1} \hat{\theta}_{im}}{M}$$

and

$$\bar{\theta} = \frac{\sum_N^{i=1} \hat{\theta}_i}{N}$$

The within variance can be calculated as

$$\hat{\sigma}_w^2 = \frac{\sum_N^{i=1} SE(\theta_i)^2}{N}.$$

where

$$SE(\theta_i)^2 = \frac{\sum_M^{m=1} (\hat{\theta}_i - \hat{\theta}_{im})^2}{M},$$

This can be thought of as the mean of the variances of the distances from the respondents to the given item.

From a technical standpoint, this definition may be more suitable for non-negative distances. In the approach suggested by Templin and Bradshaw (2013), values of the variable are simulated from a normal distribution. However, it is possible for the simulated values (i.e. the simulated distances) to be negative, which may not make sense. In contrast, this is not possible under the ICC definition.

5.2.2.1 Number of respondents, items, and iterations

The first set of conditions involves manipulating the number of respondents, items, and the kept iterations from the MCMC. These conditions are of practical interest to investigate because the reliability of these distances may suffer when this approach is applied to a dataset from a small class or a short assessment. We would like to know the threshold at which this approach may no longer yield reliable estimates of distances. Furthermore, it may not be practical to use this approach if it requires many iterations to yield reliable estimates. Figure 5.1, calculated under the definition proposed by Templin and Bradshaw (2013), suggests that as the number of respondents and items increases, so too does the reliability. The number of kept iterations does not matter so much. As expected, more respondents and items lead to higher reliability, but one should not worry too much about the number of iterations with regards to reliability. However, one should still ensure that the chain from the MCMC method has converged. Since the number of iterations does not seem to matter, we will focus our attention on other factors.

5.2.2.2 Number of respondents, items, and degree of dependence (large-scale assessments)

The second set of conditions involves the number of respondents, items, and the degree of dependence, specifically in the context of large-scale assessments which may include many items and respondents. As described in the supplementary materials of Jeon et al. (2021), the dependence term ξ_{ji} is simulated from the normal distribution, such that certain respondents are more likely to answer certain items correctly to an extent determined by the normal distribution. Thus, the degree of dependence can be increased by increasing the mean of the normal distribution, so that certain respondents affected by ξ_{ji} are much more likely to answer their groups of items correctly. Analogously, in the latent space model, this amount of dependency can be quantified by the gamma term.

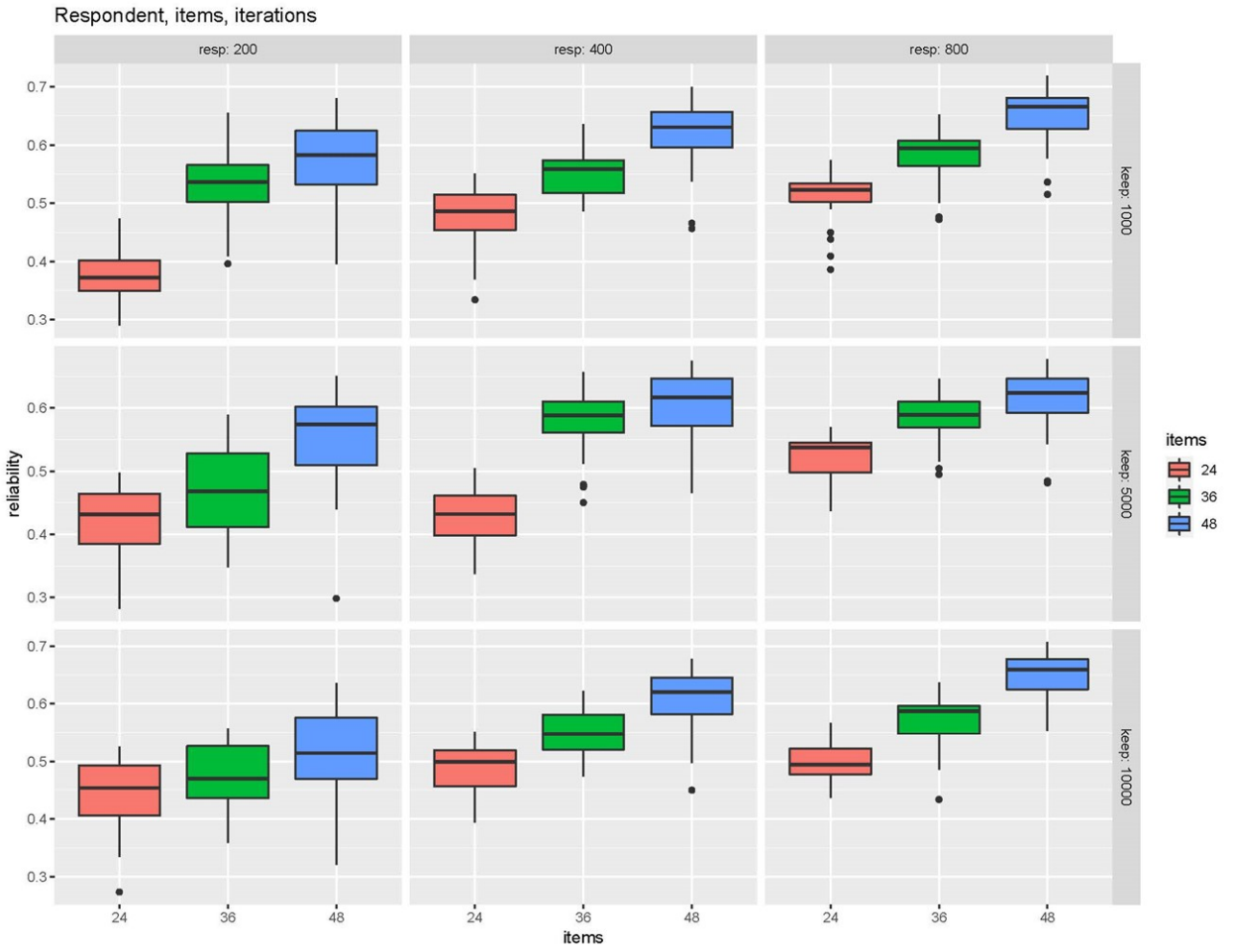


Figure 5.1: Reliability of distance measures varying by number of respondents, items, and iterations

Figure 5.2 shows the reliabilities calculated under the Templin and Bradshaw (2013) definition. The figure shows that similarly as in the first set of conditions, the more items/respondents, the higher the reliability. But the degree of dependence is also highly influential, with higher dependence leading to higher reliability. With less dependence, the latent space model is not as useful because the reliability of the distance term are much lower. With less dependence, one sees more respondents in the middle of the two clusters of items. The distance term is then not so useful. In summary, when item clusters are present, the reliability depends on how clustered those respondents are around the items. Before using this approach to identify student strengths and weak-

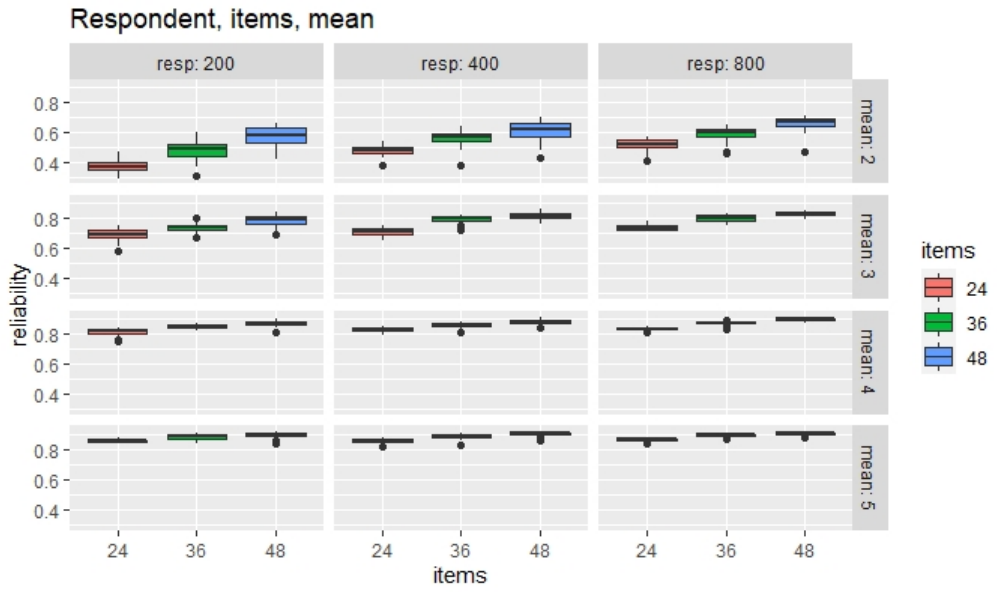


Figure 5.2: Reliability under the Templin and Bradshaw (2013) definition

nesses, it may be useful to use the model selection approach outlined in Jeon et al. (2021) first to determine whether the latent space model offers better fit due to dependencies in the data. If so, then this approach should be used to ensure that the distances are reliable.

5.2.2.3 Number of respondents, items, and degree of dependence (fewer items)

Figure 5.3, calculated under the definition proposed by Templin and Bradshaw (2013), demonstrates that the magnitude of the dependency matters much in terms of reliability. This set of conditions reduces the number of respondents (shown on the top) and the number of items (shown on the bottom) to simulate a small classroom assessment. While more items generally mean higher reliability, what really seems to increase reliability is the amount of dependency in the data. Thus, even for small classroom assessments, these profiles can be reliable if this dependency is present in the data. It is worth emphasizing again that the model selection process outlined in Jeon et al. (2021) may be undertaken first to ensure that there is enough dependency to justify the use of these profiles.

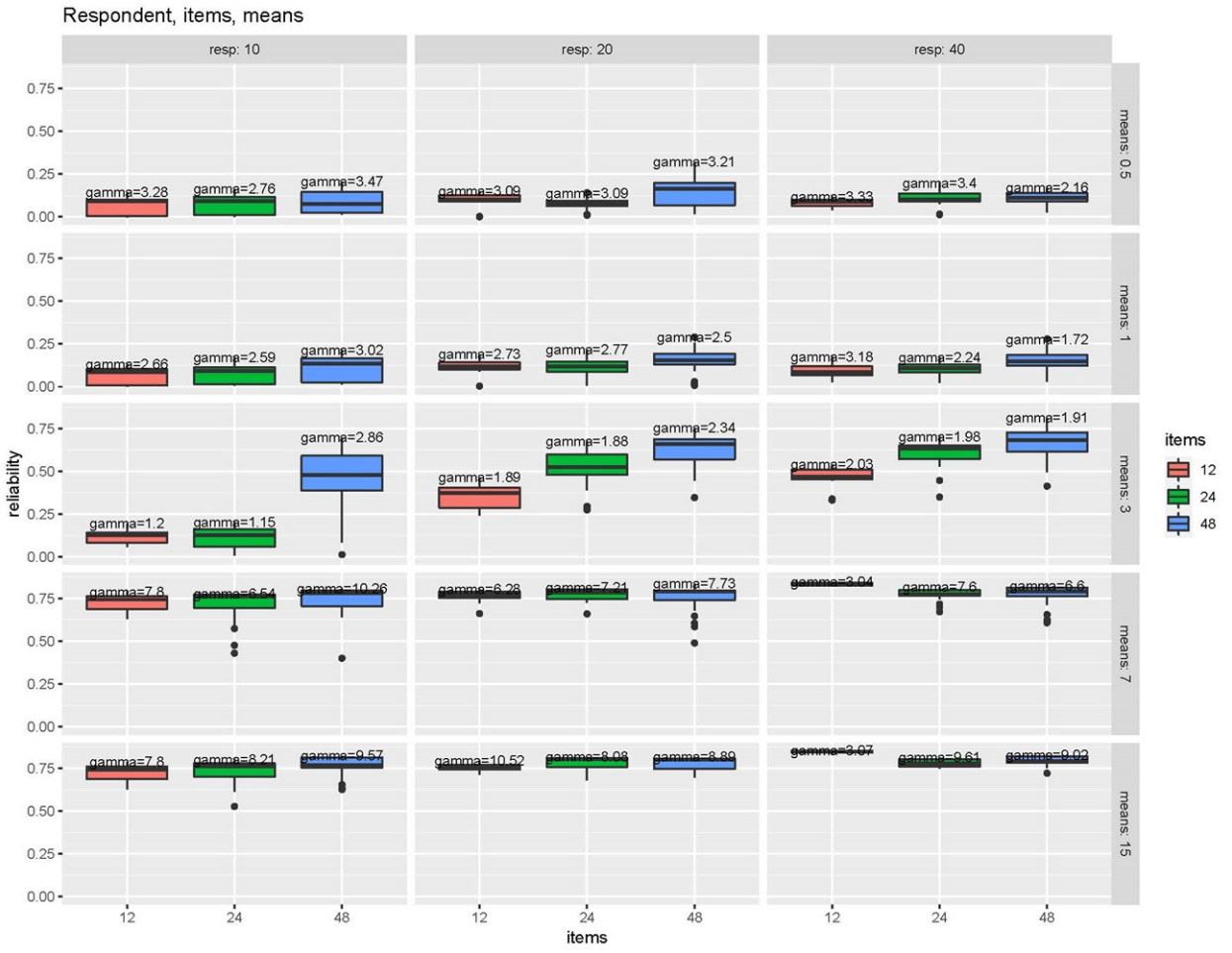


Figure 5.3: Reliability of distance measures varying by number of respondents, items, and degree of dependence along with γ estimates

