Boston College
Lynch School of Education and Human Development

Department of
Measurement, Evaluation, Statistics, and Assessment

ACCOUNTING FOR INTERSECTIONAL SOCIAL IDENTITIES:

EXPLORING THE STATISTICAL CONSTRAINTS OF MODELS

Dissertation
by

OLIVIA SZENDEY

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

May 2024

ACCOUNTING FOR INTERSECTIONAL SOCIAL IDENTITIES: EXPLORING THE STATISTICAL CONSTRAINTS OF MODELS

Olivia Szendey, Author

Michael Russell, Chair

Abstract

Intersectionality theory garners increased attention from researchers interested in understanding the many ways in which oppression impacts lived experiences. In any given present and evolving context, oppression leads to advantages for some social positions and disadvantages for others (Collins & Bilge, 2016; Crenshaw, 1989). Quantitative researchers have attempted to adapt statistical modeling methods to reflect intersectional identities as a proxy for oppression and advantage in their models (Bauer et al., 2021; Schudde, 2018). This dissertation expanded on existing knowledge about the statistical limitations of three methods of modeling intersectional analyses on a continuous outcome variable: 1) Interaction, 2) Categorical, and 3) MAIDHA (multilevel analysis of individual heterogeneity and individual accuracy).

Using a Monte Carlo simulation, four demographic data characteristics were manipulated to explore the three models under different scenarios which manipulated: a) the number of demographic categories (and thus intersections); b) the proportion of the sample represented by each demographic group; c) the within-intersectional-group variance in the outcome variable of interest; d) overall sample size. Each scenario and model were replicated 1000 times; results summarized performance of the intersection estimates and effect detection using the outcomes: bias, accuracy, power, type 1 error, and confidence interval coverage.

The fundamental questions that guided this dissertation were:

1. What are the statistical advantages and disadvantages of each model under different demographic data characteristics?
2. In what ways does each model perform differently from one another under each demographic data characteristic condition?

The findings of this dissertation contribute to intersectional quantitative research methods by providing greater insight into how each model performs under more complex data scenarios.

Acknowledgements

I have been fortunate enough to be supported and guided by an exceptional group of individuals whose contributions to both this research and my own personal development have been invaluable. I am deeply grateful for their collective wisdom, encouragement, and support they have provided.

To my dissertation committee-- Mike Russell, Andrés Castro Samayoa, and Shaun Dougherty-- your expertise, critical feedback, and unwavering support have been instrumental in shaping this research. To Mike Russell, my dissertation chair, I owe a special debt of gratitude. Your commitment to fostering an environment conducive to critical quantitative research has profoundly influenced my scholarly pursuits. Your persistent encouragement to question established methodologies and pursue research with intentionality has enhanced my skills as both critical thinker, researcher, and writer. Laura O'Dwyer, your guidance during the early stages of this process was invaluable; the laughter we all shared in dissertation seminars remains among my fondest memories of the entire program. Larry Ludlow, your support and mentorship were pivotal in navigating this journey. There was never a dull moment when you were in 336, infusing fun and levity into the times I needed it most. Melissa McTernan, without your thought-partnership and insightful problem-solving in coding the complexities of my research goals in R, this study would not have been possible.

To the MESA community, your spirit of camaraderie and support has been a source of constant encouragement. It has been a privilege to be part of a group that values community over competition. I will forever be each of your biggest fan. Haylea and Clara, my journey through this program would have been markedly different without your friendship. From uber pools to emergency ice cream, you have been the best cohort I could have wished for. Katherine, I am

thankful for your mentorship and friendship. Noman, I am grateful for your support throughout my academic journey and now as colleague at the beginning of my professional career.

To my adventure partners, the moments spent outdoors-- from mornings at the stadium to weekends in the mountains-- have been essential in maintaining my balance and well-being throughout this journey. Aidan, your encouragement to find equilibrium between academia and leisure right from the start has been a guiding principle for me. To my Seattle friends, moving cross-country while finishing my dissertation was daunting, but your welcoming community made the final stretch of this journey far more manageable.

To my family, being able to pursue graduate studies so close to home, especially during a pandemic, was a blessing. I am grateful for your unwavering support and willingness to engage in proofreading my drafts after nearly a decade off from the role of homework editors. Thomas, thank you for your unwavering support. Your partnership and constant reminder to pursue a fulfilling life have been a source of strength and inspiration.

Table of Contents

Tables

Figures

Chapter 1: Introduction

Intersectionality theory garners increased attention from researchers interested in understanding the many ways in which oppression impacts lived experiences. An individual's social position is formed through the intersection of each aspect of their identity. In any given present and evolving context, oppression leads to advantages for some social positions and disadvantages for others (Collins & Bilge, 2016; Crenshaw, 1989). Quantitative researchers have attempted to adapt statistical modeling methods to reflect intersectional identities as a proxy for oppression and advantage in their models (Bauer et al., 2021; Schudde, 2018). Applying quantitative approaches to incorporate intersectional identities has brought to light formally unseen disparities (e.g., Dillway & Broman, 2001; Hinze et al., 2012; López et al., 2018).

Although this work demonstrates the importance of employing an intersectional approach to quantitative methods when exploring the influence of oppression and advantage on lived experience, researchers' understanding of how the choice of a quantitative modeling approach influences one's ability to account for intersectionality is still emerging. In addition, this is further complicated because education researchers often work with data that has complex demographic characteristics, such as uneven proportions within categories and varying amounts of within-group variance. This dissertation simulated clustered educational datasets to examine how three methods of modeling intersectional identities perform under various demographic data scenarios. In this simulation study, I varied the number of demographic categories, the proportion of observations within each identity indicator, the within-intersectional group variance, and the overall sample size to create realistic scenarios education researchers encounter when working with demographic data. Analyses then examined the degree to which each of three

methods of modeling intersectional analyses functioned similarly or differently under these conditions.

Theoretical Orientation

Intersectionality

Intersectionality theory posits that each individual's unique intersection of identity is integral to understanding their lived experience. Intersectionality theory also understands that social positions interact with oppression to influence an individual's lived experience (Bowleg, 2012; Collins, 2007; Crenshaw, 1989). Intersectional thinking arose from analyses that centered Black women's experiences (Collins, 1986; Combahee River Collective, 1986; Crenshaw, 1989). Since these initial (re)orientations, intersectional scholarship has extended to a wide range of social positions to consider the intersections of class, race, gender, sexuality, and other demographic characteristics (Collins, 2015).

Intersectional scholars use metaphors to help describe the relationships between identities and the formation of social positions. For example, Crenshaw (1989) uses the metaphor of a traffic intersection where each road is an axis of oppression. Other scholars use the metaphor of interlocking cables to emphasize that different aspects of one's identity cannot be pulled apart when studying experiences with oppression and advantage (Collins, 1991; CRC [Combahee River Collective], 1983) or borderlands where the salience of identity may change based on the social context in which one is currently located (Anzaldúa, 1987). The metaphor a researcher adopts influences the development of their research questions, the method they use to tell the story of their data and the subsequent interpretation and explanation of results.

Based on this theoretical orientation, I explored three facets of the formation and maintenance of social positions: inseparability, mutual constitution, and social context. I used

each of these facets as a lens to consider the fit of a given method to the intersectional framework.

1) Inseparability: Intersectional theorists assume that identities are inseparable; a Black woman cannot be a woman without also being Black. Both identities influence their experience (Crenshaw, 1989, 1991; May, 2015).

2) Mutual Constitution: Identities are mutually constituted; each identity category reinforces the other(s) and is interlaced in multiple systems of oppression (May, 2015; Shields, 2008).

3) Social Context: The social context surrounding an individual will interact with the salience of their identity and impact how oppression manifests to influence their lived experience (Bonilla-Silva, 1997; Collins, 1986; Collins & Bilge, 2016).

Social positions are considered proxies for how forces of oppression advantage some groups while simultaneously disadvantaging others.

Intersectionality assumes that racism, homophobia, classism, imperialism, nativism, ableism, and other forms of oppression exist (Hancock, 2016). Thus, this research assumed that oppression causes differential outcomes between social positions. Collins (2014) defines oppression as "any unjust situation where, systematically and over a long period, one group denies another group access to the resources of society" (p. 4). While oppression is often discussed in terms of its influences on individual identities (e.g., race, sex, age, class), considering only one aspect of oppression is termed "single-axis thinking" (Crenshaw, 1989; May 2015). From the perspective of intersectionality theory, single-axis thinking is problematic because it generalizes the experiences and knowledge of some group members to represent all

group members and ignores the intersecting influence other forms of oppression have on lived experience.

Frameworks for Incorporating Intersectionality in Quantitative Methods

According to Bauer et al. (2021), two of the most influential methodological contributions that account for intersectionality in quantitative research come from McCall (2005) & Hancock (2007). McCall (2005) explains three main methodological orientations when analyzing intersectional identities: intercategorical, intracategorical, and anticategorical. Intercategorical intersectionality explores the difference between social positions, such as how the experiences of Black women differ from those of white women. Intracategorical research investigates the within-group experience of people located within a social position. This approach focuses on a particular intersection, such as Black lesbian women, and explores variation in lived experiences among members of that social position (e.g., Bowleg, 2008). Finally, an anticategorical approach rejects the idea of categorization altogether to explore the advantages and disadvantages a person's experiences without categorizing them into a demographic group.

This research focused on quantitative models applied through an intercategorical orientation, where experiences with advantages and disadvantages are compared across each social position of interest. Hancock (2007) identifies six key assumptions for conceptualizing and using demographic categories within an intersectional paradigm. These assumptions are particularly relevant to an intercategorical approach where the researcher examines the relationship between categories. These assumptions include the following:

1. Multiple background categories play a role in examining complex social problems and processes.

2. Categories should be equally attended to in research but should not always be assumed to have the same relationship.
3. These categories are constructions of dynamic individual and institutional factors.
4. Each category contains within-group variation.
5. Categories should be examined at multiple levels of analysis.
6. Attention is necessary regarding both empirical and theoretical aspects of the research question.

Together, McCall (2005) and Hancock (2007) guide the field on *how* one might apply the theory of intersectionality to quantitative methods. These foundations have clear implications for how quantitative models are applied when accounting for intersectional identities.

Re-Orientation of Quantitative Models

When applying an intersectional lens to analyses, quantitative researchers must ensure the model they select aligns with intersectionality theory. Alignment is challenging, however, because these models typically treat demographic variables as separate, and traditional generalized linear models fail to represent intersectional social positions in a manner that is consistent with intersectionality theory. For these models, dummy coding is generally used to obtain estimates for each identity indicator within a single-axis demographic category (Choo & Ferree, 2010; Rhodes, 2010; Schudde, 2018). For example, a racial variable with four identity indicators may be coded into three binary dummy variables (Black, Asian, Hispanic), with white students as a reference and coded as 0 across each variable. The compound effect of oppression and advantage is then estimated by aggregating each identity indicator estimate. Researchers may also use interaction terms where they multiply separate demographic identity indicator variables, such as Black*female, to attempt to account for the intersection of identity. The

interaction term is interpreted as the combined effect of two identities after accounting for variance "explained" by each separate identity variable. Intersectionality theorists, however, do not conceive of oppression as the linear composite of separate axes. Instead, oppression and advantage are uniquely experienced as a result of a person's intersectional social position. Traditional modeling approaches do not reflect the complex and compounding nature of oppression and advantage associated with intersectional identities; thus, they are inconsistent with intersectionality theory (Bauer et al., 2021; Bowleg, 2008; Misra et al., 2021; Schudde, 2018).

Description of the Problem

Over the past decade, the incorporation of intersectionality theory into quantitative research has gained traction; researchers are exploring ways to alter how demographic categories are incorporated into their models to reflect intersectionality. For example, many researchers center interaction terms or recode their data to include social position variables (e.g., Covarrubias et al., 2018; Jang, 2019; López et al., 2018; Nissen et al., 2021). Approaches such as these enable quantitative researchers to better align their models with intersectional thinking. However, as researchers fit different models to account for intersectional identities in their analyses, they rarely discuss the limitations of their methods (Bauer et al., 2021). The various ways of manipulating variables to represent intersectional identity constrain statistical models. For example, a greater number of interaction terms may lead to a loss of statistical power, more comparisons may increase type 1 error, and uneven distributions of participants between two (or more) identity categories may bias the coefficient and error estimates. However, most papers that employ a quantitative approach to intersectionality have not considered the influence model

selection has on their findings or have not provided detail on the extent to which their model is appropriate for their sample.