The estimated reliabilities calculated under these conditions but under the ICC definition are shown in Figure 5.6. Each point in the boxplot represents a reliability calculated for a given item in the simulated data. Note that the estimated reliability are very similar to those shown in Figure 5.3, including conditions when there are many students. These results corroborate our earlier ones.

It is worth noting that in general, the reliabilities calculated under both definitions are consistent. Figure 5.7 shows the reliabilities calculated under the ICC definition for the conditions so far explored. For easier visualization, Figure 5.8 shows the same visual but without the estimates of γ.

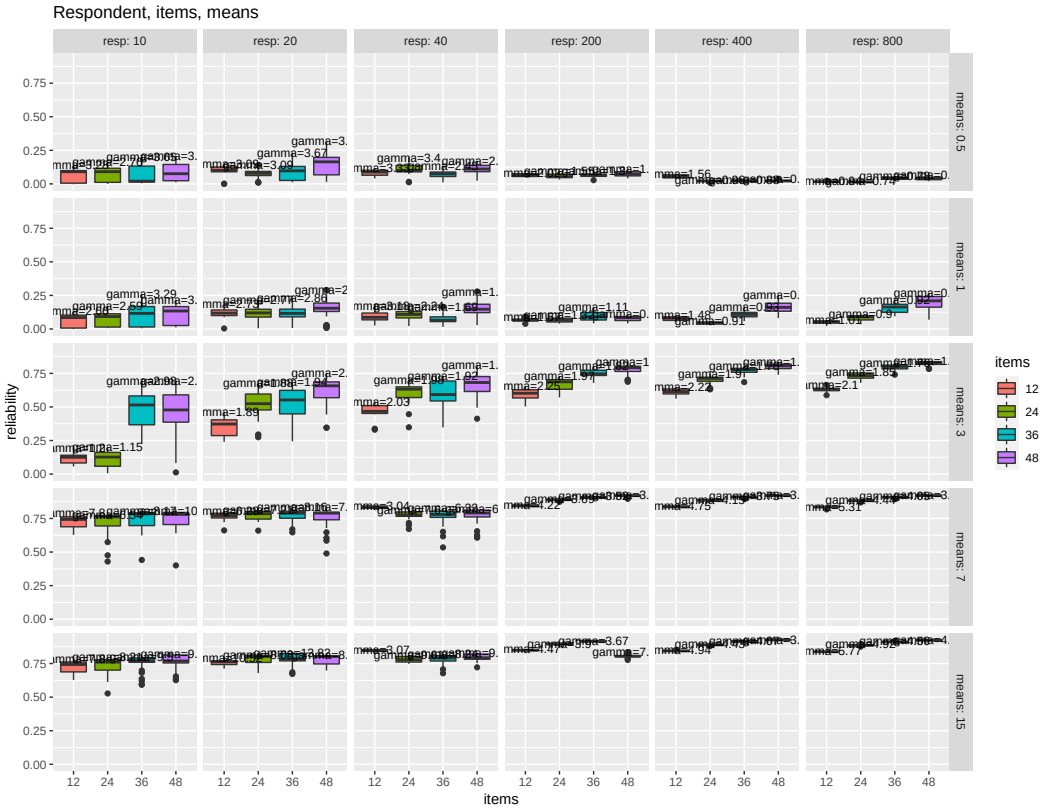


Figure 5.4: Reliability of distance measures varying by number of respondents, items, and degree of dependence along with γ estimates

5.3 Validity

Although the distances may be reliable, it is also important to understand whether these distances truly represent the strengths and weaknesses of a student with respect to the items. The assumption is that the different response patterns, as visualized on the interaction map, demonstrate some dependencies not accounted for by respondent ability and item easiness. These dependencies could stem from a combination of problematic items or from individual student strengths or weaknesses on those items (those strengths or weaknesses could stem from lack of knowledge about a domain assessed by the item). Quantitative methods alone cannot determine the source of these dependencies; additional qualitative analysis or quantitative data, such as external measures of the students' competencies on domains, would be necessary. However, it is possible

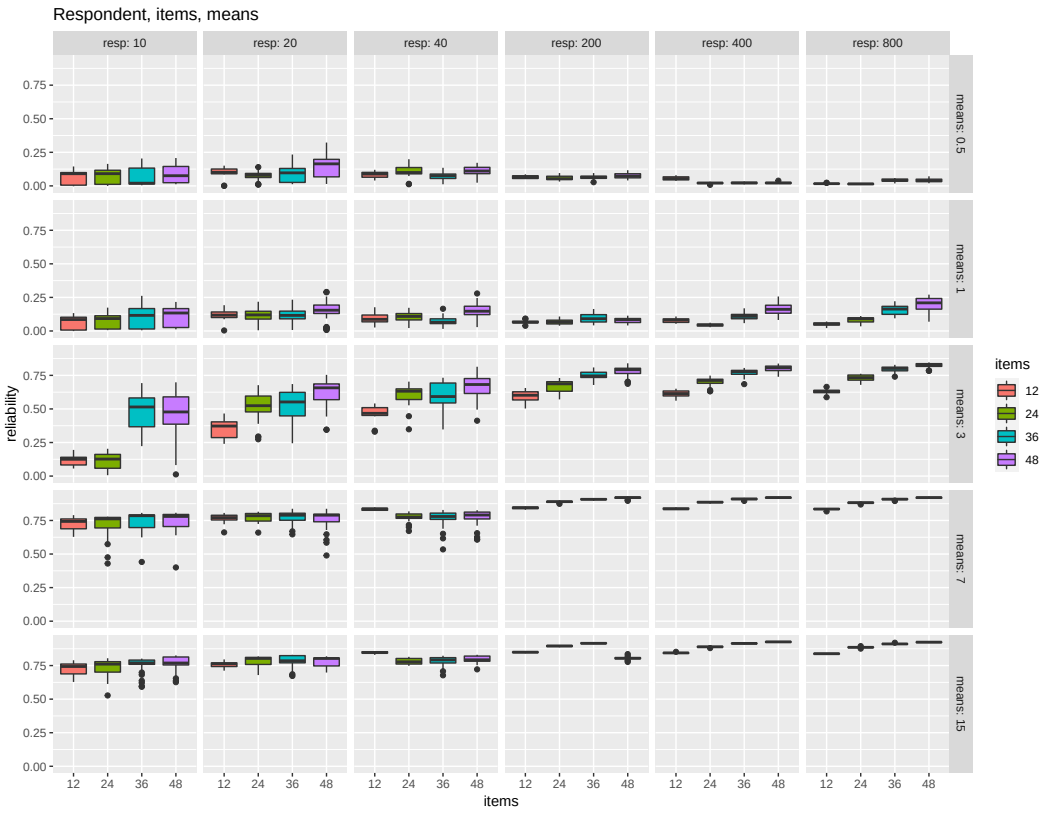


Figure 5.5: Reliability of distance measures varying by number of respondents, items, and degree of dependence along with γ estimates

to show whether the distances correspond to the dependencies found in the data, as a sort of validity check.

Recall that the data is simulated such that a local dependence term ξ_{ji} is generated for each respondent-item pair. This dependence term can represent the respondent's strength on a given item; the larger this term, the more likely the student will respond correctly. Since we expect the distance to be smaller for items that the respondent is more likely to answer correctly, we should expect a negative correlation between the dependence term ξ_{ji} and the distances. These correlations were calculated for different means of the normal distribution used to generate the dependence terms, and Figure 5.9 shows that the two terms are very strongly negatively correlated so long as the dependence is present. This shows that the distances are accurate representations of

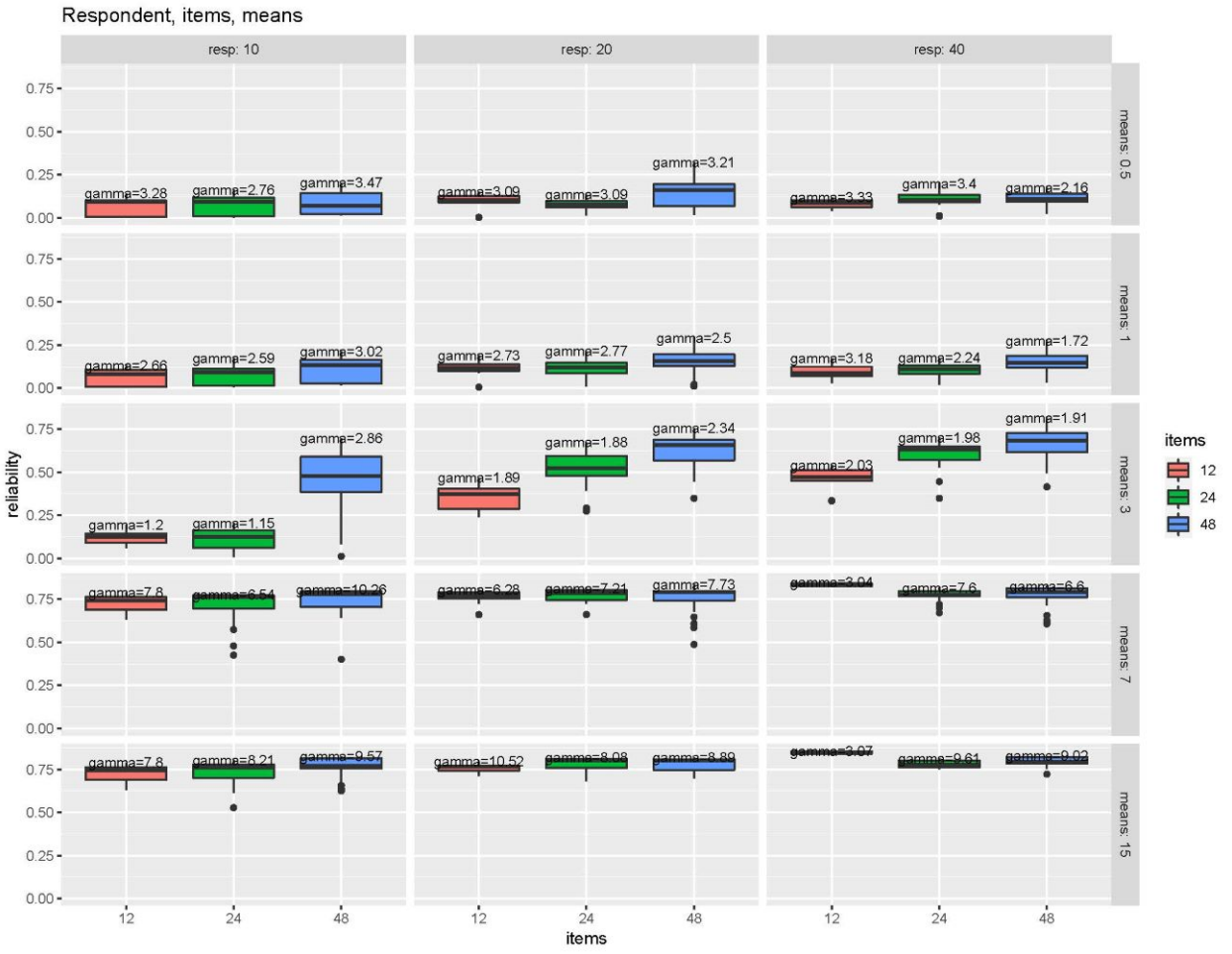


Figure 5.6: Reliabilities calculated under the ICC definition for smaller classroom assessments

the dependencies in the data.