One exception to this observation is Mahendran et al. (2022b), who conducted a simulation study to compare the accuracy with which different methods of modeling intersectional analyses explained variability in continuous outcomes. Examining the influence of model selection through a simulation study is informative for comparing methods given different data scenarios. Mahendran et al. (2022b) generated data simulations at three different sample sizes to explore the accuracy of each model for estimating each intersectional identity's relation to a continuous outcome variable. Overall, Mahendran et al. (2022b) found that when sample sizes were medium to large (>10,000 participants), the models explored provided accurate predictions for the outcome variables. Mahendran et al. (2022a) replicated the simulation procedures with models that predict binary outcomes, which also yielded similar findings.

While the simulation studies conducted by Mahendran et al. (2022a, 2022b) provided insight into the influence sample size has on prediction accuracy, additional research is needed to understand each model's performance under more complex data scenarios. Mahendran and colleagues investigated only one condition of sample size distribution and did not manipulate within-group variability. Most demographic categories they examined had relatively even sample sizes except for their racial category, where they simulated 80% of the participants into the "white" category and the other 20%into the "people of color" category. Their study focused only on single-level data, whereas educational datasets are often complex with clustered structures where students are nested in schools. Each of these factors could impact the accuracy of the models. Furthermore, these factors may change other characteristics of the models (such as power or type 1 error), which Mahendran et al. (2022b) did not investigate.

Present Study

This dissertation expanded on existing knowledge about the statistical limitations of three methods of modeling intersectional analyses on a continuous outcome variable. The statistical methods explored in this dissertation focused on what McCall (2005) coins "intercategorical complexity," where the researchers adopt the existing demographic categories to examine inequality among demographic groups. This research compares three methods of modeling intersectional analyses that can be used to explain a continuous outcome variable: 1) Interaction, 2) Categorical, and 3) MAIDHA (multilevel analysis of individual heterogeneity and individual accuracy).

1. Interaction Model:

Interaction terms are created by multiplying two (or more) binary or otherwise coded demographic variables (Hinze et al., 2012). These interaction terms are often interpreted as moderated multiple regression, where one variable moderates the other's relation to the outcome. Generally, interactions are not tested in statistics unless there is a significant relationship between each independent variable and the outcome variable. However, researchers working under an intersectional framework often include all interactions, regardless of the statistical significance (i.e., p-values) of the main effect covariates (Bowleg, 2008; Scott & Siltanen, 2017).

2. Categorical Model:

Each participant is categorized into a categorical variable representing the intersection of their multiple axes of identity. This method handles demographic variables differently than the interaction model by creating or recoding them *a priori*. For example, López et al. (2018) created categorical variables based on race, gender identity, and SES (e.g., black-female-high SES,

where respondents that fall into those three categories are coded as "1".) Then, each category is compared to a reference category (or value, depending on the coding strategy).

3. MAIDHA (Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy)

In this method, intersectional social strata are created based on intersectional identities. The strata are created similarly to the intersectional variables in the categorical approach by combining multiple identity variables. However, the intersectional strata are used as a clustering variable (level 2) in a multilevel model instead of serving as individual predictors. In this method, individuals (level 1) are clustered within their intersectional social strata identities (level 2) (Evans et al., 2018; Merlo, 2018).

Research Questions

This dissertation explored the utility of three methods of modeling intersectional analyses under different demographic data characteristics when modeling a continuous outcome variable in a clustered context.

The fundamental questions that guided this dissertation were:

3. What are the statistical advantages and disadvantages of each model under different demographic data characteristics?
4. In what ways does each model perform differently from one another under each demographic data characteristic condition?

To answer these research questions, I simulated datasets to understand how four demographic characteristics influence the technical quality of the estimates provided by each model. The characteristics of focus include a) the number of demographic categories (and thus intersections); b) the proportion of the sample represented by each demographic group; c) the within-intersectional-group variance in the outcome variable of interest; d) overall sample size.

Across these four characteristics, I built 54 data characteristic scenarios that were applied to each of the three methods, yielding 162 combinations of method and scenario. For each of the 162 combinations of scenario and method, I produced 1000 replications. I used the 1000 replications to summarize, under each combination of scenario and model, the performance of the intersections using the outcomes: bias, accuracy, power, type 1 error, and confidence interval coverage. For each outcome, I had values in which I flagged extreme estimates. I synthesized the results and performed a descriptive analysis to determine patterns across scenarios and models. The findings of this dissertation contribute to intersectional quantitative research methods by providing greater insight into how each model performs under more complex data scenarios.

Significance

As the use of quantitative methods in education to examine relationships between intersectional identities and outcomes of interest expands, researchers are deepening their understanding of the socio-historical forces of racism and oppression in education (e.g., Covarrubias et al., 2018; Jang, 2019; López et al., 2018; Nissen et al., 2021). However, the statistical models typically used by educational researchers today were not designed with intersectionality in mind (Bowleg, 2008). Recognizing the factors specific to intersectionality theory that impact the technical characteristics of results produced by various statistical modeling techniques is needed to help inform the advancement of social justice goals in education.

This dissertation identified the advantages and disadvantages of three methods of modeling intersectional analyses. This dissertation revealed the ways and settings in which each method is beneficial for unveiling disparities between social positions. This research holds the potential to advance intersectional applications in quantitative research by providing an analysis that compares three models across several conditions specific to demographic data. By

comparing three different models, I provided researchers with a deeper understanding of how the choice of model aligns with the complexity of their intersectional data structure and, thus, the extent to which they will be able to answer their research question(s) with a given method.

Chapter 2: Theoretical Orientation & Literature Review

The main focus of this dissertation was to explore how the complex structure of demographic data in education influences three methods of modeling intersectional analyses. This literature review explores intersectionality in quantitative modeling to provide the appropriate background knowledge for this research topic. First, I detail what intersectionality theory is and how it is applied in my research. I then explore the theoretical foundations of incorporating intersectionality in quantitative research. Both of these pieces set the foundations for how we can further explore the capabilities of intersectionality in quantitative methods. I then examine what is currently known about each method's statistical limitations and conceptual fit with intersectionality. While the main focus of this research is on advancing understanding of the utility different methods have for examining statistical relationships through an intersectional lens, it is not possible to explore each model without first exploring its conceptual fit with intersectionality. While something may be statistically sound, it does not necessarily translate to a conceptual fit with intersectional theory. Therefore, both the conceptual and methodological components of each analysis approach are necessary to consider in this review. This literature review sets the foundations to explore how the complexity of demographic data in educational contexts influences the estimates provided by each method of modeling through the lens of intersectionality.

Intersectionality

The history of intersectionality does not fit a neat timeline and arguably should not be linearly traced (see Nash, 2019). Intersectional thinking stems from Black women whose experiences were not represented in feminist and civil rights movements (Collins & Bilge, 2016; Collins, 2019). For example, at the 1851 women's convention, Sojourner Truth argued that as a

Black woman who did not fall into the traditional roles of femininity, she felt that she was not fully included in feminist movements; she famously declared, "Ain't I women?" (see hooks, 2015). With this declaration, Truth called attention to how women's movements fail to account for and represent the experiences and needs of Black women (Crenshaw, 1989a). Black feminist, Julia Cooper, is another example of an influential 19th-century voice who advocated for Black women. Cooper (1982) chastised those who spoke up against racism but failed to account for Black women's unique experiences with oppression. She argued, "Only the Black woman can say, when and where I enter … then and there the whole Negro race enters with me" (Cooper, 1982, p 31). Cooper indicates that uplifting Black women supports the entire civil rights movement; however, the experiences of Black women are often left behind because they face oppression on the basis of both race *and* gender (Crenshaw, 1989a).

Intersectional thinking emerged through activism in the latter half of the 20th century; women of color conversed with civil rights activists through Black Power, Chicano Liberation, Red Power, and Asian American movements. Collins & Bilge (2016) provide examples of core ideas of intersectional thinking in several texts, such as Toni Cade Bambara (1970), who points out that race, class, and gender need *simultaneous* attention for Black women to have a chance of being liberated from oppression. In addition, Frances M. Beal (1970) interrogates Black women's experiences through the *interlocking* systems of racism, capitalism, and patriarchy. As a collective community of Black feminists, the Combahee River Collective (CRC) developed intersectional critiques of social movements. The CRC published "The Combahee River Collective Statement" (CRC (Combahee River Collective), 1983); as a Black feminist statement, it highlights a structure of interlocking oppressions similar to Beal's expression while expanding to consider homophobia and heterosexism (Collins & Bilge, 2016). Collins & Bilge (2016)

explain that the CRC statement is the first document to frame an intersectional lens of identity as a tool for resistance. The ideas and liberatory goals from previous generations of women of color have shaped the metaphor today that is known as intersectionality.

Intersectionality as a Metaphor

A metaphor is a representative literary device that uses spacial relations to provide a mental map of a complex idea, such as human experiences and social interaction. Metaphors provide analytical value that aids in understanding how social structures and power relations are produced; they offer new angles to envision social relations (Collins, 2019). Intersectional theorists use metaphors to describe how multiple forms of oppression interact (Collins, 2019, 1991). In this section, I describe several prominent metaphors to demonstrate intersectional ideas: interlocking systems, borderlands, and traffic intersections (Anzaldúa, 1987; Collins, 1991; CRC (Combahee River Collective), 1983; Crenshaw, 1989). Each metaphor provides a visual explanation of how intersectional identities lead to unique forms of oppression.

In their 1977 statement, the Combahee River Collective introduced the language of "interlocking" to describe how multiple forms of oppression interact with the experiences of Black women struggling with racial, heterosexual, gender, and class oppression (CRC (Combahee River Collective), 1983). Through the term "interlocking," the CRC explains how joint oppression systems cannot be pulled apart. In 1991, Patricia Hill Collins drew on interlocking as a metaphor to discuss the complexity of embedded power relations (Collins, 1991). The imagery Collins (1991) describes of interlocking systems complements her argument that multiple memberships in privileged or subordinated identity groups does not mean that forms of oppression are similar or interchangeable. Thus, the oppressive system of racism that

disadvantages Black women due to their racialized identity cannot be separated from the system of sexism disadvantaging them due to gender identity.

Gloria Anzaldúa (1987), in her book *Borderlands/La Frontera: The New Mestiza,* introduced the metaphor of borderlands to express intersectional thinking from a Chicana/a perspective. Borderlands are the spaces near and between borders where multiple physical areas interact. These physical areas contain different ideas, cultures, and systems of power. Borderlands are a meeting place that reflects complex hierarchical power relations (Anzaldúa, 1987). Anzaldúa (1978) uses borderlands as a spatial metaphor to explore identity without categorizing it while still recognizing that power and oppression are shaped by identity. While anticategorical, this spacial identity still takes different forms as each individual moves between contexts.

In 1989 Kimberle Crenshaw introduced the term *intersectionality* by drawing on a traffic intersection metaphor. The language of intersectionality was quickly adopted as a name to capture the intersectional thinking that had been present within activism and scholarship for decades. As a term, intersectionality took hold and was adopted broadly to provide a name and frame for theorizing social positions and experiences with oppression and advantage associated with those positions. In her introduction of intersectionality, Crenshaw (1989) describes the relationship between identity and oppression as a traffic intersection:

> Discrimination, like traffic through an intersection, may flow in one direction, and it may flow in another. If an accident happens in an intersection, it can be caused by cars traveling from any number of directions and, sometimes, from all of them. (Crenshaw, 1989, p 149)

The traffic intersection metaphor demonstrates what is lost by only considering one axis of oppression. That is, if we limit the focus to women *or* people membered Black's experiences, we miss out on Black women in the intersection who are experiencing oppression from multiple axes.

Some theorists critique the traffic intersection metaphor's simplicity and lack of context (Carastathis, 2016; Garry, 2011). Carastathis (2016) extends the interpretation of the traffic intersection to consider an accident where no driver claims fault. In this interpretation, Carastathis (2016) explains that we cannot track oppression to a single source; we cannot determine when oppression is due to sexism versus racism (or other forms) because it is often due to the collision of both forms. Similarly, Garry (2011) adds a roundabout to the metaphor to capture the complexity of intersectionality. The roundabout helps to paint the picture of how axes of oppression blend and represent experiences from a mixture of identities as a driver changes locations within the roundabout. Conceptualizing intersectionality as a metaphor--the intersection of roads and vehicles, interlocking identities, or the fuzzy region forming a border--conveys the complexity of identity and oppression experienced by an individual in a given context.

Intersectionality as Heuristic

An influential heuristic shifts perspectives and research practices. Collins (2019) explains that as a heuristic, intersectionality provides a set of assumptions and rules of thumb for researchers. As a heuristic, intersectionality provides a tool for exploring intersectional social problems and designing research studies. May (2015) explains intersectionality "accentuates its problem-solving capacity, one that is contextual, concerned with eradicating inequity, oriented toward unrecognized knowers and overlooked forms of meaning, attentive to experience as a

fund of knowledge, and interrogative (focused on asking questions, incrementally and continuously)," (p 19). Intersectionality can restructure how we ask research questions, design methods, and analyze data to better align with the intersection of groups.

The metaphor choice shapes the way intersectionality is understood. The heuristic builds off the metaphor to suggest practical ways to execute research projects. For example, the metaphorical use of an accident at a traffic intersection (Carastathis, 2016; Crenshaw, 1989) may lead to an interaction model choice where the model multiplies separate components to represent the cumulative impact of forces of oppression as two cars collide. As a heuristic, the metaphor of automobiles colliding treats categories as separate (i.e., the individual roads leading to the intersection) and then investigates them in a combined form. They are multiplied instead of added together because the specific source of oppression distributes itself unevenly (i.e., a car from one road may have been driving faster than another). The extension of the traffic intersection provided by Garry (2011) with a roundabout intermeshing axes of identities may lead a researcher to choose a method where there is no separation of the demographic identity variables because the roundabout means that we cannot attribute the cause of the crash to one axis of oppression. Therefore, it aligns with the idea that the axes of oppression are blended to create what Garry (2011) calls a "distinct mixture," where entirely new variables are created to represent membership in multiple identity categories. Finally, the interlocking systems metaphor (CRC (Combahee River Collective), 1983; Collins, 1991) may lead to an analysis that investigates oppression at multiple levels. Each of these metaphors will guide a researcher to make modeling decisions and handle quantitative variables differently based on how they apply the heuristic of intersectionality.

Intersectionality as a Paradigm

The use of intersectionality as a heuristic by a growing body of researchers is contributing to a paradigm shift within academic fields (Collins, 2019). A paradigm shift occurs when a field reorganizes its research practices. Collins (2019) explains:

> When applied to intersectionality, the concept of a paradigm shift suggests that intersectionality convincingly grapples with recognized social problems concerning social inequality and the social problems it engenders; that its heuristics provide new avenues of investigation for studying social inequality; and that it has attracted a vibrant constellation of scholars and practitioners who recognize intersectionality as a form of critical inquiry and praxis. This newly formulated, heterogenous community of inquiry both resonates with the metaphor of intersectionality as a collective identity and relies on heuristic thinking for social problem solving (p. 42).

Collins (2019) explains that scholars in various fields are switching from a traditional view of separate conceptualization of inequality forces to a view that embraces the interconnection of multiple axes of power. For example, intersectionality has brought to light a greater complexity in what was formally known about disparities in STEM education and thus is shifting the way researchers both theorize and study inequality in STEM (e.g., Pearson et al., 2022; Van Dusen et al., 2022; Van Dusen & Nissen, 2020; Wilson & Urick, 2022)

Intersectionality contributes to paradigm shifts within existing research frames, but it may also be emerging as a paradigm in its own right (Collins, 2019). Through continued heuristic application, intersectionality changes the way of theorizing and researching entirely within a field or subfield of academic study. As a paradigm, Collins (2019) explores the core constructs and guiding premises of intersectionality. The provisional core constructs she presents include 1)

relationality, 2) power, 3) social inequality, 4) social context, 5) complexity, and 6) social justice (Collins, 2019; Collins & Bilge, 2016).

1. Relationality: There is a relational process that connects categories of identity. Instead of distinguishing identity axes (such as race or gender), the focus is on examining their interconnection to explore the relations between race and gender.
2. Power: Power divisions based on categories of identity produce social groups that cannot be understood as separate categorical identities. Multiple axes of power lead to interlocking identities, which mutually construct a person's experience with power systems.
3. Social Inequality: Power relations produce social inequalities. We assume that these social inequalities exist, are produced by experiences with oppression and advantage, and are constantly evolving.
4. Social Context: Social context influences the identities and forms of oppression most at play in a given situation and impacts lived experiences and outcomes. Researchers must consider the influence of historical, intellectual, and political contexts that influence and produce present inequality.
5. Complexity: Intersectionality is complex; it intertwines themes of power, inequity, relationality, and context. Researchers are working under this paradigm to better represent the complexity of social structures. This complexity means intersectional theorists can never produce a tidy instruction manual or a set of methods that tell others how to "apply intersectionality."
6. Social Justice: Historically, social justice has been central to intersectional thinking. While there are different beliefs about how intersectionality handles social justice, many

> scholars believe that research cannot be intersectional if it does not work toward social justice. Meanwhile, many scholarly papers today work to tease out how to apply research methods that may not directly engage with social justice. Still, those researchers believe this foundational work will produce more applications that directly lead to social justice and liberation.

In addition to core constructs, Collins (2019) also introduces provisional guiding premises of the paradigm for practitioners to follow:

> (1) Race, class, gender, and similar systems of power are independent and mutually construct one another.
>
> (2) Intersecting power relations produce complex, interdependent social inequalities of race, class, gender, sexuality, nationality, ethnicity, ability, and age.
>
> (3) The social location of individuals and groups within intersecting power relations shapes their experiences within and perspectives on the social world.
>
> (4) Solving social problems within a given local, regional, national, or global context requires intersectional analyses. (p. 44)

These four guiding premises offer a starting point for understanding shared assumptions under an intersectional framework. Collins argues that these premises, coupled with the core constructs, are the foundation for an intersectional paradigm. Overall, the metaphors, heuristics, and paradigm (shifts) of intersectional thinking serve as practical cognitive architectures. They lead us from concepts to strategies and finally onto a framework for using intersectionality in research practice.

From Social Position to Lived Experience

Developing Social Positions. I refer to social positions as the intersection of identities, such as age, sex, gender, race, ethnicity, or class. I focus on three key ideas to define social positions and their relation to lived experiences. First, the identities that create a social position are inseparable. Second, the axes of oppression affecting an individual are mutually constituting and thus co-construct a lived experience. Third, the tangible and social context shapes an individual's social position. After examining these three facets, I discuss how systems of oppression form the lived experience for any given social position.

Inseparability of Identity. The notion of inseparability of identity is foundational to intersectional thinking. Most intersectional metaphors orient themselves around inseparability, such as an interlocking system that is incapable of being separated. Although researchers often discuss oppression in terms of its influences on individual identities (i.e., racism, sexism, ageism, classism), only considering one aspect of oppression functions as "single-axis thinking" (Crenshaw, 1989; May, 2015). Separating a Black woman's identity into that of being Black and that of being a woman does not adequately represent a Black woman. Her identity as a person who is Black intersects with her womanhood to create unique forms of oppression only experienced by individuals with similar social positions. When researchers disaggregate identity, we miss out on unique experiences at the intersection of multiple axes of identity (and thus forms of oppression).

Single-axis thinking falsely universalizes the experiences and knowledge of some group members to represent the experiences, needs, and claims of all group members; this can lead to unseen disparities. The wage gap provides a helpful example of how single-axis thinking can mislead the reality of wage divides. Women make, on average, $0.82 for every dollar the average

man makes. Black individuals make less than non-Hispanic whites; for every dollar a non-Hispanic white makes, a Black individual earns $0.76. However, these numbers do not paint the picture of individuals at the intersection at which women who are Black operate. Focusing only on Black women, we see that they earn $0.63 compared to non-Hispanic white men (Kochhar, Rakesh, 2023; Wilson & Urick, 2022). Independently investigating wages separated on a single axes by gender or race obscures the stark wage differences of multiple intersecting categories.

In her legal analysis of workplace discrimination cases, Crenshaw (1989) demonstrates how Black women fail to receive a judgment in their favor because the cases focus on a single axis, such as civil rights or gender discrimination. For example, in *DeGraffenreid v General Motors*, five Black women alleged discrimination, but the courts examined their claim through the single-axis lenses of racial discrimination and gender discrimination. They ruled that there was no discrimination on account of race and there was no discrimination on account of gender. However, the women were experiencing discrimination through their multiple identities as Black women. Crenshaw argues that the unique social position of a Black woman experiences greater inequality than considering either identity (Black or women) separately.

Mutually Constitute. Intersectional thinkers often use the language "mutually constitute(d)" to recognize that identity categories are not independent of one another. Shields (2008) explains that categories of identity *reinforce* each other. “The formation and maintenance of identity categories is a dynamic process in which the individual herself or himself is actively engaged. We are not passive “recipients” of an identity position, but “practice” each aspect of identity as informed by other identities we claim” (Sheilds, 2008, p. 302).

May (2015) explains that our identities are interlaced in multiple systems of oppression which co-construct the lived experience. Similar lived experiences of individuals grouped in

social positions are due to how oppression interacts with the formation of each aspect of our identity.

Social Context. Multiple levels of context situate an individual by influencing how they identify and how others identify them. The identity they are assumed to have influenced their experiences with oppression. Within each level of context, the salience of an individual's identity may shift. Each characteristic of someone's identity will influence their experiences, but at some levels, one identity may seem more prominent than the other one it co-constructs. Bonilla-Silva (1997) discusses how, depending on the context, a particular aspect of identity may take precedence over others. For example, in U.S. systems, the racial struggle may be more salient than gender or class, whereas, in Brazil, class is more salient. Similar aspects of identity salience are prominent across social structures within the United States and throughout an individual's development, depending on their context.

Another way to consider the relationship between context and oppression is through a sociological lens. The sociological lens reveals how oppression and advantage operate at multiple contextual levels, from the individual to their larger context and the policies that orient their experience. A framework of a socio-ecological approach is provided by McLeroy et al. (1988), where a higher education researcher may consider the: 1) individual intrapersonal knowledge, attitudes, and behaviors of an individual; 2) the relational/ interpersonal social support structures surrounding that individual; 3) the institution factors and formal rules and regulations; 4) community factors and relations among the organization and institution; and 5) local, state, and national laws/ policies. This multi-level focus is similar to Bronfenbrenner's ecological systems theory that demonstrates the role micro (individual), meso (locale), and macro (societal) level factors have in influencing human experience (Bronfenbrenner & Evans,

2000). The micro level captures levels 1 and 2 of the sociological lens; the meso level captures 3 and 4, and the macro is similar to level 5.

Regardless of the framework applied, recognizing the multiple levels of power dynamics helps us contextualize an individual's experience in systems of advantages and disadvantages. For example, a light-skinned Latina may not find her ethnicity salient in primarily white spaces since she may pass with racial/ ethnic privilege. However, this identity may be more salient in a Latina/o/x-dominated social space (such as a Latino/a/x Student Alliance at a university) as it creates and forms bonds with others who share similar cultural values. Her ethnicity may be salient in this situation because it creates a sense of belonging. Whereas, perhaps in a white-dominated space where her ethnicity may not be as recognized, an identity such as class may appear to have a larger influence on her experiences. However, the class identity exists within and is mutually constituted by racism and sexism. While, it might appear more salient, her financial status is co-constructed with systems of racism and sexism.

We can widen the lens for this example to investigate the community/public policy levels. This student's experience in her Latino/a/x Student Alliance at her university may be different from another woman in a similar social position at a different university in perhaps another part of the country. The policies and procedures at the national, state, and college levels will influence an individual's experience on campus. For example, affirmative action or DACA programs may have helped both women attend college; but these policies may be looked upon more favorably in certain locales compared to others. Within the institution, the Latino/x/a student alliance may be more supported (and thus funded, creating better experiences) on one college campus than another based on the unique policies and events that influence that campus.

Oppression. Collins (2014) defines oppression as "any unjust situation where, systematically and over a long period, one group denies another group access to the resources of society" (p. 4). Similarly Bell (2016) defines oppression as "the term we used to embody the interlocking forces that create and sustain injustice" (p 29). The intersections of our identities and their context influence how an individual or group is subjected to oppression and advantage (Zinn & Dill, 1996). The advantage may lead to greater opportunities, which Shields (2008) argues is different from avoiding disadvantages. A group may experience advantages in one context and disadvantages in others.