5.4 Summary

Based on simulated data, it seems that the reliability of the distances ultimately depends on the amount of unobserved dependence in the data. Increasing the number of respondents and items only seems to improve reliability marginally. These findings are true across the reliability coefficients estimated from both the approach suggested by Templin and Bradshaw (2013) and the definition of the ICC. Therefore, it is important

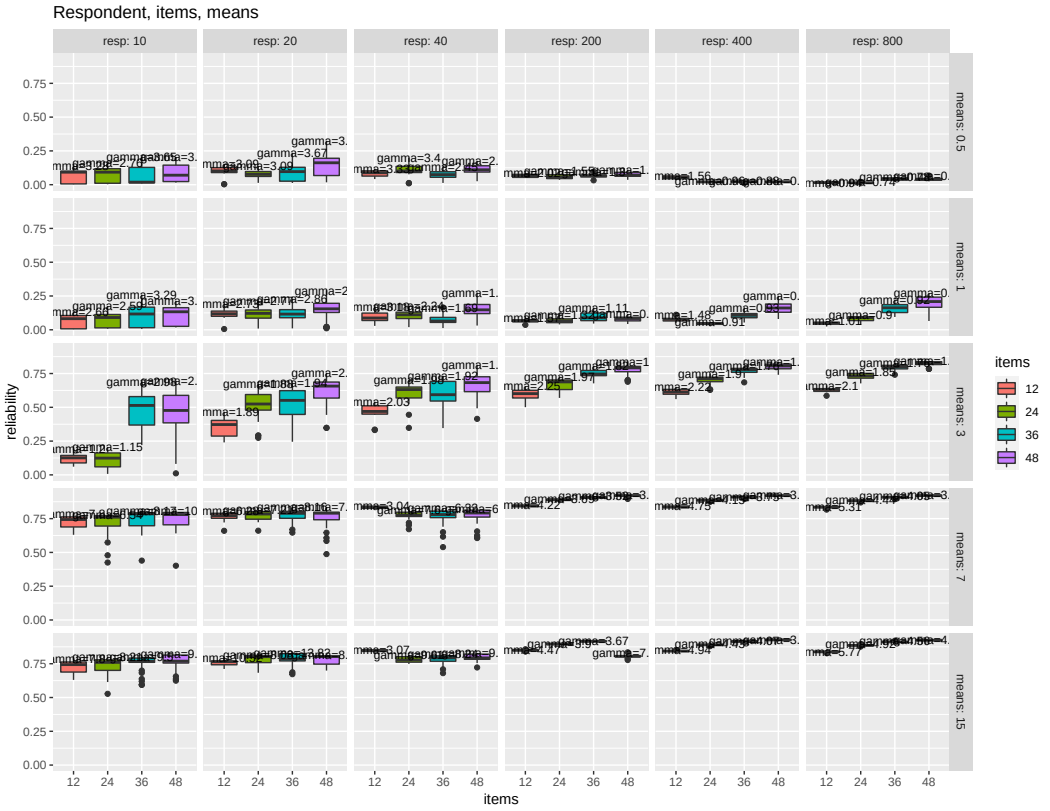


Figure 5.7: Reliability of distance measures varying by number of respondents, items, and degree of dependence along with γ estimates using the ICC definition

to consider using the model selection process suggested by Jeon et al. (2021) to first decide whether the latent space item response model is suitable for the data. If so, this suggests that there is considerable amount of unobserved dependence present in the model, thereby making the distance term necessary and more reliable for practical use. The prior investigations have shown that this approach is suitable for not only large-scale assessment data but also smaller classroom assessments, since the reliability of the distance terms depends not so much on the number of items or respondents. Finally, since validity requires reliability, establishing the reliability of these distance terms is a necessary step in establishing the validity of these measures.

Furthermore, preliminary investigations of the validity of the distance terms suggest that the distance term does indeed correlate well with the simulated dependence

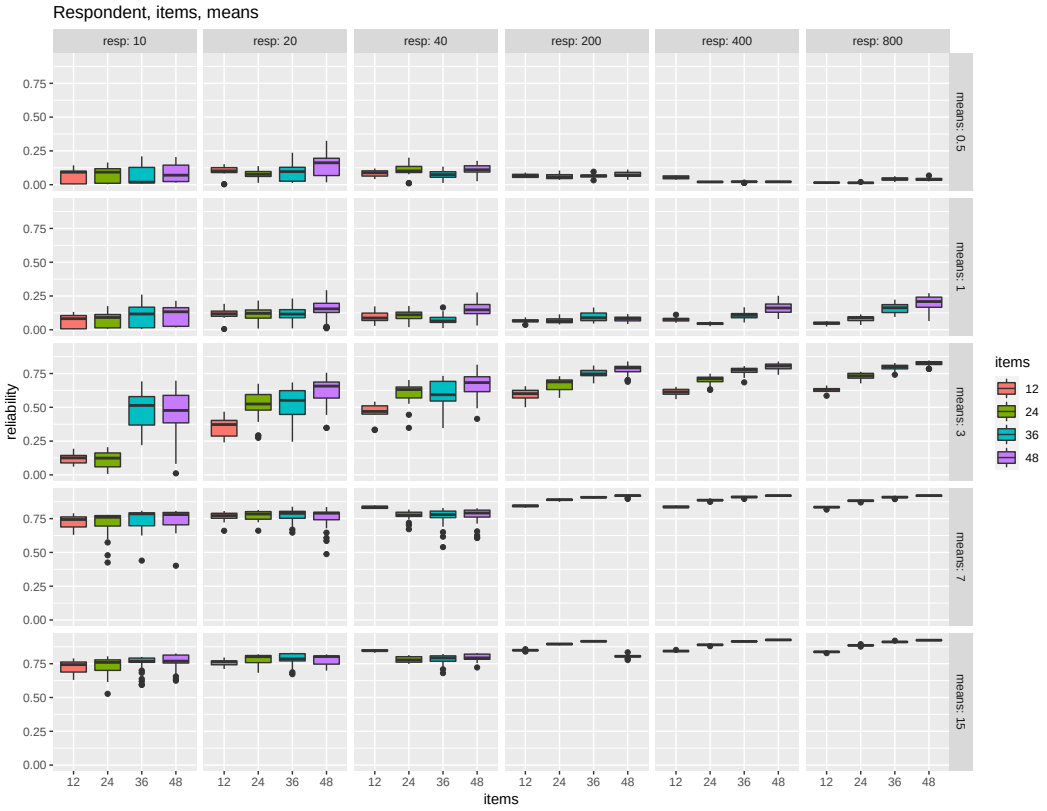


Figure 5.8: Reliability of distance measures varying by number of respondents, items, and degree of dependence along without the γ estimates using the ICC definition

terms in the simulated data. While Jeon et al. (2021) did compare the item response data with the estimated distance terms, this is a more rigorous investigation of the correspondence between the estimated distance term with the dependence to be accounted for in the data. Thus, the evidence presented in this chapter suggests that the model accurately captures unobserved dependencies in the data and can be used reliably, under certain circumstances, to create maps and profiles.

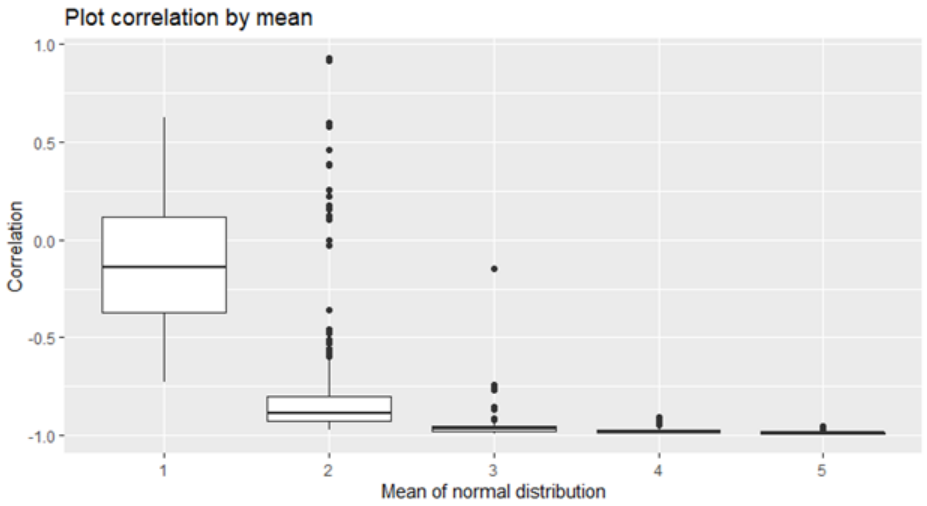


Figure 5.9: Boxplots between degree of dependence and the correlation between the degree of dependence and the pairwise distance.

CHAPTER 6

Empirical Applications and the Shiny Application

The interaction map approach provides useful item and student diagnostic information that is more easily comprehensible to practitioners through visualizations. For example, the interaction map provides a broader overview of item groupings to allow test designers to assess whether items are assessing the same concepts as desired. Users may also note certain groupings of students who are struggling with certain items. The profiles can help users conduct a deeper dive into the performance of these students. These features make more practical and illuminating the model originally presented by Jeon et al. (2021).

In this chapter, I compile the features of the interaction map approach introduced in previous chapters and demonstrate the utility of this approach with empirical applications to real-life educational assessment data. I create interaction maps and profiles from the data with accompanying interpretations and practical insights. I also present a prototype of a web-based application that can allow users to use this approach for their own data which they can upload to the application. I demonstrate the user interface and features of the application and optimization of the code that allows the application to run faster.

6.1 Empirical Application

This method will be applied to data from the CourseKata platform which was developed by researchers from UCLA and the California State University, Los Angeles (Stigler et al., 2020). CourseKata features an online textbook called "Statistics and Data

Science: A Modeling Approach" which is organized into different chapters, each of which contains review questions. Instructors from various high schools and universities can use this textbook to provide instruction to students. To reinforce their understanding of statistics concepts, students can answer multiple-choice, free-response, and coding exercises throughout the book. The process and response data from classes that use the textbook are available to select researchers on the CourseKata website.

We use the item response dataset from a class taught in the San Mateo Union High School District in Fall 2020. The interaction maps and profiles will be created for review questions at the end of chapters 2 through 5. These items cover the concepts taught in each chapter. For simplicity, if a student responded more than once to a given item, only the first attempt is used in the interaction map.

The data from the CourseKata platform is ideal for demonstration purposes for various reasons. First, the textbook uses formative assessments which this approach is designed for. Second, the textbook is divided into various chapters, so the interaction map and profiles can validate the groupings of the items by chapter. Third, students progress through the textbook sequentially; hence, it would be interesting to gauge the learning progression of students by looking at different interaction maps created at different timepoints. At each timepoint, I present a case study of two students, showing their profiles to elucidate their learning trajectories. I show useful insights that could be provided to the CourseKata users using this interaction map approach.