Multiple frameworks help us examine how oppression operates within different contexts. Collins (2014) focuses on the hierarchical structure of power and its operation at multiple levels. Under the framework of the "matrix of domination," Collins (2014) describes the four domains of power that organize the systems of oppression that individuals experience. Collins (2014) and Collins & Bilge (2016) present the four domains of power and explain their influence on our social positions.

1) Interpersonal Domain: This is the unique social position in which an individual operates; the multiple axes of oppression shape a person's identity (their interests, experiences, needs, and desires.)
2) Disciplinary Domain of Power: Social positions dictate how rules are communicated to us, which rules are (or are not) implemented, and when. These rules both explicitly and implicitly send individuals to various life paths and provide options that may not be viable to others in different social positions.

3) The Structural Domain of Power: How an organization or specific context is organized to reflect power, perpetuate inequality, and promote particular individuals/ organizations over others.
4) Cultural (or Hegemonic) Domain of Power: Society can hide that playing fields are not level. However, the cultural domain claims that society at its core is unjust, and conditions for groups in various social positions across contexts are not equivalent.

The domains of power offer an example of mutual constitution: each domain of power operates in conjunction with, and thus reinforces, the remaining three. The multiple forms of oppression operating are part of a greater system that is not in control of an individual. The matrix of domination helps to show that there are rarely pure "winners" or "losers;" instead, most people experience advantages from one level of oppression and disadvantages from another. The context shifts an individual's experience with the world due to the interconnected forces of power and oppression described in the matrix of domination. An individual's context can influence their lived experience because of how oppression interacts with the salience of their identities.

Theoretical Underpinnings: Intersectional Heuristics for Quantitative Methods

The theoretical underpinnings of intersectional heuristics draw from Leslie McCall and Ange-Marie Hancock, who have laid the foundations for applying quantitative methods to an intersectional framework. Leslie McCall presents typologies of methods, and Ange-Marie Hancock presents assumptions for research within an emerging intersectional paradigm.

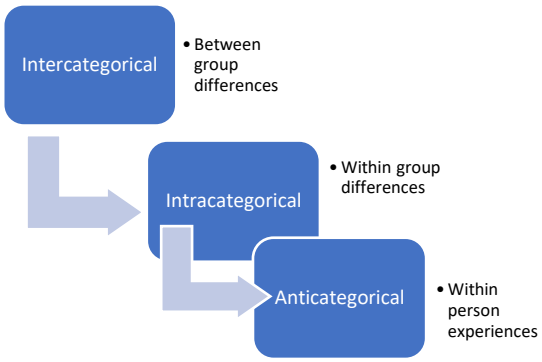
Typologies of Methods

Leslie McCall (2005) explains three methodological approaches when accounting for intersectional identities: Intercategorical, Intracategorical, and Anticategorical. While research in practice may not fit neatly into each category, it is helpful to consider these three approaches to

intersectional analysis. Figure 1 demonstrates my conceptualization of the spectrum of approaches McCall (2005) describes. The purpose of this visual is to represent how the research focus narrows from one approach to the next. These arrows should not be misconstrued as hierarchical but instead represent how the constriction of the research focus.

Figure 1

Visualization of McCall (2005) Methodological Approaches



The layered blocks show how the approaches start broad with the inclusion of many demographic categories of research participants, then narrow to focus on specific groups and individuals, and finally reject categorization of identity altogether to focus on each person as a unique individual. I use blocks and arrows to demonstrate that the same data source can be used to form each approach, but who the sample focuses on may narrow. The space between intercategorical and intracategorical indicates that these are two distinct approaches, whereas the connection of intracategorical and anticategorical represents a spectrum from investigating a particular category to denying the use of categorical framing altogether.

Intercategorical. Intercategorical analysis focuses on the intersections of a complete set of multiple categorical social positions to investigate advantages and disadvantages between groups. Within intercategorical research, every demographic category of interest based on theory

is cross-classified to form unique intersectional groupings. For example, López et al. (2018) applied an intercategorical approach to study achievement gaps at a university located in the Southwest. They re-coded their single-axis data into social position variables based on race-ethnicity, class, and gender. This approach led to categories such as “Black female high-income,” “Black female low-income,” and “Asian male high-income.” McCall (2005) describes this type of categorization as provisional to "empirically chart" the inequalities between groups. Therefore, Lopez et al.’s (2018) analysis describe differences between social position groups as a proxy for forces of inequality. The intercategorical approach is beneficial for understanding the experiences of multiple social positions. However, it comes with the caution of overemphasizing between-group experiences without recognizing the within-group differences (May, 2015).

Intracategorical. The intracategorical approach examines a specific social position, such as research on Black women, and explores the lived experiences (and differences in experience) of members in that position. Intracategorical approaches still member individuals into a social position, but the focus is on the within-group experience instead of the relationship between positions. For example, Bowleg (2008) studies Black lesbian women, a social position that she explains has been historically understudied. She documents shared aspects of the lived experience of Black lesbian women that were previously unexplored. An intercategorical approach can reveal the complexity that a group experiences due to the intersection of oppression. Intracategorical approaches are valuable in acknowledging that a social position does not lead to a homogenous experience; recognizing within-group differences is important in intersectionality research.

Anticategorical. Through the critiques of categorization, anticategorical complexity arose. Drawing from the idea that categories are not a reflection of reality due to their social

construction, anticategorical approaches attempt to avoid categorization altogether. Thus, an anticategorical approach is taken when researchers investigate an intersectional social position without categorizing it. Anticategorical intersectionality focuses on the within-person experience and how forms of oppression manifest in a participant's lived experience (McCall, 2005). By rejecting categories, anticategorical researchers seek to avoid group-level generalizations.

Emerging Paradigm

Hancock (2007) presents a compelling list of assumptions for quantitative researchers to consider when attempting an intersectional analysis. In these assumptions, Hancock (2007) details how researchers can conceptualize and use multiple intersecting variables to guide analysis. Hancock (2007) explains these assumptions are valuable for how they guide researchers as they operationalize indicators of inequality within and between social positions. Given their categorical focus, these assumptions apply to McCall's (2005) conceptualizations of intercategorical and intracategorical analysis, but are not applicable to an anticategorical approach. Hancock argues that intersectional research should jointly address all six of these assumptions:

1. Multiple background categories play a role in examining complex social problems and processes.
2. Categories should be equally attended to in research but should not always be assumed to have the same relationship.
3. These categories are constructions of dynamic individual and institutional factors.
4. Each category contains within-group variation.
5. Categories should be examined at multiple levels of analysis.

6. Attention is necessary regarding both empirical and theoretical aspects of the research question.

Hancock's (2007) assumptions lay a foundation that encourages quantitative methodologists to explore the intersecting categories and who is within them while simultaneously centering each participant's context. These assumptions support the notion that intersectional research is not a method, but instead a way of theorizing a problem and designing a study. In conjunction, Cho et al. (2013) argue that what "makes an analysis intersectional—whatever terms it deploys, whatever its iteration, whatever its field or discipline—is its adoption of an intersectional way of thinking about the problem of sameness and difference and its relation to power" (p. 795). Hancock's (2007) six assumptions demonstrate that critical thinking is necessary at every step of the research process to mold our quantitative methods into an intersectional paradigm.

Resituating the Paradigm

Researchers must frame their studies and select methods that best support their understanding of intersectionality within the context of their research. Intersectional thinking challenges traditional quantitative approaches because of the complexity with which intersectionality conceives of how social positions, oppression, and context influence one's lived experiences. To move into an intersectional paradigm, our quantitative approaches and the assumptions guiding them must shift. Quantitative researchers must question the normative assumptions of their training, which often have anchors in positivism. Positivism is an epistemology that assumes that robust research methodology can lead to neutral and bias-free results. Intending to uncover "truth," positivism assumes that rigorous research can lead to objective reality. As I will explore, many of the assumptions of this approach do not hold when working within an intersectional paradigm. This section probes strategies for integrating

intersectional thinking into quantitative methodology. First, I detail how an additive approach is incompatible with intersectional thinking. Then, I examine how quantitative researchers can better account for the three facets of social positions: inseparability, mutual constitution, and context. Finally, I will offer some limitations of the intercategorical approach.

Additive Approach

Traditional modeling methods often rely on an additive approach. When investigating the relations between different demographic variables, an "additive approach" treats identity as the sum of the individual effects of each single-axis form of oppression (Choo & Ferree, 2010; Rhodes, 2010; Schudde, 2018). To understand the experience of someone in multiple identity categories, the researcher adds covariate coefficients to estimate the combined experience. For example, consider an additive approach to the question, "*How do Black females experience sense of belonging on their college campuses?"* The coefficient for a variable representing females and the coefficient for a variable representing Black racialized identity are added to yield an aggregate estimate of the experience of Black membered females. Using this approach, a researcher may assume that if being Black is positively related to the outcome and being female is positively associated with the outcome, then the effect of being Black and female would also be positively associated with the outcome. The additive approach is problematic because it treats intersectional identity as separable forms of oppression and advantage. In this scenario, the researcher does not investigate the unique position of being a Black female but assumes that their experiences are an aggregate of the effect of two distinct aspects of their identity.

Mutually Constituted Categories

In addition to accounting for inseparable identities in the model, researchers must interpret identity categories jointly. However, researchers who formerly focused on one

demographic category often interpret their findings with preference given to that category (Hancock, 2007). Collins (2019) describes this case as prioritizing one "master" category and adding other categories to it, such as prioritizing a discussion of gender and later interpreting racial findings. Focusing first on one category creates the impression that forms of oppression influencing the categories are separate rather than mutually constituted. To honor the mutual constitution of categories, researchers must apply joint attention to both categories throughout the research process (Collins, 2019; Hancock, 2007).

Inseparability of Categories

Intersectionality posits that identity categories are inseparable. The additive approach treats the experience for members of an intersectional group as a linear combination of sexism and racism. It ignores the potential intersections among forms of oppression and advantage as a result of the intermeshing of one's gender and racialized social position. Most statistical methods require the assumption of independent covariates. Given the intersectional understanding of inseparability, treating variables as independent terms violates statistical modeling assumptions and can lead to inaccurate covariate estimates and interpretations (Shields, 2008). Because categories of identity (and thus forms of oppression) are inseparable, researchers must adapt their approaches to better represent demographic categories. While data may have been collected from separate categories, one solution to better align with inseparability is to manipulate the variables to member participants into social positions. For example, the categorical method involves re-coding separate demographic categories (such as gender or race) to be unique variables representing intersectional identities (such as Black-female). In this way, each aspect of identity is no longer conceived of as independent from every other aspect.

Incorporation of Context. Within an intersectional paradigm, research approaches must acknowledge the context of the participants. Hancock (2007) explains that intersectional research should incorporate context at multiple levels. Research should explore how these levels interact to create a unique lived experience (Bowleg, 2012). To best represent context, researchers may model relevant contextual variables. For example, we know that Black students (individual level) at predominantly white universities (institutional level) report a lower sense of belonging compared to Black students at Historically Black Colleges and Universities (HBCUs) (Hurtado et al., 1998). A researcher could account for this interaction between college contexts and individuals in a multilevel model which nests students in the college they attend (Raudenbush & Bryk, 2002). This approach can allow a researcher to model HBCU enrollment at a second level. The researcher can then investigate cross-level interactions between the individual's identity and their HBCU enrollment. By considering multiple levels, research can portray how an individual operates within a larger ecosystem.

While researchers may not always work with data capable of exploring the full context and its cross-level interactions, they must situate the literature and the results within the greater context (Bowleg, 2008; Bowleg, 2012; Cuadraz & Uttal, 1999). For example, Bowleg (2012) integrates multiple research studies in her explanation of Black lesbians' experiences in health. Through this incorporation of literature, she is able to explain how experiences with structural racism limit her participants' ability to feel comfortable disclosing their sexual identity, which further defines the full access to the health care they receive (Bowleg, 2012). Pairing intersectional research with a model that conceptualizes the influence of context and cross-level interactions is beneficial; such as the previously discussed socioecological model (McLeroy et

al., 1988), ecological systems model (Bronfenbrenner & Evans, 2000), matrix of domination (Collins, 2014) or another framework which similarly theorizes interrelations between levels.

Reframing the research question(s) can set up a dialog that explores how context shapes an individual, rather than the popular narrative that directs attention to the experience of a group after the context is controlled. For example, Johnson & Jabbari (2022) study Black student suspensions in majority white schools. They frame their research question, "How do the math performances and beliefs of suspended and non-suspended students from varying racial groups change as a high school's white student enrollment increases?" In this question, they acknowledge that context influences individual-level experiences. Johnson & Jabbari (2022) are then able to use existing literature on social inequality to help explain their results.

When applying an intersectional lens, Bowleg (2008) suggests that research questions should pointedly highlight discrimination and prejudice beyond looking at demographic differences. Cole (2009) suggests asking what role inequality plays in a given context; this type of question enables the researcher to draw attention to how intersectional group membership positions a person's experience in the context of the inequalities they experience. One way to do this is to frame questions around group similarities to help situate how the outcomes reflect institutional and cultural influences (Cole, 2009). This framing can enable the research to transcend beyond investigating differences towards a discussion on how forces of oppression and advantage are causing and shaping the differences observed. Overall, research questions are a powerful tool to focus research on context.

When addressing context, researchers should also consider their use and interpretation of covariates. Often contextual variables are interpreted as "control variables" to describe the "effect" of a demographic category after the contextual covariate is controlled for, or the

variance associated with the covariate is partialed out (Spector & Brannick, 2011). For example, a researcher who is modeling academic achievement may use a variable that indicates the type of locale an individual is from (urban, suburban, rural). They may explain that after controlling for race, the locale has a significant impact on student achievement. Or, if they are focusing on the demographic category, they may state after controlling for the locale that the effect of race is not related to student achievement. But, the locale someone lives in and the resources within that locale are directly related to systemic oppression. The experience of a given social position is shaped by context; context often dictates how forms of oppression manifest. Therefore, by attempting to "control for" or "partial out the effects of" a locale, the researcher may actually be explaining away forms of oppression. Although it is not possible to incorporate all aspects of context that influence someone's advantages and disadvantages, researchers can come closer to an accurate depiction of reality by integrating literature, using theory to shape their analysis, and reframing their narrative around oppression.

Categorization

Essentialism treats identity as though it is constant across contexts. Most analyses reflect an essentialist conception of identity: identity is treated as a fixed, permanent, and stable characteristic within a sample or population. However, intersectional scholarship focuses on *subjective* identities. The social context around a person will influence how they view or express their identity. Categorizing variables is necessary for most statistical models; but categorization requires discrete boundaries that may be challenging to establish accurately. Demographic variables—such as racialized identity, gender, and ethnicity—do not offer reliable categorizations because they are social constructs. (Viano & Baker, 2020; Zuberi & Bonilla-Silva, 2008). The social construction means each demographic construct's definition is unstable

over time (as society changes, meanings can shift). For example, racialized identity is a dynamic construct in political and social history (James, 2008). Therefore, the category that someone members in today may be different from the category they member themselves in tomorrow. Categorization is an inherent limitation of quantitative methodology because it stratifies social constructs into categories and treats them as fixed (James, 2008).

In the context of secondary data, I use the terminology "membering" individuals into categories, which assumes that this categorization is imprecise and that the participants are not given full autonomy over their demographic placement. Because demographic variables are often socially constructed, membering individuals into categories is fraught with error (James, 2008; Kaplan, 2014; Zuberi, 2001). For example, Viano & Baker (2020) investigated administrative data collection in schools and found a lack of reliability in how an individual is classified based on the wording of the demographic question and who is in charge of membering individuals into categories. Therefore, while this research relies on categorization, it is essential to recognize its imperfections. Despite limitations, categories can allow groups to serve as a *proxy* for oppression and advantage experienced and include social positions in research that previously had not been explored (Hancock, 2007).

Categorization is helpful when exploring those experiences because intersectional thinking understands that the lived experience of members who experience similar intersections of oppression are more similar to each other than people in other social positions. The membered overlapping identity categories can help us explore oppression's complexity and how it operates in diffuse and differentiated ways (Cho et al., 2013). Nonetheless, despite the analytic utility afforded by forming categories and membering individuals into a given group, researchers must recognize that these categories are provisional and fluid.

Subgroup Variation

When researchers create categorical demographic variables, they place individuals who vary in a heterogeneous manner into a single homogenous group. For example, an Asian demographic category comprises people with a variety of cultural values. Researchers often discuss covariate estimates with the implicit assumption that the aggregate result encompasses the experience of all members of the group. This is in contrast with an intersectional paradigm where we assume there is variation within the subgroup (Hancock, 2007; McCall, 2005). For example, Hancock (2007) points back to the broad "Asian" demographic category. She explains that within education, the needs of Southeast Asian students are often unseen because their experiences are membered into a more general Asian category; the "Asian" aggregate outcomes do not necessarily reflect the experiences of Southeast Asian students (Hancock, 2007).

To avoid homogenous interpretations, researchers should examine variability within each category. For example, including descriptive information for each category (e.g., group deviations or residuals) can help demonstrate the variability that exists around the mean score estimated for all group members. For groups that seem to have a wide degree of variation, it may be helpful to provide graphs, further disaggregate descriptive statistics, or perform supplemental analysis. Even when the focus of analysis is between groups, examining the within-subgroup variation can provide direction for further research. For example, in an intercategorical approach, researchers may notice that the Asian category has relatively high outcomes but with high amounts of variation. Therefore, their subsequent intracategorical study may focus on intersectionality within the Asian demographic to further uncover formerly unseen experiences.

Beware: Intersectionality as a Testable Hypothesis

Researchers have several areas to consider as they move into an intersectional paradigm. Specifically, quantitative researchers with positivist-based training must avoid treating intersectionality as a "testable hypothesis." As Hancock (2013) explains, positivist-oriented researchers may approach intersectionality as a hypothesis that is tested to determine whether an intersection "exists." For example, Bauer et al. (2021) argue that treating intersectionality as a testable hypothesis results from the author's lack of clarity on using theory to inform their analysis. In intersectionality, we assume differences in exposure to advantages and disadvantages are a result of a social position. Instead of focusing on *whether* an intersectional group exists, intersectional researchers focus on *how* experiences differ between and within social positions.

This theoretical work guides the field in how one might apply quantitative methods within an intersectional paradigm. As a research paradigm, intersectionality is more than methods. Further, simply including intersectional groupings in an analysis does not qualify the analysis as intersectional. Rather, an intersectional analysis is informed by theory, which guides the entirety of the research process.

Conceptual and Statistical Limitations and Advancements

This research compares three methods of modeling intersectional analyses; 1) interaction model, 2) categorical model, and 3) MAIDHA (multilevel analysis of individual heterogeneity and discriminatory accuracy). This dissertation compares each method's conceptual and statistical limitations, specifically in the context of how they handle complex demographic data. In this section, I suggest several areas of expansion for both conceptual application and statistical considerations. Then, I discuss existing comparisons of intersectional methods. In this review, I observe that the complex structure of demographic data is a topic that has been understudied and

is in need of further exploration. This section provides a necessary foundation for where the current state of research is for each of these models and where it can progress.

Interaction Model

In an interaction model, the product of two variables, X*Z, is used to create an interaction term (Cohen & Cohen, 1983). Researchers who use this model to represent intersectional experiences investigate the interaction of single-axis demographic variables. To do this, they first code the demographic identity indicators into variables, most often through a process of dummy coding (Daly et al., 2016). Dummy coding incorporates categorical variables into a regression model by creating k(categories) - 1 variable for a demographic category. Each variable is coded as 1 if that individual falls into that category or 0 if they do not, with a reference category coded as 0 across all variables of a given category. For example, for a set of variables representing gender identity, a researcher may have three categories in their dataset—male, female, and nonbinary. A female variable is created where the female is coded as "1," and all other participants as "0", and a nonbinary variable is created where nonbinary individuals are "1" and all other respondents would be "0". Therefore, there is no "male" variable, and instead, males are coded as 0 across both the two gender identity variables.

To investigate interactions between dummy-coded variables, a researcher creates interaction terms for all variables of one demographic category (such as gender identity) across all variables of another demographic category (such as racialized identity). For example, a researcher may use "male" and "white" as reference categories for racialized and gender identities. They then create interaction terms by multiplying the value of each racial category with each gender category (i.e., "female"*"Black," "nonbinary"*"Black," "female"*"Asian,"

"nonbinary"*"Asian," and so on). The interaction terms are then included in the statistical model, and a coefficient for each interaction term is estimated.

Conceptual Implications.

Additive Components. The interaction model uses additive components to create interaction terms; the approach begins by examining separate demographic variables. Intersectionality does not understand social identity categories as distinct where gender is separate from racialized identity (e.g., white and female); instead, intersectionality recognizes multiple identities as inseparable (e.g., white-female). Thus, including the discrete demographic terms in the model and interpreting each term separately violates the assumption of inseparability. Researchers can advance the use of this approach to have greater conceptual alignment by only focusing on the interpretation of the interaction effect (Scott & Siltanen, 2017).

Reference Category Exclusion. The interaction model does not allow for simultaneous comparison of all possible social identities because reference groups are not comparable. For example, if I use gender and race in an interaction model—with "white" and "male" as reference groups, I may find a significant interaction of "Black"*"female." However, I would not know if there was an interaction between the most privileged categories of "white"*"male" or the combination of privilege and disadvantage such as "Black"*"male." Therefore, when using a coding approach that requires reference categories, the interaction terms do not account for all possible interactions among categories.

Multiple Moderated Regression. An alternate approach to conceptualizing interaction terms employs "Multiple Moderated Regression" or "MMR" (Jaccard & Turrisi, 2003). Typically, an interaction term is only employed in a model if the main effects are significant.

Then, that interaction is retained in the model only if it is also significant (Scott & Siltanen, 2017). In MMR, a significant interaction term means that a third variable moderates the relationship between an independent variable and an outcome variable. Thus, the interaction of X and Z on outcome variable Y is interpreted as Z moderating the relationship between X and Y or that the slope of Y on X differs across values of Z (Aguinis et al., 2005). To interpret the interaction terms of demographic variables, such as "female"*"Black," a researcher may explain that when a respondent identifies as female compared to male, there is a fixed unidirectional change in the outcome variable moderated by their identity as Black compared to white. MMR does not align with intersectional thinking because it suggests prioritizing one identity over the other. For example, an excerpt from Jaccard & Turrisi (2003) reads, "and how these ethnic differences vary as a function of gender. In this case, ethnicity is the independent focal variable, and gender is the moderator variable" (p 4). In this example, Jaccard & Turrisi (2003) center ethnicity in the analysis. Hancock (2007) argues that prioritizing one identity over another is a research pitfall when attempting to account for intersectionality. This prioritization conflicts with the assumption of mutually constituted identity categories.

Another problem with interaction terms is that they are used only if the single-axis variables are significant. Under MMR, if one or both single-axis variables are not significant, then an interaction term based on the two single-axis variables is not included. Intersectional researchers can explore ways to conceptualize interaction terms outside the MMR hypothesis. One way to better align with intersectional thinking is to include the interaction terms regardless of the main effects (Scott & Siltanen, 2017). While we risk losing statistical power by including every interaction term, at a minimum, researchers can include interactions for social positions

that they theorize to be salient in a given research context. Under an intersectional paradigm, it is often reasonable to include race, ethnicity, gender identity, and class variables.

Another way to better align with intersectionality is to shift our interpretation of the interaction terms. For example, Schulman et al. (1999) use interaction terms in their study on physician recommendations for cardiac catheterization. Using the reference groups "white" and "male," Schulman et al. (1999) explain that Black women are more likely than white men to experience cardiac catheterization. In their discussion, they do not prioritize or discuss moderation of racialized or gender identity categories (Schulman et al. 1999). The joint result of the interaction is assumed to be more influential than the individual influence of each demographic variable.

Statistical Implications.

Power. A model with high statistical power is more likely to detect a significant interaction if it exists. Interaction terms often influence the statistical power of a model, as the number of interaction terms in a model increases, the statistical power decreases. Auginis & Gottfredson (2010) explain that for interaction terms, the overall cell size of each group and the overall sample size must be considered. Statistical power increases when categories are relatively equal in sample size. A Monte Carlo simulation study by Alexander & DeShon (1994) found that when sample sizes are unequal across subgroups, the large-sized subgroup has more error variance, which violates the assumption of homogeneity and causes the power to decrease. Bell et al. (2019) explain that sample size is less problematic for two-way interactions (e.g., race*gender). For example, in an analysis that uses three categories of gender identity—male, female, and non-binary—the model will have higher power if each group has a similar sample size. However, a researcher will likely find that the number of participants membered into the

non-binary category is drastically lower than that of male or female, leading to uneven group sample sizes. However, with greater model complexity, sample size constraints frequently lead to underpowered models when there are three-way (e.g., race*gender identity*class) interactions. Educational researchers often work with demographic data with a large sample size in some categories (i.e., white) and small sample sizes in other categories (i.e., indigenous/ native). Therefore, the disparities in sample size may present limitations for this approach.