6.1.1 Chapter 2

Chapter 2 of the textbook is called "Understanding Data." This is largely an introductory chapter since Chapter 2 teaches students the structure of datasets (i.e. what variables are) and basic manipulation of data frames. Figure 6.1 is the interaction map created from the chapter 2 review items. The items are denoted in red, prefixed by the chapter number and suffixed by the item number.

Students 24 and 54 are highlighted in red and green respectively. An inspection

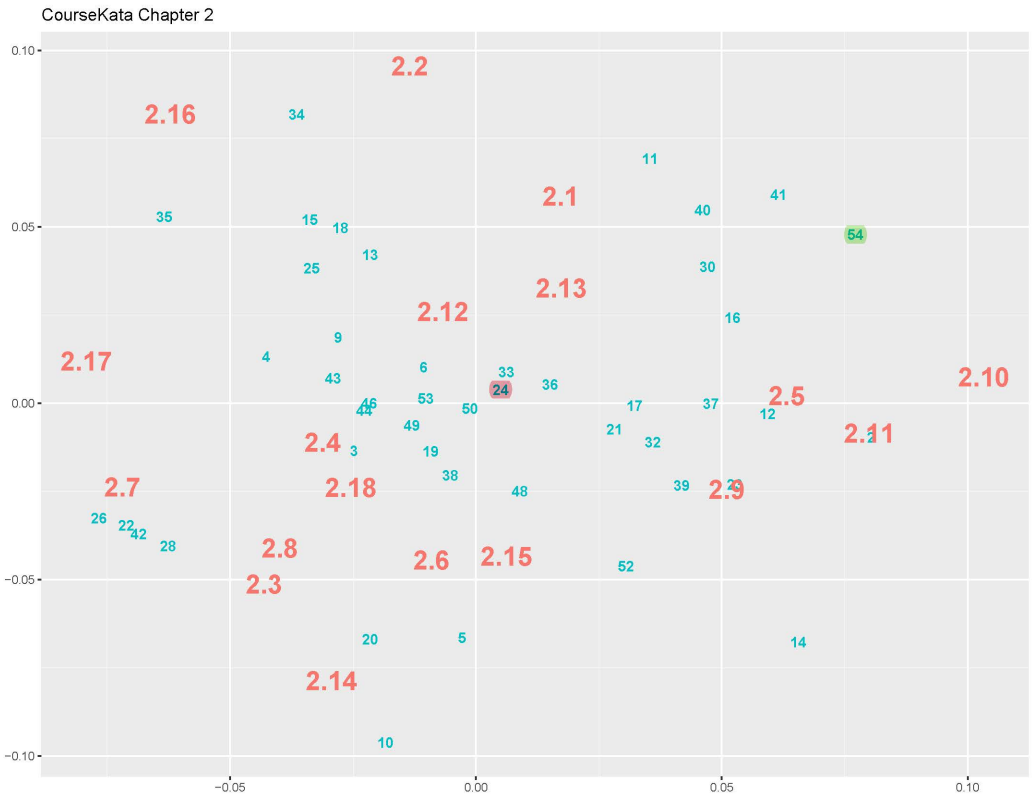


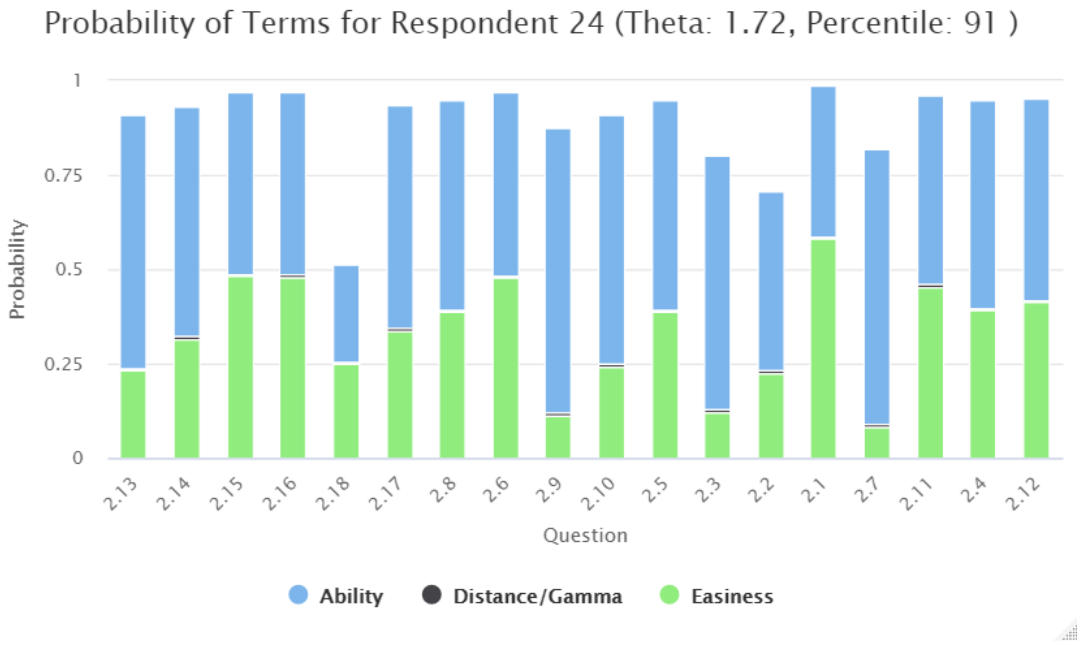
Figure 6.1: Interaction map of items from Chapter 2

of their positions suggests that student 24 is equally likely to answer Chapter 2 items correctly while this is not so much the case for student 54. The profiles in Figure 6.2 seem to corroborate this observation. In general, student 24 has high estimated correct response probabilities for all Chapter 2 items, suggesting that student 24 is generally stronger with Chapter 2 items and concepts than is student 54, whose correct response probabilities for Chapter 2 items are roughly unequal.

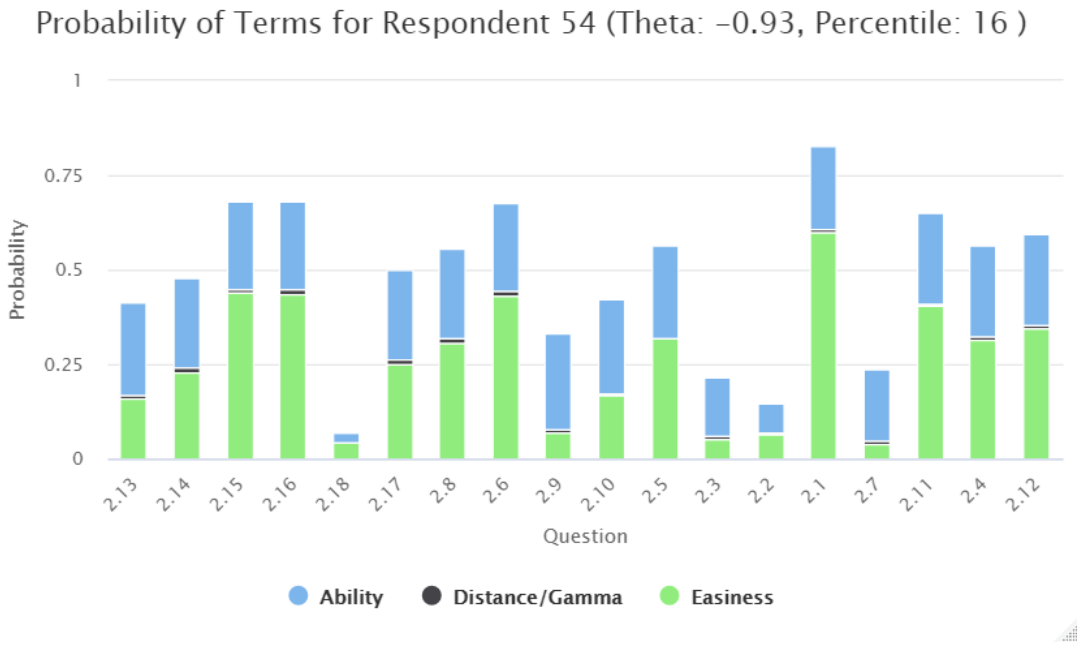
The response patterns are shown in Table 6.1.

6.1.2 Chapters 2 and 3

Now imagine that this interaction map was created later in the academic term after students have completed Chapter 3 ”Examining Distributions” items. Chapter 3 items involve finding summary statistics and visualizing the distribution of datapoints.



(a) Respondent 24 profile



(b) Respondent 54 profile

Figure 6.2: Respondents 24 and 54 profiles for Chapter 2 items

Respondent	24	54
2.1	1	1
2.2	1	0
2.3	0	0
2.4	1	0
2.5	1	1
2.6	1	0
2.7	1	0
2.8	0	0
2.9	1	0
2.10	1	1
2.11	1	1
2.12	1	0
2.13	1	1
2.14	1	0
2.15	1	0
2.16	1	1
2.17	1	0
2.18	1	0

Table 6.1: Response patterns of students 24 and 54 for Chapter 2 items

Figure 6.3 shows the interaction map with both items from chapters 2 and 3.

It is worth noting that the Chapter 2 items are on the left-hand side of the interaction map while Chapter 3 items are on the right-hand side. This suggests that the items indeed assess different concepts. However, there is some overlap. For example, items 2.16 and 3.20 and 3.3 are close to one another, suggesting that students who answer one of those items correctly are also likely to answer the other two correctly. Item 2.16 asks students "You run the following command: `RandomLakes <- sample(FloridaLakes, 10)`. What will be the result?" Item 3.20 asks students "If you were interested in proportions rather than counts, which argument would you add to your code above?" And finally, item 3.3 asks students "What would the following R code do, beyond creating a histogram?" One possible explanation is that these items all assess students' abilities to program in R. However, an item analysis might be useful in further understanding why this is the case and whether this is desirable.

In this updated latent space, it appears that while student 24 had been strong with Chapter 2 items, they are now weak on Chapter 3 items, since student 24 is closer to the Chapter 2 item cluster. Meanwhile, student 54 is between the two chapter clusters, suggesting that they are equally likely to answer both sets of items correctly. Their profiles are shown in Figure 6.4. The heights of the bars for Chapter 3 items are slightly lower than those of Chapter 2 items for respondent 24. The distance terms also play a greater role as seen by the larger black bars. Meanwhile, for student 54, the correct response probabilities are also low for Chapter 3 items as they were for Chapter 2 items. This is consistent with what we saw in the interaction maps thus far; since student 54 was weak on Chapter 2 items, and they were equally likely to answer Chapter 3 items correctly based on their interaction map position, it would make sense that they would be weak on Chapter 3 items as well.

At this point in time, the instructor might consider providing more support to both students but also note that the performance of student 24 appears to be declining.

The response patterns are shown in Table 6.2.