Combining Categories. To increase a model's statistical power, researchers often create a new variable that combines multiple categories of identity (McClelland & Judd, 1993). For example, many education studies focus their racial analysis on white compared to other races or ethnicities because this combination of categories can lead to more equal sample sizes. However, combining categories directly contrasts with the core principles of intersectionality. By treating separate identities as one, this approach ignores the unique experiences associated with each distinct identity. In addition, combined categories increase the within-group variance, which will introduce greater error into the model (McClelland & Judd, 1993). This additional error can lead to issues of detection and interpretation of interactions, reduction of power, and even the introduction of spurious interaction effects (Busemeyer & Jones, 1983; Maxwell & Delaney, 1993). Thus, researchers may consider accepting a higher type 1 error rate and avoid collapsing demographic categories.

Categorical Approach

The categorical approach involves the use of intersectional categorical variables. For example, a researcher interested in the relationship between gender and race creates variables representing "female-Black," "male-Black," "non-binary-Black," "female-Asian," and so forth. For a researcher using an existing dataset, this would mean recoding formerly separate

demographic terms into intersectional categorical variables. While this approach produces conceptual benefits, it also yields additional statistical constraints that I explore.

Conceptual Implications. Conceptually, the categorical approach improves on some of the limitations discussed with the interaction model. This model aligns with the notion of the inseparability of identity by creating the variables before building the model and does not include variables representing single-axis identity categories. Instead of multiplying variables to create interactions, the researcher creates intersectional variables apriori. In addition, the categorical approach improves the issue of reference category exclusion because there is only one reference category across all intersections of a demographic variable. While conceptually different, this approach yields coefficients that are similar to the interaction model (Evans et al., 2018). While a researcher will likely not find substantially different regression coefficients with this model, they will be able to view more results through a larger number of comparisons available, potentially explain a greater amount of the variation, and demonstrate alignment with intersectionality.

Reference Comparisons. Categorical demographic variables require a reference group or value. Like the interaction approach, categorical models often rely on dummy coding (i.e., López et al., 2018). The reference category frequently defaults to the most advantaged group, which is often white males. This selection may make sense if you want to understand the most prominent differences between groups. However, defaulting to the most privileged group means that every other group is implicitly (or explicitly) described as "less than." Evans et al. (2018) argue that using the most privileged group as a reference category reinforces the idea of privilege being a “default” social achievement. However, researchers can frame their questions to shift interpretation. If the research question(s) are framed in terms of the outcomes being a product of the individual, then it’s likely the results will be interpreted from a deficit narrative.

Alternatively, when the question is framed in terms of the impact of oppression, the focus is on the overall systems of oppression that influence the experiences of those located in a given social position. Thus, the limitations of the reference group can be mitigated with carefully articulated research questions.

Alternatively, researchers can also use categorical coding strategies to shift the narrative of their results. While dummy coding is the most frequently used strategy, there are other approaches that do not rely on an individual reference group, such as effect coding and orthogonal coding. An example of each coding procedure is provided in Table 1. Effect coding compares each category to the unweighted average of all categories (Hardy, 1993). Effect coding is well-suited for categories with relatively even sample sizes. However, an unweighted average presents an issue when social position variables differ notably in sample size. In education settings, the most privileged groups often have the highest representation; thus, effect coding may still use a comparison value derived from this privileged group. To counter this, Te Grotenhuis et al. (2017) suggests following a weighted effect scheme to better account for unequal observations across categories. Sweeney & Ulveling (1972) introduce a weighted effect coding scheme based on sample sizes. Another approach by Daly et al. (2016) uses population weights to adjust the mean comparison value.

Contrast coding is an orthogonal method where specific independent comparisons are determined apriori (see Cohen & Cohen, 1983). For example, suppose a researcher is interested in the differences between males and females of similar social positions. They may set up contrasts that compare “Black-male” with “Black-female’ and “White-male” with “White-female,” and so forth. However, there are limited cases where contrast coding is relevant because researchers need theoretical justification for the contrasts they employ.

Table 1

Example of Coding Procedures

	Additive Variable		Recoded Intersectional Variable		
	Race	Gender	Black-female	Black-male	white-female
Dummy Code	white	male	0	0	0
	Black	male	0	1	0
	white	female	0	0	1
	Black	female	1	0	0
Effect Code	white	male	-1	-1	-1
	Black	male	0	1	0
	white	female	0	0	1
	Black	female	1	0	0
Orthogonal Code	white	male	0	0	1
	Black	male	0	1	0
	white	female	0	0	-1
	Black	female	0	-1	0

Statistical Implications.

Multiple Comparisons. A categorical approach inevitably compares multiple variables. When we simultaneously compare multiple groups using the same data set, the probability of type 1 errors increases (see Shaffer, 1995). Therefore, when working with intersectional social positions and thus creating multiple groups, researchers need to be mindful of the type 1 error rate. One way to account for an inflated type 1 error rate is to adjust the p-value for each estimated covariate. A popular solution is a family-wise adjustment; Shaffer (1995) explains this method as treating each set of comparison groups as a family and dividing the intended alpha level by the number of members. For example, if a researcher investigates comparisons among six social position variables with an intended alpha level of .05, the new adjusted alpha level is .008.

Another option is the Benjamini & Hochberg "B-H" method which uses ordered p-values to control the false discovery rate (Benjamini & Hochberg, 1995). Russell et al. (2021) apply the

B-H procedure to an intersectional study of differential item functioning; after adjusting for multiple comparisons, they still report more differences with intersectional variables compared to an analysis of the additive components. Regardless of the method, researchers need to be aware of the issues of multiple comparisons, especially when looking at the intersection of three or more demographic variables.

Multilevel Models

Researchers use multilevel models to account for the nesting of individuals in a shared context, such as a school or a classroom. For the purpose of secondary data analysis, this can be referred to as a clustered dataset. Multilevel models are necessary because we assume that individuals within a given school are more similar to each other than they are to those from other schools.

The hierarchical model operates with multiple levels. The first level is the individual (students), and the second level is the shared context (schools). Models can have more than two levels if there are additional clustered contexts (states). Alternatively, they can be cross-classified within a level if multiple contexts interact at the same level (neighborhoods and schools). Multilevel modeling partitions the variance to demonstrate what the individual level (level 1) explains by accounting for what is explained at level 2 (Goldstein et al., 2010). Thus, in a hierarchical model, a researcher can simultaneously examine the main effects for each level and interactions across levels (Hoffman & Walters, 2022). Education researchers often work with datasets where students are clustered within schools, classrooms, districts, or other contexts. This research will consider each model in the context of clustered datasets. Therefore, multilevel models are appropriate, and both the conceptual and statistical fit with intersectionality is explored. The models previously explained, interaction and categorical, are modeled at level 1,

with educational clusters (schools) modeled at level 2. The MAIDHA model, as discussed in the next subsection, is designed as a hierarchical model, and for a clustered dataset, the school context will be cross-classified at level 2.

Conceptual Implications. Hierarchical modeling is conceptually relevant to intersectional thinking because social positions and the salience of identity depend on the context an individual operates within. Raudenbush (1989) draws on work from Bidwell & Kasarda (1980) to explain how schooling is a multilevel process because actors within the organizations determine the distributions of resources, including time, people, and materials. Educational structures encompass other sources of inequality. For example, the school someone attends is related to their neighborhood, which can often be connected back to policies created to segregate housing based on racialized identities.

Statistical Implications. In multilevel models, most researchers cite rules of thumb for sample sizes (Hox, 1998; Maas & Hox, 2004, 2005; or Keft, 1996). These researchers discuss sample size providing suggestions for the number of units at level 2 and the number of units within each level 2 unit. The earliest commonly cited rule of thumb is from Kreft (1996), who introduced the 30/30 rule, which is a minimum of 30 units at each level (i.e., 30 groups with 30 individuals in the group). Hox (1998) adds details to Kreft's (1996) suggestion by explaining that 30/30 is suitable for an investigation of the fixed effects for level 1 predictors. However, Hox (1998) explains when cross-level interactions are of interest, then a 50/20 rule is more appropriate, and when the focus is on level two fixed effects, a 100/10 rule is a better fit. Maas & Hox (2004, 2005) ran a series of simulation studies on the influence of sample size to provide further evidence for these "rules of thumb" commonly followed. They found that when there are ten groups with a sample size of 5, the fixed effects are not biased (but the level 2 variance is)

(Maas & Hox, 2004, 2005). Maas & Hox (2004, 2005) explain any group size under 100 will lead to bias in the standard error estimates such that they are too small, but a group size of 50 is still reasonable in practice. Clarke & Wheaton (2007) echo this 100/10 rule by examining conditions from 50 to 200 level 2 units and 2 to 20 level 1 unit per cluster. They found that following the 100/10 rules avoids most bias in the parameters and errors.

Recent research continues to back up these early findings on sample size in multilevel models. McNeish & Stapleton (2016) explored research findings regarding small cluster sizes in multilevel models and found that overall, models were most impacted by the number of level two units. McNeish & Stapleton (2016) found that across articles, fixed effects were the least affected by the number of clusters, and level 2 fixed effects (and cross-level interactions) tend to be overestimated when the number of clusters falls below 15 (e.g., Baldwin & Fellingham, 2013; Stegmueller 2013). When the number of clusters is small, researchers agree that the resulting standard errors will be downwardly biased (McNeish and Stapleton, 2016). Overall, these results show the importance of the number of level two units; with a low number of level two units, the model estimates are susceptible to underestimated standard errors and bias on the variance components. McNeish (2017) explains that when the standard error estimate is too small, the test statistic will be inflated, which will lead to p-values that are too small. This ends up inflating the Type-I error rates for the fixed effects (McNeish, 2017). Working with demographic data in education may offer greater complexity to the understanding of sample sizes. In education research, there are often uneven demographic categories per level two cluster, and they may be unevenly distributed.

MAIDHA

The MAIDHA model is a "multilevel analysis of individual heterogeneity and discriminatory accuracy." This approach to conducting intersectional analyses was introduced by Evans (2017) and coined by Merlo (2018). MAIDHA is a hierarchical model where individual respondents and their separable identities (Level 1) are nested within intersectional social strata (Level 2). From a data management perspective, intersectional social strata are created nearly the same way as the categorical variables such as "Black-female." However, the use of these variables is entirely different; MAIDHA clusters intersectional variables at the second level of a hierarchical model. Therefore, MAIDHA does require a traditional coding procedure and thus does not compare results to a reference category or value. MAIDHA can better promote the idea of within-group differences (in addition to between-group differences) due to the estimates that the hierarchical model provides (Evans, 2019; Merlo, 2014).

Typically, researchers nest individuals within level 2 when they share a tangible context that creates a similarity between them, such as a school, classroom, or neighborhood. Often, researchers do not consider demographics such as gender, race, or ethnicity to be something that is used to cluster. However, it is theoretically relevant to cluster on intersectional social strata, given that models require that error terms are not correlated. Within an intersectional paradigm, the forces of oppression are understood to be shared among members of an intersectional social position; each group can be assumed to have correlated errors. Thus, under intersectional theory, this is similar to the clustering structure of physical or structural contexts.

Cross-Classified Model. A cross-classified model uses multiple clustering variables at the same level, such as schools and neighborhoods, where there are different combinations of the two with which a participant may be associated. A cross-classified model is appropriate when

observations are nested in multiple contexts crossed at level 2 (Hoffman & Walters, 2022). The researcher will likely use a cross-classified model to apply the MAIDHA method in a dataset with natural clustering structures such as neighborhoods or schools (Evans, 2019c). For example, for a model with neighborhoods and schools, researchers would cross-classify the two contexts at level two, but in this case, social strata are cross-classified with the second contextual cluster. A cross-classified model is conceptually aligned with intersectionality theory because it allows for proxies for oppression to interact at the same level as the context in which an individual is positioned. When using a cross-classified model, Evans (2019c) explains that researchers should allow the social stratum effects to differ across the second clustering variable and the second clustering variable effects to be able to differ across social strata.

Conceptual Implications.

Interaction Effects. Multiple researchers claim that an advancement MAIDHA offers is the ability to estimate "interaction effects" using the stratum-level residual (level 2 variance) (Evans et al., 2018; Merlo, 2018). Evans et al. (2018) explain that interaction effects identify the extent to which the inclusion of social strata contributes to explaining the outcome variable beyond that of the additive model. An interaction effect can be interpreted as the extent to which the intersectional social positions account for the system of disadvantage that impacts the outcome variable above and beyond what is explained by an additive-only model. Evans et al. (2018) caution that this estimate cannot be a direct measure of the influence of "intersectionality" because interpreting the interaction effects in this way requires the assumption of no-omitted variable bias; thus, all relevant axes of identity must be in the model (Evans et al., 2018). All axes of identity which are salient towards the intersectional experience in the research context would need to be included, and this may prove difficult in fields where the theory on

intersectional identities and their relation to outcomes is still in development. Despite this limitation, the interaction terms have been the central focus when interpreting MAIDHA results (e.g., Evans, 2019a, 2019b; Evans et al., 2018; Merlo, 2018). Merlo et al. (2018) detail how this estimate can aid decision-making in the public health sector. Evans et al. (2018) argue that it is ideal in their own public health research context to "determine simultaneously whether all intersectional identities exhibit evidence of an interaction (or intersectional) effect" (p. 65). However, this claim and what interaction effects can and cannot explain is a point of contention in the literature.

Lizotte et al. (2020) present a critique of the MAIDHA approach with a specific focus on the utility of the interaction effects. Lizotte et al. (2020) argue that the so-called interaction effects are not interpretable, and thus, this approach lacks utility. They explain that the stratum-level residuals in the model cannot be interpreted as interaction effects because "the fixed effects in MAIHDA do not represent population average effects; rather, they reflect effects under an implicit re-weighting of the data given all intersections are of equal size" (Lizotte et al., 2020, p. 4). This is due to the fact that the individual demographic identity variables in level 1 (fixed effects) also determine the social strata membership that a person is in at level 2, leading to "over-adjusting where it is not possible to know whether to attribute a difference in mean outcome to a difference in group membership or to individual-level social position effects" (Lizotte et al., 2020 p. 5). Further, the estimates of stratum residuals are conflated due to unmodeled interaction terms between the demographic identities at level 1. Lizotte et al. (2020) conclude that the emphasis placed on interaction effects is not of value, and therefore MAIDHA does not provide additional information beyond a categorical model. In their rebuttal, (Evans et al., 2020) clarify that the fixed effects estimates are precision-weighted grand means. While they

acknowledge the limitation of no omitted variable bias as something not only present in MAIDHA but all multilevel models, they do not provide a rationale for the omission of interaction terms. They explain the language around the interaction effect in that it "tells us whether the stratum is advantaged or disadvantaged relative to what might have been predicted for it based on additive effects alone" (Evans et al., 2020, p 6). Overall, Evans et al. (2020) argue for the use of MAIDHA and discuss the value of its use in intersectional analysis. This discussion demonstrates that defining what MAIDHA is and is not capable of – specifically, what the interaction effects can tell us— remains disputed.

In addition to ongoing scholarly discussions, I argue that interaction effects are not fully aligned with intersectional thinking. The interaction effects, as described in the literature, are a way of testing for the relevance of intersectionality in a given context. But, applying an intersectional lens to research requires the assumption that intersectional oppression exists. Evans (2018) acknowledges this limitation, "that our framing of quantitative, intercategorical intersectionality falls into what some scholars have called 'intersectionality as testable explanation' (Hancock, 2013), in that it involves an assessment of whether statistically significant interaction effects are detectable." (p. 66). Evans (2018) further explains that this approach is intended to be exploratory and does not require hypothesis testing of interaction effects. Therefore, researchers need to be careful with their interpretation of the interaction effect so that it does not lead to a narrative of testing for intersectionality. Applied through an intersectional lens, one may interpret whether the oppression experienced has a meaningful impact on the outcome variable, but this does not negate the fact that intersectional groups exist. Therefore, researchers must be clear that this is not a test of whether intersectional groups exist

but rather a test of the extent to which oppression and advantage experienced by the intersectional group is salient in explaining a given research context.

Statistical Implications.

Model Fit. The multilevel structure of MAIDHA presents advantages for model parsimony. In MAIDHA, additional demographic categories increase the number of terms in the model linearly instead of geometrically (Evans et al., 2018). In an exploration of model fit, Evans et al. (2018) found that the number of parameters added to the MAIDHA model did not increase the Bayesian Information Criterion (BIC). MAIDHA remains robust even when additional intersectional categories are added; thus, the model is scalable.

Multiple Comparisons. MAIDHA may also be more robust to issues of multiple comparisons. Evans (2019a) explains that multilevel models have greater precision with weighted estimation and borrow strength due to the inclusion of random effects. Through this approach, Evans (2019a) explains that MAIDHA models "automatically adjust estimates for social strata based on the number of respondents at those intersections, down-weighting extreme estimates based on too few respondents and therefore providing more reliable (if conservative) estimates" (p. 96).

However, Bell et al. (2019) argue that this is only applicable when the level 2 residual terms are independent and identically distributed. However, in this approach, that assumption may not be met because, depending on the context, the salience of axes of identities may vary. While intersectionality does not test the saliency of each axes of identity, if one particular axes is salient while no others in the model are, that salience will influence the ability for model shrinkage to occur. For example, if gender is a salient aspect in predicting a given outcome, then

> the ‘male’ intersections will be more similar to each other than to the ‘female’ intersections…. This would have the effect of not only making those intersections appear different (indeed, we would want them to do this), but also affect the estimates of other residuals because shrinkage would be incorrectly applied. (Bell et al., 2019, p. 89)

Bell et al. (2019) substantiate this argument with a Monte Carlo simulation that assesses the influence of related residuals. By comparing the accuracy in residual statistical significance, they found there were some benefits in the statistical shrinkage to increase model accuracy compared to a main-effects approach, but it was still less than desired (Bell et al., 2019).

To mitigate the influence of the broken assumption, Bell et al. (2019) suggest applying an iterative approach by first adding two-way interactions, then three-way. Bell et al. (2019) recommend statistically comparing the amount of level-2 variation explained with model comparison statistics such as the Deviance Information Criterion (DIC). A reduction of level 2 variance approaching zero (as seen by Jones et al., 2016) indicates that adding additional axes of identity to the model will not account for further variance explained in the outcome. Bell et al. (2019) explain, “by seeing how the level-2 variance decreases as increasing orders of interactions are included, it would be possible to see how ‘deep’ the intersectionality goes – whether it is the result of two variables interacting, three, or more” (p. 95). Thus, the intersections included are the most salient axes of oppression related to the outcome of interest, and additional intersections would not further increase the variance explained by the model. Therefore, this would represent the depth of intersecting axes of oppression that may be relevant to explaining advantage and disadvantage for the given outcome. This approach can be used to complement the axes of identity a researcher has already determined may be salient in their context, based on theory.

Interpreting Intersectional Experiences. Since MAIDHA uses intersectional identities as the clustering variable, the analysis fundamentally changes how intersectional identities are explored. Unlike interaction and categorical models, intersectional identities are not entered in as covariates. Instead, to understand the differences between intersectional strata, researchers must examine the magnitude of the level-2 intercepts. Instead of comparing the coefficients for the covariates, the slope intercepts are examined to estimate the strata's relationship with the outcome variable.

Sample Size. Depending on the goal of the analysis, research on multilevel modeling suggests a minimum of 30 level two units to produce accurate estimates. This means there should be a minimum of 30 intersectional social strata. However, for a researcher whose theory has led them to analyze a race-gender intersection, where they have 2 categories of gender identity and 5-8 categories of racialized identity, they may only end up with 10-16 social strata (level 2 units). Therefore, this approach may be limited to research contexts where it is theoretically relevant to explore a high number of social strata.

In addition, the multilevel structure requires sufficient cell sizes to make accurate estimates. As previously explained, the model operates best with a relatively even sample distribution. Evans (2019b) suggests that each cell contains at least 20 observations; however, she does not provide a rationale for that number. Given that this is a newer method, additional research is needed to better understand the advantages and limitations of this modeling approach (Evans et al., 2020; Lizotte et al., 2020). Until then, the scholarly debate will continue regarding the best approaches for model building, model use, and model interpretations.

Empirical Comparisons of Methods

Conceptually, the literature does not point to any one method as being more advantageous than others when conducting quantitative analyses through the lens of intersectionality; each method has both advantages and disadvantages. The approach used to answer a given research question depends on how the researcher aims to balance statistical constraints with the issues they wish to explore with the data. While the use of intersectionality in quantitative methods is still emerging, several researchers have empirically compared methods to offer a deeper understanding of the statistical constraints of each method.

Claire Evans (2019a) provides a comparative study that demonstrates empirical differences between an interaction model and a MAIDHA approach. The main question of interest in her study was whether MAIDHA produced fewer statistically significant results compared to an interaction model. Evans (2019a) sought to explore the implications of adding more dimensions of social identity, such as including gender, race, or class in a model, and how those additions influence statistical significance. Using the Add Health dataset, she compares the results of several outcome variables for MAIDHA and the conventional interaction model. When comparing models, Evans (2019a) found that MAIDHA was less likely to reveal interactions (when examining at the interaction effect) compared to the interaction model (when examining significant covariate interactions). Overall, Evans (2019a) found that when additional dimensions of identity were added, formally unseen intersectional experiences were uncovered. Evans (2019a) acknowledges that this fits the theoretical narrative of intersectionality. However, researchers must still be careful about employing too many dimensions because type 1 error inflation may occur due to multiple comparisons.

Mayuri Mahendran and her colleagues have published the first two simulation studies on empirically evaluating the accuracy of intercategorical data analysis approaches that account for intersectionality (Mahendran et al., 2022b, 2022a). Mahendran et al. (2022a) focus on binary outcomes, and Mahendran et al. (2022b) explore continuous outcomes. In these studies, they investigate several methods, including the interaction model, categorical model, and MAIDHA. For both binary and continuous outcome variables, they found that at large sample sizes (n > 50,000), all three methods could accurately estimate the outcome, and MAIDHA performed better than the other methods at smaller sample sizes (2,000 to 5,000).

When applied to a real dataset, Mahendran et al. (2022a) found that the categorical model (which they label the main effects model) and the MAIDHA model produced very different estimates.

> For example, for white female respondents aged 18 to 39 with poverty-level income, the estimated prevalence varied from 7% to 20%. Comparing the two best performing methods at smaller samples from the simulation, MAIHDA and main effects, the estimated prevalence were also different. For example, among Black male respondents age 60+ with non-poverty income, main effects estimated 76.5% while MAIHDA estimated 65.3%. For female Hispanic respondents age 60+ with poverty-level income, main effects estimated 50.5% while MAIHDA estimated 60.3%. (Mahendran et al., 2022a, p 8)

Mahendran et al. (2022b) suggest that MAIDHA is more accurate. This may demonstrate how MAIDHA can surpass the accuracy of the other methods at large sample sizes due to the ability of multi-level shrinkage.

While there have been several studies that compare the models, they do not provide all of the information necessary to make conclusions about the statistical utility of each model. First, Evans' (2018) study only compares two models (categorical and MAIDA) on a singular dataset. This is helpful to show how the models present themselves differently, but the results are limited to the constraints of that dataset. The studies by Mahendran et al. (2022a, 2022b) were useful for comparing overall model accuracy to each other and demonstrating the capabilities of each model. However, they did not vary the sample size within each cluster. Therefore, their design does not necessarily reflect the complexity that researchers accounting for intersectionality in education may face.

Education researchers often grapple with starkly unequal distributions of sample sizes across groups. However, within intersectional methods, researchers have yet to investigate the proportion of representation and variance within single-axis categories. In addition, each of these comparisons has been in the context of a single-level dataset. It is possible that the introduction of clustered data influences the accuracy of model parameters as well as other outcomes. It is necessary to explore the implications of uneven sample sizes within clustered intersectional models. While there have been a number of studies that focus on the influence of sample size in multilevel models, no study has yet to take into account the unique features that intersectionality introduces into a model.