Respondent	24	54
3.1	0	0
3.2	0	1
3.3	0	0
3.4	1	1
3.5	0	0
3.6	0	1
3.7	0	0
3.8	0	0
3.9	0	0
3.10	0	0
3.11	0	0
3.12	0	0
3.13	0	0
3.15	1	0
3.16	0	1
3.18	0	0
3.19	1	0
3.20	0	0

Table 6.2: Response patterns of students 24 and 54 for Chapter 3 items

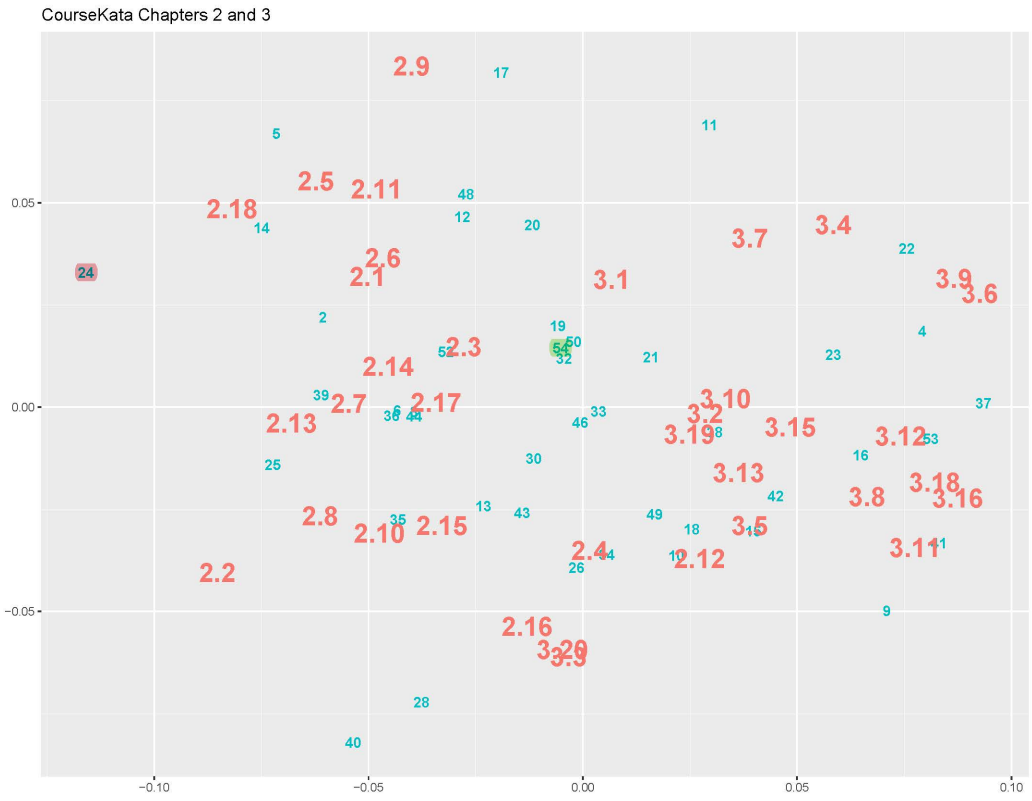
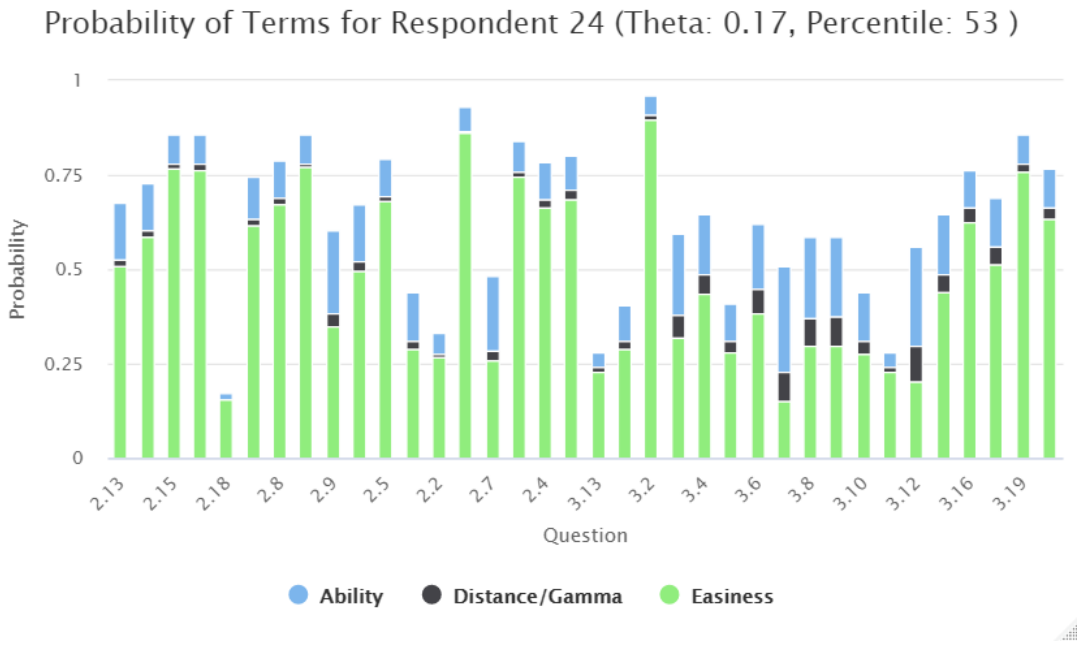


Figure 6.3: Interaction map of items from Chapters 2 and 3

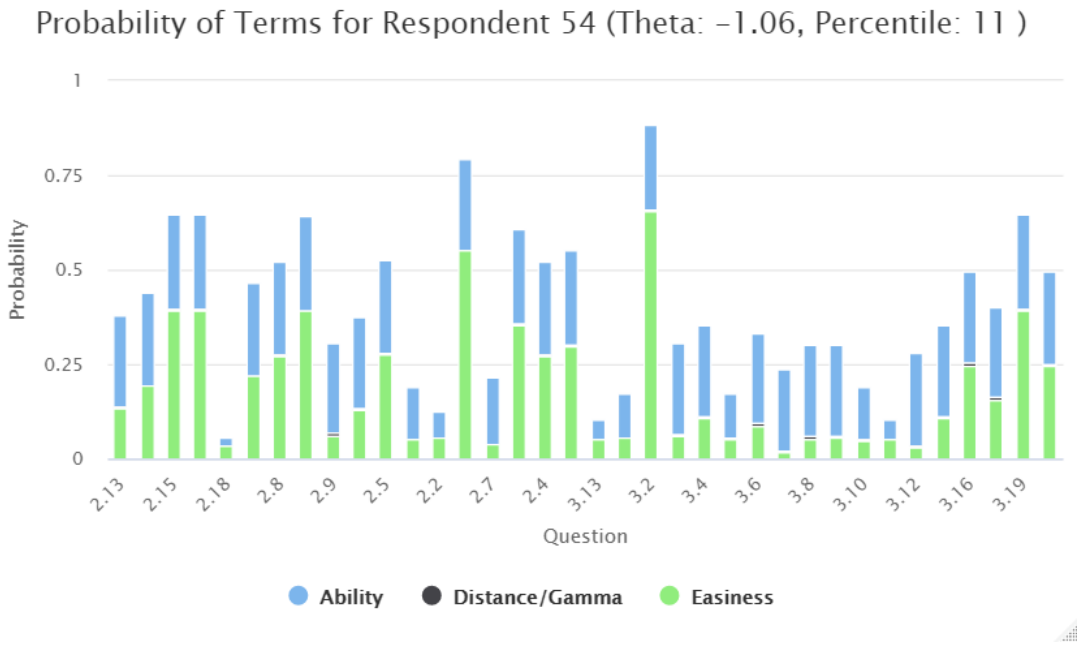
6.1.3 Chapters 2, 3, and 4

Chapter 4 is "Explaining Variation." This chapter involves looking at the relationship between variables to explain variation in one variable.

The interaction map, as seen in Figure 6.5, shows the items clustered by chapter. The Chapter 3 items are situated between the Chapter 2 and Chapter 4 items. Interestingly, item 2.2 (which asks students "If you're told that there's measurement error in how one of your variables was recorded, which of the following could be true?") is situated far from the rest of the items. This suggests that students who answer this item correctly are not necessarily also likely to answer other items correctly. This may be due to the fact that other items do not assess this concept of measurement error. Additional item analyses could probe into the reasons for this. Regardless, the designers of the CourseKata textbook might consider whether this item fits within Chapter 2 or within



(a) Respondent 24 profile



(b) Respondent 54 profile

Figure 6.4: Respondents 24 and 54 profiles for Chapters 2 and 3 items

the exercises in general. There are also items such as 4.19 and 4.20 which seem farther from the Chapter 4 cluster, which may not be surprising since those items involve the concepts of random sampling and assignment, which are not covered by the other items in Chapter 4.

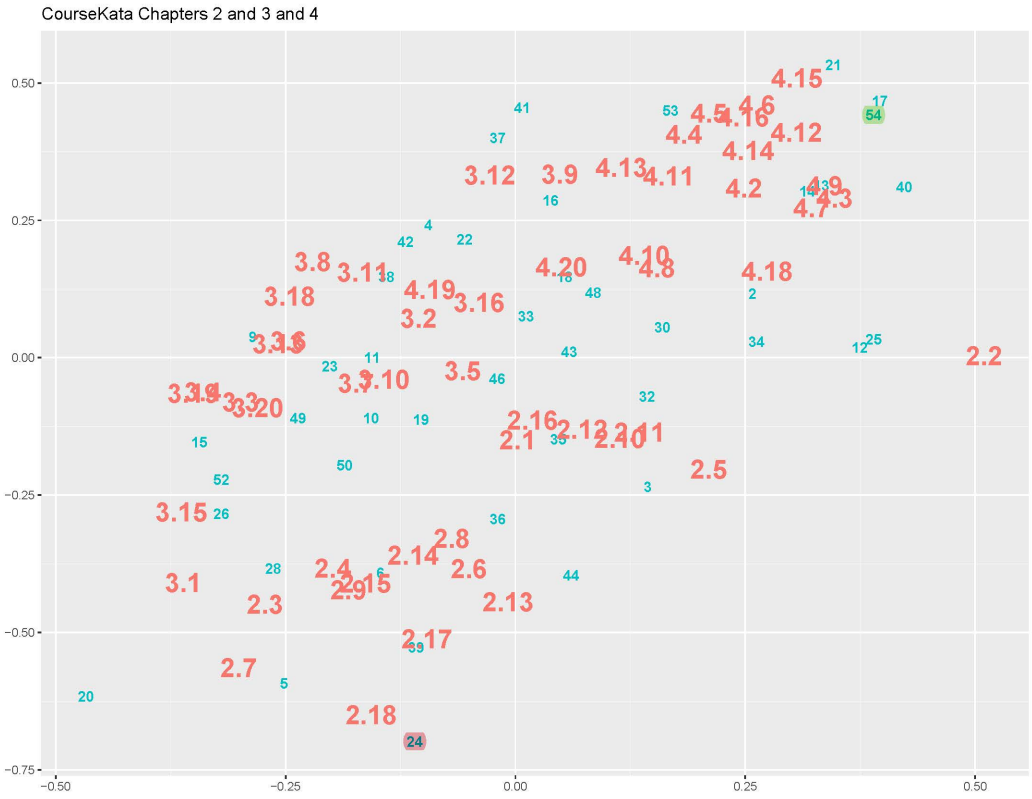
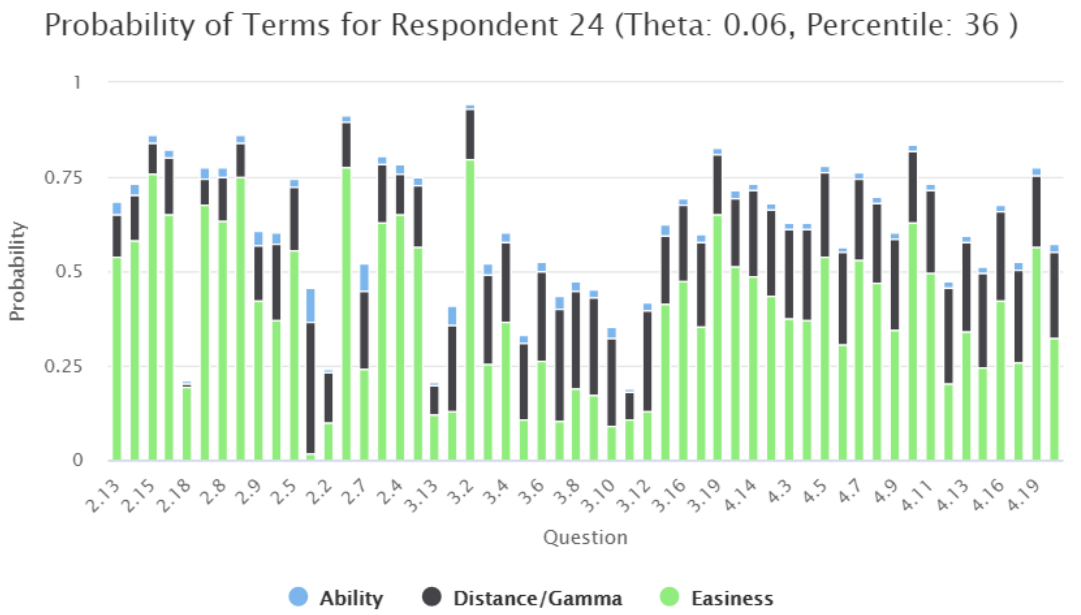


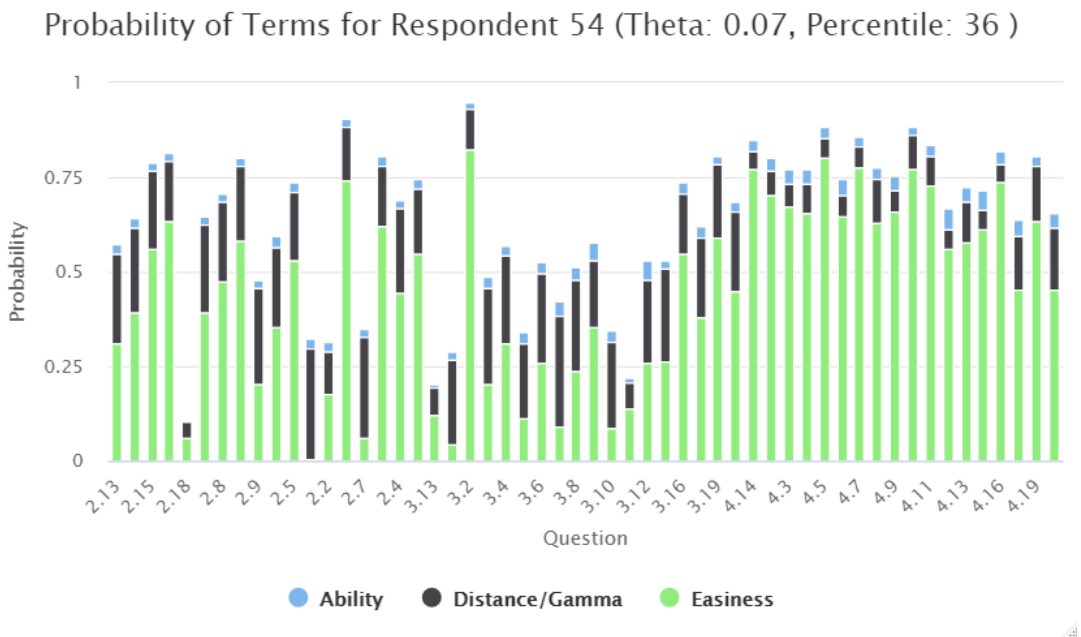
Figure 6.5: Interaction map of items from Chapters 2, 3, and 4

At this point in time, it seems that student 24 is falling behind even further since they are much farther from the Chapter 4 item cluster. In contrast, student 54 is much closer to the Chapter 4 items, suggesting that they are improving. The profiles in 6.6 seem to corroborate this observation. Student 54 has more items with higher correct response probabilities (above 75 %) than does Student 24.

The response patterns are shown in Table 6.3.



(a) Respondent 24 profile



(b) Respondent 54 profile

Figure 6.6: Respondents 24 and 54 profiles for Chapters 2 and 3 and 4 items

Respondent	24	54
4.2	1	1
4.3	0	1
4.4	0	1
4.5	0	1
4.6	0	1
4.7	1	1
4.8	0	0
4.9	0	1
4.10	0	1
4.11	0	1
4.12	0	1
4.13	0	1
4.14	0	1
4.15	0	1
4.16	0	1
4.18	0	0
4.19	1	0
4.20	0	0

Table 6.3: Response patterns of students 24 and 54 for Chapter 4 items

6.1.4 Chapters 2, 3, 4, and 5

Finally, the interaction includes Chapter 5 ”A Simple Model” which introduces students to the idea of models, such as using the mean to model variation. The interaction map in Figure 6.7 reflects again the differences among the chapter items. The Chapter 5 items form a distinct cluster on the left-hand side of the map, perhaps unsurprising since Chapter 5 items constitute the second part of the textbook called ”Modeling Variation.” However, item 2.16 is close to this cluster, suggesting that students who answer this item correctly are also likely to answer Chapter 5 items correctly. Previously, we had noted that item 2.19 was distinct from the rest of the items in Chapter 2 and that it asked students ”You run the following command: `RandomLakes <- sample(FloridaLakes, 10)`. What will be the result?” Interestingly, it seems that students who answer this item correctly are also likely to answer Chapter 5 items correctly. Again, an item analysis could uncover why this is the case.

Figure 6.8 shows the same interaction map but with the items sized and colored by their easiness. Items that are overall easier for students have brighter colors and are represented by larger bubbles. This version of the interaction map allows users to see the overall easiness of items. It seems that the Chapter 5 items, represented by the bubbles on the left-hand side, are actually generally easier than the items in the other chapters. Some of the easiest items include item 3.1 which asks students how to create a distribution of a dataset (the following item immediately shows students how to create a histogram in R) and item 2.1 which asks students what data is.

Figure 6.9 shows the profiles for both students. At this point in time, it seems that respondent 54 has drastically improved since the beginning of the term, having much higher correct response probabilities for Chapter 5 items. This can be seen in the interaction map since they are between the Chapter 4 and 5 items. In contrast, student 24 is struggling with Chapter 5 items, with lower bars and a position farther from the rest of the items.

Using this interaction map approach, the instructor can not only gain item di-

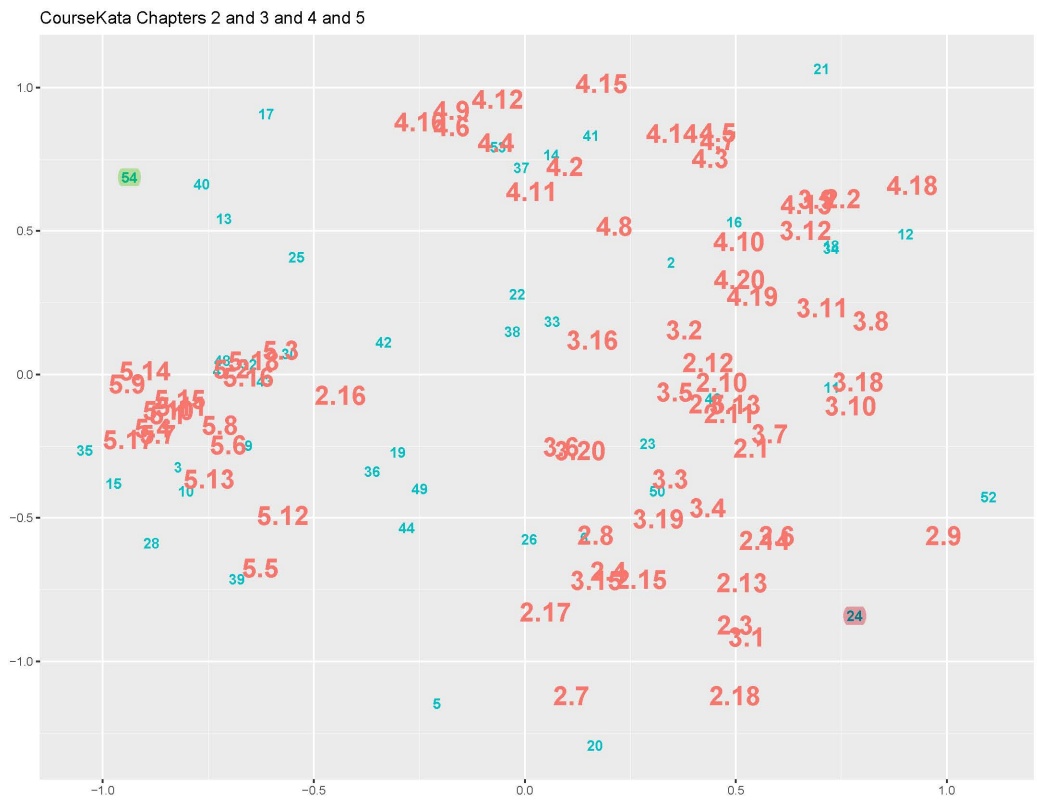


Figure 6.7: Interaction map of items from Chapters 2, 3, 4, and 5

agnostic information, as to whether items are answered similarly by respondents (to gauge whether they seem to be assessing the same concepts) but also understand student performance on an item-level. Additionally, this approach, used at different points in time, can delineate the learning trajectory of students. In this case study, the instructor is better able to see how student 24 increasingly struggled with the progression of the concepts while student 54 gradually improved.

The response patterns are shown in Table 6.4.

6.1.5 Comparison with existing approaches

In this section, I compare existing approaches with the interaction map approach to demonstrate the unique insights that the latter provides.

Respondent	24	54
5.1	0	1
5.2	0	1
5.3	0	1
5.4	0	1
5.5	1	1
5.6	0	1
5.7	0	1
5.8	0	1
5.9	0	1
5.10	0	1
5.11	0	1
5.12	0	1
5.13	0	1
5.14	1	1
5.15	0	1
5.16	0	1
5.17	0	1
5.18	1	1

Table 6.4: Response patterns of students 24 and 54 for Chapter 5 items

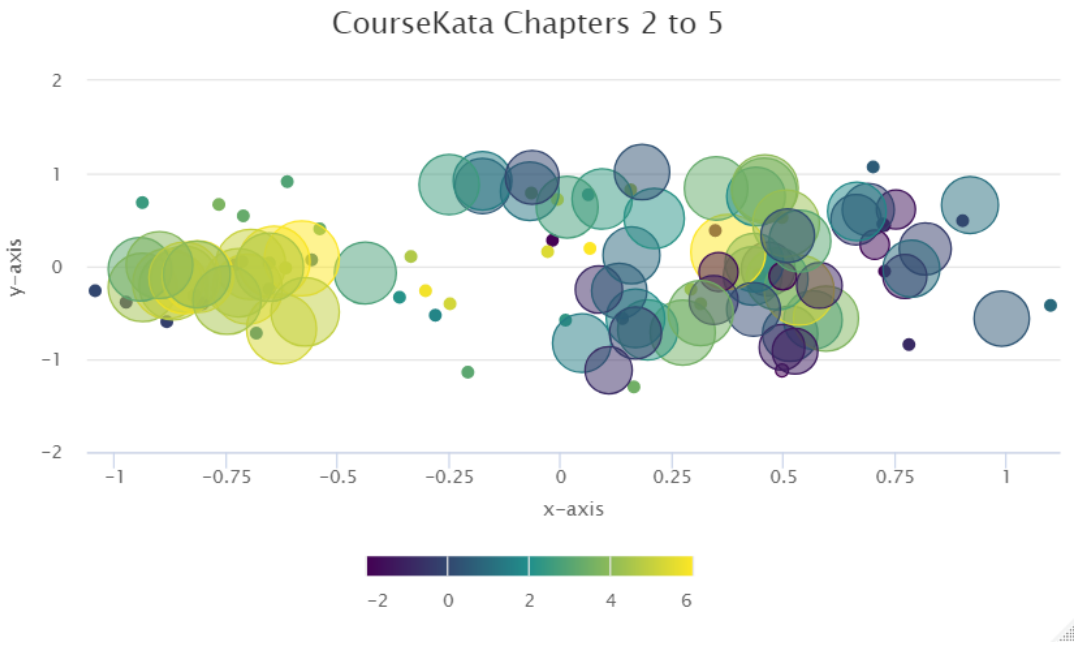
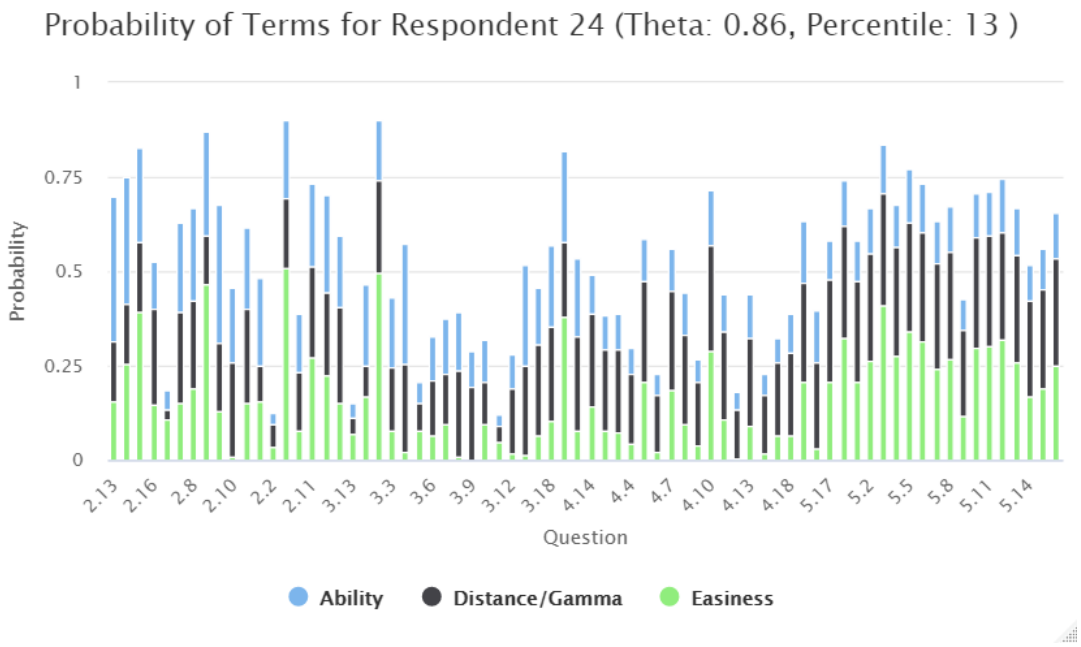


Figure 6.8: Interaction map of items from Chapters 2, 3, 4, and 5 with main effects

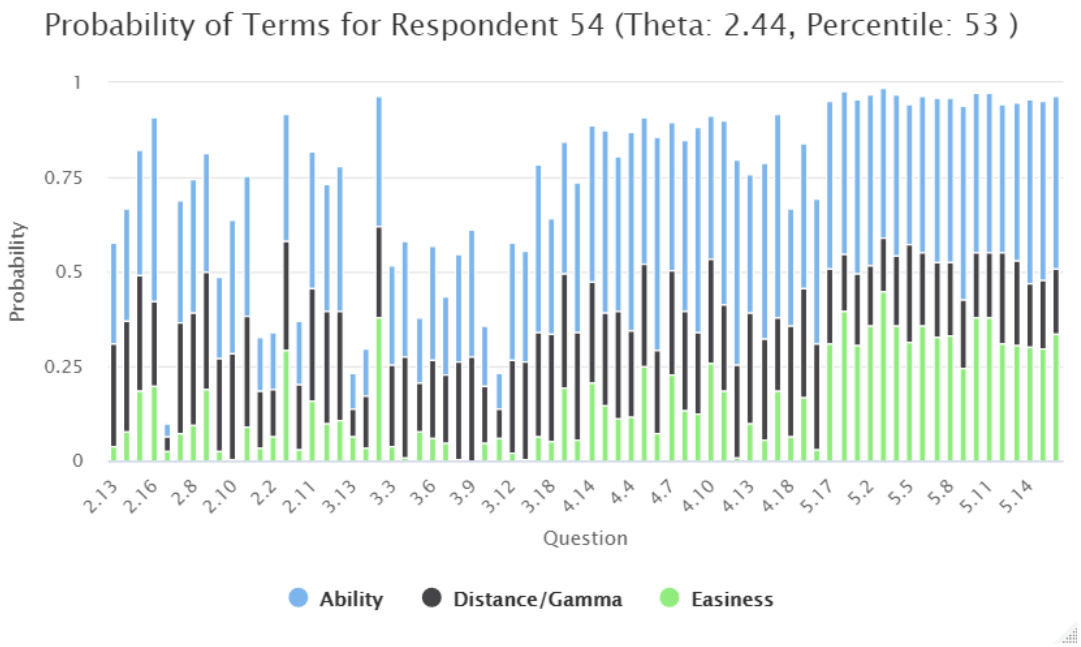
6.1.5.1 Proportion of correct responses (CTT)

One other common approach to identifying student strengths and weaknesses is to simply calculate the proportion of correct responses for each respondent for each chapter. That way, an educator could identify which chapters a student is having trouble with. This is in line with classical test theory (CTT) which evaluates the abilities of students in such a way.

The main problem with this approach is that there is a circular dependency problem (Fan, 1998). The proportion of correct responses for a respondent may offer an incomplete picture of a respondent's true ability. For instance, a respondent may have only answered easy items correctly and answered harder items incorrectly, yet this fact would not be reflected in a simple proportion. One could then attempt to compute for each item the proportion of respondents that answered it correctly to derive an estimate of item easiness. Unfortunately, this proportion may obscure the true difficulty of an item because that proportion does not account for the abilities of respondents. Accounting for these abilities requires the proportion of correct responses; therein lies the dependency.



(a) Respondent 24 profile



(b) Respondent 54 profile

Figure 6.9: Respondents 24 and 54 profiles for Chapters 2 and 3 and 4 and 5 items

In contrast, like other IRT approaches, the interaction map approach provides estimates of these respondent abilities and item easiness without these problems. This approach also provides finer-grained item-level information so that users also know which specific items a student is stronger on (information not provided by the proportion). Taken together, it is possible to know which items are overall easier or harder and to know which students are stronger or weaker on them. This cannot be done with simple proportions.

6.1.5.2 Multidimensional IRT

In contrast to the much simpler proportion method, a multidimensional IRT model could also provide these main effects from item response data. However, the multidimensional IRT model requires the dimensions to be specified a priori. Although the CourseKata designers may hypothesize about the different dimensions, we have seen some interesting and unexpected findings in our analysis of our CourseKata data which may not have been foreseen. In contrast, the interaction map approach does not require this a priori knowledge. In fact, the item groupings emerge organically in the interaction map without such prior specifications. For more complicated assessment data where the items are not grouped by chapters or the design factors are unknown, the interaction map approach may be superior. Finally, the interaction map approach provides an appealing visualization of unobserved dependencies in the data. To our knowledge, no similar visualization is possible from multidimensional IRT models.

6.2 Shiny Application

In the final section, I present a prototype of an application implementing the method using the Shiny package in R (Chang et al., 2021). The goal of this application is to allow users to try all the features introduced in the previous chapters for their dataset of choice. The application can be found at https://ohrice.shinyapps.io/LatentSpace/, a

picture of which can be seen in Figure 6.10.

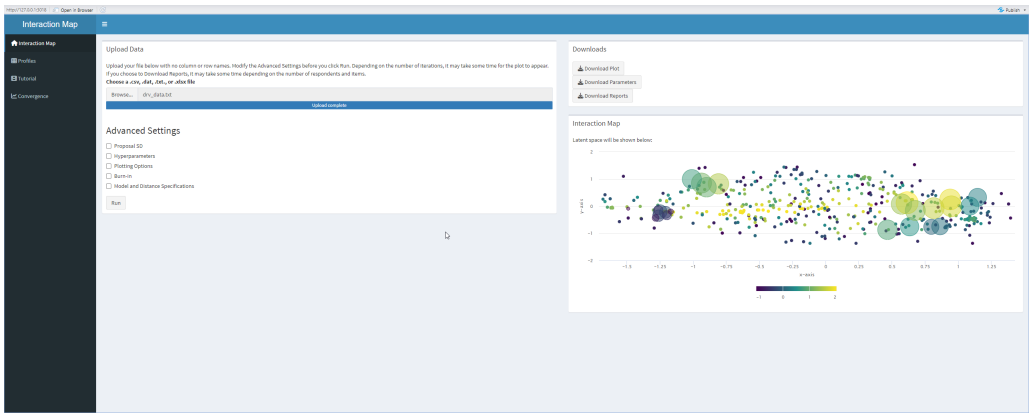


Figure 6.10: Screenshot of the Shiny application after uploading the DRV dataset

6.2.1 User interface and features

The user interface uses the shinydashboard package (Chang & Borges Ribeiro, 2021) which allows one to create attractive dashboards within the Shiny framework. One notable feature is that multiple tabs can be created for easier navigation and to prevent clutter. In the application, one tab allows users to upload their data, modify run settings (such as the number of iterations), and view the interaction map and download profiles and estimated parameters. Another tab allows the user to view individual profiles. Finally, another tab allows more advanced users to view convergence plots of the various estimated parameters.

6.2.2 Estimation

A Bayesian approach is used in the Shiny application to estimate the parameters in the latent space model using Monte Carlo Markov Chain (MCMC) methods. Prior distributions are specified for the parameters of interest. Jeon et al. (2021) specified normal priors for θ_j, β_i, and $log(\gamma)$, an inverse-gamma distribution for σ^2 (a hyperparameter representing the variance in the prior for θ_j), and multivariate normal distributions for

the respondent and item positions. One would draw posterior samples of each desired parameter and summarize them using statistics such as the posterior means. Since the exact conditional posterior distributions of these parameters are only known to be proportional to the prior distributions and the likelihood, one can use the Metropolis-Hastings algorithm with Gibbs sampling to draw the posterior samples of the desired parameters. Specifically, for a given parameter θ:

1. Draw a candidate value θ^* from a symmetric proposal distribution
2. Accept the candidate value with probability $min(1, \frac{f(\theta^*|...)}{f(\theta^t|...)})$ where $f(\theta)$ is the conditional posterior distribution at the value of θ given all the other parameters, and θ^t is the previous value of θ in the Markov chain

These steps are repeated for each parameter in a certain order (Jeon et al. suggest updating in the order θ_j, β_i, γ, $\mathbf{z_j}$, $\mathbf{w_i}$, then σ^2), with the posterior distribution conditional on the most recent set of parameters, until satisfactory convergence and mixing has been attained. This can be ascertained from trace plots and the Gelman-Rubin statistics. Otherwise, the variances of the proposal distributions (multivariate Gaussians) can be adjusted to ensure the posterior distributions have been adequately explored.

One major consideration is that the log-odds of a correct response is invariant to translations, reflections, and rotations of respondents and items. For example, in a one-dimensional example, the log-odds of a correct response for a respondent at position 1 and item at position 5 would be the same as the log-odds of a correct response for the same respondent at position 4 and item at position 8 because the log-odds depends on the distance between the respondent and item (which is 4 in this case). A given configuration of the latent space could be transformed in infinitely many ways to yield the same log-odds (Hoff et al., 2002). Therefore, to ensure identifiability, Procrustes matching (Gower, 1975) must be used to post-process the MCMC samples. Specifically, Procrustes matching is used with a reference set of positions that attained the highest

log posterior density; the posterior samples are then transformed with respect to this reference set.

6.2.3 Optimization

The original code used to implement the interaction map approach in Jeon et al. (2021) is slow. For example, running 20,000 iterations with a burn-in period of 10,000 iterations takes about 40 minutes with the original code running on the DRV dataset (with 24 items and 418 respondents). Using the R profiler and the microbenchmark package (Mersmann, 2021), I managed to optimize the code in various ways.

- In the calculation of the likelihood function, I replaced `apply()` functions with their `rowSums()` or `colSums()` equivalents, which are faster.
- To create the pairwise bipartite distance matrix, instead of using the `pdist()` function which required lengthy computations according to the profiler, I wrote a new function using the Rcpp package (Eddelbuettel & François, 2011), which allows C++ code to be integrated into the R code.
- Most importantly, I converted the input data from a data frame to a matrix, which considerably requires less memory and is more efficient.

These fixes alone reduced the runtime from 40 minutes to about 5-7 minutes, a considerable improvement which greatly improves the practicality of this Shiny application.

6.3 Summary

The application of the interaction map approach to the CourseKata dataset revealed many useful features of this approach. Firstly, this approach can be used to yield item diagnostic information, showing whether certain items may not fit with other

items in the chapter. The overall easiness of items can also be visualized. Secondly, this approach can yield diagnostic information about students. Their performances can be compared and visualized using profiles. Additionally, by comparing the interaction maps across different timepoints, it is also possible not only to see a cross-sectional view of their performances but also a longitudinal view which charts their educational trajectory, thereby allowing instructors to potentially intervene for students whose performances may diminish over time. Further extensions of this approach might consider visualizing the changes in positions of the students in the latent space with regards to the items across various timepoints. Such an approach could better elucidate the learning progression of students.

Finally, the Shiny application makes the aforementioned extensions of the interaction map and profiles accessible to a wide variety of users, including teachers, test designers, and other educational practitioners. Future studies could involve user interface (UI) or user experience (UX) testing to gauge the usability of this tool. Usability studies could help illuminate which features users would find the most helpful to incorporate into this tool.

CHAPTER 7

Conclusion

In this dissertation, I have presented various practical extensions of the latent space item response model. These extensions include improved visualizations of the interaction maps and the student profiles. These extensions also involve modifications to the model which can influence the types of maps and profiles generated. Together, these modifications and visualizations have accompanying interpretations that can help educators better understand their students' performances and tailor their instruction to meet their needs. Additionally, I have shown the conditions under which the model may provide reliable results as well as evidence that the model does indeed capture unobserved dependencies in the data. Finally, I demonstrated the utility of this approach with empirical applications to a real-life dataset along with a web-based application that allows users to try these features for themselves.

There are many other extensions possible that can improve the practicality of this model. One future area of research involves evaluating how the model could accommodate continuous or ordinal data. So far, the model has only been used for binary responses (e.g. correct or incorrect data). To extend the model for more practical usage, the model would ideally be able to handle not only binary data but also continuous data, such as response times on assessments. The increased use of online assessments also provides larger quantities of process data. For example, the CourseKata platform collects information on the response time of students. An extension of the model would allow researchers to visualize unobserved dependencies in the data. It is possible, for instance, that certain groups of students are spending more time on certain items or chapters. This is useful formative information that CourseKata instructors might want

to investigate to understand why those students are spending more time than their peers.

Another area of research includes extending the model to track the changes in the latent position of respondents and items to better capture learning progression. Although this idea was explored briefly in the empirical application, it can be cumbersome to create multiple interaction maps and manually compare changes in the positions of respondents. Further work could explore better visualizations and modeling of this progression.

As big data becomes increasingly more prevalent in the work of educational researchers and practitioners, it is imperative that they have the tools to make sense of this deluge of data to make the best possible decisions for their clients or students. Visualizations are one such user-friendly way to make better sense of this data. It is my hope that these extensions of the latent space item response model, as manifested in the Shiny application, may assist these researchers and practitioners. Whether the goal is to design better assessments or to help foster personalized learning, the proposed interaction map approach has great potential to improve the ability of educators to support their students' learning.

REFERENCES

Behrens, J. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Bowers, A. J., & Krumm, A. E. (2021). Supporting the initial work of evidence-based improvement cycles through a data-intensive partnership. *Information and Learning Sciences*, *122*(9/10), 629–650. https://doi.org/10.1108/ILS-09-2020-0212

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research, and Evaluation*, *15*(7). https://doi.org/10.7275/r4zq-9c31

Bowers, A. J., Shoho, A. R., & Barnett, B. G. (2014). Considering the use of data by school leaders for decision making: An introduction. In A. J. Bowers, A. R. Shoho, & B. G. Barnett (Eds.), *Using data in schools to inform leadership and decision making* (pp. 1–16). Information Age Publishing Inc. https://doi.org/10.7916/D8862F32

Broderick, A., Mehta-Parekh, H., & Reid, D. K. (2005). Differentiating instruction for disabled students in inclusive classrooms. *Theory Into Practice*, *44*(3), 194–202. https://doi.org/10.1207/s15430421tip4403_3

Ceres, P. (2020). A 'Covid slide' could widen the digital divide for students. *Wired*. https://www.wired.com/story/schools-digital-divide-remote-learning/

Chang, W., & Borges Ribeiro, B. (2021). *Shinydashboard: Create dashboards with 'shiny'* [R package version 0.7.2]. https://CRAN.R-project.org/package=shinydashboard

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for R* [R package version 1.6.0]. https://CRAN.R-project.org/package=shiny

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika*, *83*(1), 89–108. https://doi.org/10.1007/s11336-017-9579-4

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. https://doi.org/10.1080/01621459.2014.934827

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. https://doi.org/10.1177/0146621613488436

Data Quality Campaign. (2019). Making data work for personalized learning: Lessons learned. https://files.eric.ed.gov/fulltext/ED607216.pdf

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183. https://doi.org/10.1177/0146621608320523

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8–26. https://doi.org/10.1177/0146621610377081

Dorn, E., Hancock, B., Sarakatsannis, J., & Viruleg, E. (2020). COVID-19 and student learning in the United States: The hurt could last a lifetime. *McKinsey and Company*. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-and-student-learning-in-the-united-states-the-hurt-could-last-a-lifetime

Draney, K., & Wilson, M. (2007). Application of the saltus model to stagelike data: Some applications and current developments. *Multivariate and mixture distribution rasch models: Extensions and applications (statistics for social and behavioral sciences)* (pp. 119–130). Springer.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Gelman, A., & Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? https://arxiv.org/pdf/2012.00174.pdf

Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, *40*, 33–51. https://doi.org/10.1007/BF02291478

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Halim, Z., & Muhammad, T. (2017). Quantifying and optimizing visualization: An evolutionary computing-based approach. *Information Sciences*, *385-386*, 284–313. https://doi.org/10.1016/j.ins.2016.12.035

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2-3), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Hoover, N. R., & Abrams, L. M. (2013). Teachers' instructional use of summative student assessment data. *Applied Measurement in Education*, *26*(3), 219–231. https://doi.org/10.1080/08957347.2013.793187

Jennings, A. S., & Jennings, A. B. (2020). Comprehensive and superficial data users: A convergent mixed methods study of teachers' practice of interim assessment data use. *Teachers College Record*, *122*(12), 1–46. https://www.tcrecord.org/Content.asp?ContentId=23503

Jeon, M., Jin, I. H., Schweinberger, M., & Baugh, S. (2021). Mapping unobserved item–respondent interactions: A latent space item response model with interaction map. *Psychometrika*, *86*(2), 378–403. https://doi.org/10.1007/s11336-021-09762-5

Jin, I. H., & Jeon, M. (2019). A doubly latent space joint model for local item and person dependence in the analysis of item response data. *Psychometrika*, *84*(1), 236–260. https://doi.org/10.1007/s11336-018-9630-0

Kang, S., & Bowers, A. J. (2021). NSF Education Data Analytics Collaborative Workshop: How educators and data scientists meet and create data visualizations. In A. J. Bowers (Ed.), *Data visualization, dashboards, and evidence use in schools: Data collaborative workshop perspectives of educators, researchers, and data scientists* (pp. 68–84). Teachers College, Columbia University. https://doi.org/10.7916/d8-jj2g-e225

Kunst, J. (2022). *highcharter: A wrapper for the 'highcharts' library* [R package version 0.9.4]. https://CRAN.R-project.org/package=highcharter

McCarthy, E. M., Liu, Y., & Schauer, K. L. (2020). Strengths-based blended personalized learning: An impact study using virtual comparison group. *Journal of Research on Technology in Education*, *52*(3), 353–370. https://doi.org/10.1080/15391523.2020.1716202

Mersmann, O. (2021). *Microbenchmark: Accurate timing functions* [R package version 1.4.9]. https://CRAN.R-project.org/package=microbenchmark

Pane, J., Steiner, E., Baird, M., Hamilton, L., & Pane, J. (2017). *Informing progress: Insights on personalized learning implementation and effects*. RAND Corporation. https://doi.org/10.7249/RR2042

Park, J. Y., Cornillie, F., van der Maas, H. L. J., & Van Den Noortgate, W. (2019). A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in Psychology*, *10*(620). https://doi.org/10.3389/fpsyg.2019.00620

Pastor, D. A., Barron, K. E., Miller, B., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8–47. https://doi.org/10.1016/j.cedpsych.2006.10.003

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine*, *4*, 321–334.

Reeves, T. D., Wei, D., & Hamilton, V. (2021). In-service teacher access to and use of non-academic data for decision making. *The Educational Forum*. https://doi.org/10.1080/00131725.2020.1869358

Roberts-Mahoney, H., Means, A. J., & Garrison, M. J. (2016). Netflixing human capital development: Personalized learning technology and the corporatization of k-12 education. *Journal of Education Policy*, *31*(4), 405–420. https://doi.org/10.1080/02680939.2015.1132774

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., & Schmidt, J. A. (2018). tidyLPA: An R package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *Journal of Open Source Software*, *3*(30), 978. https://doi.org/10.21105/joss.00978

Schwendimann, B. A., Rodríguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., Gillet, D., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. https://doi.org/10.1109/TLT.2016.2599522

Sedrakyan, G., Mannens, E., & Verbert, K. (2019). Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *Journal of Computer Languages*, *50*, 19–38. https://doi.org/10.1016/j.jvlc.2018.11.002

Selwyn, N., Pangrazio, L., & Cumbo, B. (2021). Attending to data: Exploring the use of attendance data within the datafied school. *Research in Education*, *109*(1), 72–89. https://doi.org/10.1177/0034523720984200

Sievert, C. (2020). *Interactive web-based data visualization with r, plotly, and shiny*. Chapman; Hall/CRC. https://plotly-r.com

Spiel, C., & Glück, J. (2008). A model-based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 45–65). Hogrefe Huber Publishers. https://doi.org/10.7916/d8-jj2g-e225

Spiel, C., Glück, J., & Gössler, H. (2001). Stability and change of unidimensionality: The sample case of deductive reasoning. *Journal of Adolescent Research*, *16*(2), 150–168. https://doi.org/10.1177/0743558401162003

Stigler, J. W., Son, J. Y., Givvin, K. B., Blake, A. B., Fries, L., Shaw, S. T., & Tucker, M. C. (2020). The Better Book approach for education research and development. *Teachers College Record*, *122*(9), 1–32. https://doi.org/10.1177/016146812012200913

Tang, X., Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 108–135. https://doi.org/10.1111/bmsp.12144

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275. https://doi.org/10.1007/s00357-013-9129-4

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67. https://www.jstor.org/stable/2237638

U.S. Department of Education, Office of Educational Technology. (2017). Reimagining the role of technology in education: 2017 National Education Technology Plan update. https://tech.ed.gov/files/2017/01/NETP17.pdf

Wilkerson, S. B., Klute, M., Peery, B., & Liu, J. (2021). How Nebraska teachers use and perceive summative, interim, and formative data. http://ies.ed.gov/ncee/edlabs

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276–289. https://doi.org/10.1037/0033-2909.105.2.276

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research*, *8*, 1–22. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.514.9188&rep=rep1&type=pdf

Wilson, M., & Lehrer, R. (2021). Improving learning: Using a learning progression to coordinate instruction and assessment. *Frontiers in Education*, *6*, 276–289. https://doi.org/10.3389/feduc.2021.654212