A comparison of intersectional methods in a clustered data setting is necessary for researchers to determine the most appropriate model given their design and data. Therefore, this dissertation seeks to extend research on how methods of accounting for intersectional analyses handle complex demographic data (and thus sample size issues) in a clustered data context. This research considers uneven sample sizes, high variance, and overall model complexity in a series

of simulations. The purpose of this dissertation is to contribute to our knowledge of the statistical advancements and limitations of multilevel models.

Chapter 3: Methods

This chapter details the methodological choices made for this dissertation. The study's objectives were to evaluate how each of the three methods of modeling intersectional analyses (1. Interaction, 2. Categorical, and 3. MAIDHA) perform, given various complexities in demographic data characteristics. In the section titled "Objectives," I present the research questions, discuss the study's goals, and provide a general overview of the simulation design. Then, in "Simulation Procedures," I outline the simulation design, including the data characteristics used to form the scenarios, data generation procedures, and simulation study outcomes for evaluating performance in a given scenario. Next, in "Methods of Modeling Intersectional Analysis," I provide the statistical components of the models investigated. Finally, in “Reporting of Results & Proposed Comparative Analysis,” I describe the approaches used to analyze the statistical characteristics associated with each method under each data condition, as well as summarize findings across all analyses.

Objectives

Demographic data in education often have uneven sample sizes and notable differences in variability within sub-groups. It is imperative to understand how intersectional models both perform and compare to each other under different complexities in demographic data. In Chapter 2, I described three methods of modeling intersectional analyses. These methods of modeling include: 1) Interaction, 2) Categorical, and 3) MAIDHA. This simulation study aims to understand how these three methods of modeling intersectional analyses function under various demographic data characteristics.

The research questions addressed in this dissertation were:

1. What are each model's statistical advantages and disadvantages under different demographic data characteristics?
2. In what ways does each model perform differently from one another under each demographic data characteristic scenario?

Where, the demographic data characteristics are a) the number of demographic categories; b) the proportional representation of identity indicators within demographic categories of the total sample; c) the within-intersectional-group variance; and d) the total sample size.

A series of Monte Carlo simulations were conducted to generate data sets that were used to compare the performance of each method of modeling intersectional analyses under different conditions, each of which was designed to mimic the complexity of demographic data in education. Scenarios were created based on a combination of each of four demographic data characteristics. There were 54 unique scenarios that were defined by a combination of four demographic data characteristics. Across all repetitions, this study simulated a hierarchical data context where students were evenly clustered in groups designed to represent schools.

For each of the 54 different scenarios, 1000 datasets were generated based on true analytic parameters for each intersectional group, yielding a total of 54,000 datasets. For each dataset, three analytic methods were applied. This yielded a total of 162 combinations of methods and scenarios, and thus a total of 162,000 models were built. For each of the 162 combinations of methods and scenarios, five simulation outcomes were estimated: bias, accuracy (mean square error), type 1 error, power, and coverage. Model fit through Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) was also retained.

Simulation Procedures

This simulation study examined how demographic data characteristics influence estimates, designed to be proxies of intersectional oppression, provided by three methods of modeling intersectional analyses. In this section, I first describe the scenarios designed to reflect various conditions of complex demographic data. Next, I explain the procedures for generating true values and datasets. Finally, I describe the outcomes of the simulation study and the saved values used to compare models within and across scenarios.

"Demographic" categories were created to mimic: A) racialized and ethnic identities; B) gender identity; and C) financial status, presented in Table 2. Since the data employed for this study was simulated and this research did not examine actual relationships among the three demographic categories and/or various intersectional groupings, these three identity categories are henceforth referred to as A, B, and C. Each category is represented by separate variables, annotated using the notation of a letter to represent a given demographic category and a number to represent an identity indicator within that category. As an example, B1 is used to represent demographic category B (e.g., gender identity) and identity indicator 1 (e.g., female). For scenarios with two demographic categories, only A and B were used to build datasets and model results, whereas for scenarios with three demographic categories C was used as well.

It is important to note that these proxy demographic categories are imperfect. For example, in writing this I recognize that six categories of racialized identity do not capture every identity, there are more than two gender identities, and the differences between financial status's may be arbitrary. The specific choices of categories were chosen based off the current IPEDS data collection, to represent what may be "typically" collected in an education context. However, these categories have several limitations, and researchers should critically consider what identity

indicators are theoretically relevant to the issue(s) they are studying and for which categories they are able to collect sufficient information on when designing an intersectional analysis.

Table 2

Demographic Variable Notation

Category	Notation
A. Racialized and Ethnic Identities	A0*
	A1
	A2
	A3
	A4
	A5
	A6
B. Gender Identity	B0*
	B1
C. Financial Status	C0*
	C1
	C2

Note. * indicates the reference category for dummy-coded variables.

Simulation Scenarios

The simulation scenarios were designed to reflect variation in four characteristics of demographic data, including the number of demographic groups, the proportion of the sample represented by each demographic group, the within-intersectional group variance in the outcome variable, and the total sample size. Each scenario and its conditions are summarized in Table 3.

Table 3

Scenario Conditions

Scenario Characteristic	Description	Additional Details
Number of Demographic Categories	**1.** Two Categories (A and B) **2.** Three Categories (A, B, and C)	1. A (seven categories), and B (two categories) 2. A (seven categories), B (two categories), and C (three categories)
The proportional representation of identity indicators within demographic categories	**1.** P1: Even **2.** P2: Imbalanced **3.** P3: Extremely Imbalanced	The ratios of the sample in each identity indicator are presented in Table 4
Within Category Variance in the Outcome Variable	**1.** Small: Little to no variance for all intersectional groups **2.** Large: High variance for all intersectional groups **3.** Mixed: Mixed variance	Where high variance is (σ^2 = 5.0). Mixed variance was set so 10% of the groups had high variance Displayed in Table 5
Overall Sample Size	**1.** Small sample **2.** Medium size sample **3.** Large sample	**1.** 5,000 **2.** 10,000 **3.** 20,000

Condition 1: Number of Demographic Categories. Intersectionality theory emphasizes the importance of employing theory to inform the relevance of specific intersectional groups in an analysis. Thus, the inclusion of an intersectional social position in an analysis must be driven by theory. Many educational researchers theorize that lived experiences of students differ across gender, racialized identity, and class/economic status. As a result, these three demographic

categories are often included in analyses of educational experiences. Given this practice, this study models two sets of demographic data, one based on two demographic categories (A and B), and another based on three demographic categories (A, B, and C). Moreover, the number of identity indicators within each category differs, such that category A has seven identity indicators, category B has two identity indicators, and category C has three identity indicators.

Condition 2: Demographic Data Representation. There were three variations of demographic data representation across identity indicators. The first condition reflects an even representation of identity indicators, where the proportional representation of each identity indicator within a demographic category is equal. The second condition models relatively imbalanced identity indicator distributions, which were loosely based on 2021-2022 IPEDS data on college enrollment (U.S. Department of Education, 2021). After collecting the IPEDS proportions, I adjusted the lowest possible proportions to be .050 to provide representation of categories during this study. Finally, in the third condition, extremely unbalanced representation was designed such that some identity indicators have a relatively high proportional representation (as high as .80), and others had a relatively low proportional representation (as low as .005). Table 4 presents the breakdowns for the distribution of identity indicators within each demographic category across each of the three conditions. Categorical intersectional variables and intersectional social strata were formed from these additive variables. Thus, the proportion for a participant in intersectional group A1/B1/C1 would be equivalent to the three individual proportions multiplied, where, in a relatively imbalanced scenario, this is 0.050*.590*.467= 0.014 proportion of respondents out of the total sample.

Table 4

Distribution of Sample Size within Each Dummy-Coded Demographic Variable

Identity Indicator	Even	Relatively Imbalanced	Extreme
A0	0.143	0.432	0.630
A1	0.143	0.050	0.050
A2	0.143	0.073	0.110
A3	0.143	0.137	0.050
A4	0.143	0.208	0.050
A5	0.143	0.050	0.055
A6	0.143	0.050	0.055
B0	0.500	0.410	0.200
B1	0.500	0.590	0.800
C0	0.333	0.145	0.100
C1	0.333	0.467	0.700
C2	0.333	0.388	0.200

Condition 3: Within Category Variance. Often in educational research, the datasets we work with have some intersectional groups with very little variance in an outcome variable, while others have a large variance. To reflect this variation, the variance within intersectional groups was manipulated in three ways. In one set of scenarios, each intersectional group was designed to contain a small amount of naturally occurring variance. In a second set of scenarios, each intersectional group was simulated to contain high variance ($\boldsymbol{\sigma^2}$ = 5.0). In the third set of scenarios, the amount of within-category variance differed where 10% of the unique

combinations of intersectional group and school ID were randomly selected to have high variance ($\boldsymbol{\sigma^2}$ = 2.0), listed in table 5. All other groups had a small amount of (unchanged) variance.

Table 5

Intersectional Groups with High Variance for the Mixed Variance Condition

	Intersectional Group	Designed Difference
Two Demographic Categories	A1B1	Negative
	A3B1	No difference
Three Demographic Categories	A2B0C0	Positive
	A6B1C1	Negative
	A6B1C2	Negative
	A0B0C1	No difference
	A1B0C1	No difference
	A4B1C2	No difference

Condition 4: Sample Size. Three sizes are simulated: 5,000, 10,000, and 20,000.

Models

This section describes the statistical components for each of the three methods of modeling intersectional analyses examined in this study; a conceptual overview of each method is presented in Chapter 2, under the section titled "Conceptual and Statistical Limitations and Advancements." Hierarchical linear models were developed for each method to examine their functioning within a clustered data structure.

I first describe the null model for each method before presenting the full model. The unconditional model predicts the outcome variable for a given "school" cluster without adding

any predictor variables. Typically, an unconditional model is used to estimate the interclass correlation coefficient (ICC), which is estimated as $\frac{\tau_{00}}{\tau_{00}+\sigma^2}$, where τ_{00} is the estimated between-group variance and σ^2 is the estimated within-group variance. Categorical and interaction models have the same unconditional model and thus are described together. The MAIDHA model has a different unconditional model due to its cross-classified structure.

Figure 2

Unconditional Model for Interaction and Categorical Models

Level one

$$Y_{ij} = \beta_{0j} + r_{ij}$$

Level two

$$\beta_{0j} = y_{00} + u_{oj}$$

Mixed model

$$Y_{ij} = y_{00} + u_{0j} + r_{ij}$$

In this model, y_{00} is the fixed component of the model, the predicted grand mean of the outcome variable, $u_{0j}$is a random level-2 effect (for school *j*), B_{0j} is the intercept, and r_{ij} is the error term.

Model 1: Interaction

In the interaction model, the dummy coded identity indicator variables for each demographic category were multiplied to produce interaction terms representing unique intersectional groupings of demographic categories. For example, the dummy coded variable representing female gender identity was multiplied by each of the five dummy coded variables representing each racialized identity indicator to produce five interaction terms that represent the intersection of gender identity and each indicator of racialized identity. The number of terms

(dummy-coded demographic variables and interaction terms) varied based on the scenario, as described under simulation conditions. Given the intent to vary the number of demographic categories (and thus the number of variables), the models had either 6 or 12 interaction terms. Including the individual additive identity indicator variables entered in separately, this approach yields models containing a total of 13 or 21 terms.

Figure 3

Interaction Model

Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij}) + \cdots \beta_{kj}(X_{ij}) + \beta_{1j}(X_{ij}) * \beta_{kj}(X_{ij}) + r_{ij}$$

Level 2

$$\beta_{0j} = y_{00} + u_{0j}$$

$$\beta_{1j\ldots kj} = \mathrm{y}_{10\ldots \mathrm{k}0}$$

Mixed Model

$$Y_{ij} = \beta_{0j} + Y_{10}(X_{ij}) + \cdots Y_{k0}(X_{ij}) + Y_{k0}(X_{ij}) * (X_{ij}) + u_{0j}$$

Where k is the number of dummy-coded demographic variables in the model.

Model 2: Categorical

In the categorical model, intersectional groupings (e.g., Black-female, white-male, etc.) were created *a priori*. Rather than entering single-axis demographic variables and interaction terms, the intersectional groupings were entered directly into the model. As a result, depending on the demographic data scenario, there were either 13 or 41 categorical intersectional variables entered into the model.

Figure 4

Categorical Model

Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij}) + \cdots \beta_{kj}(X_{ij}) + r_{ij}$$

Level 2

$$\beta_{0j} = y_{00} + u_{0j}$$

$$\beta_{1j\ldots kj} = y_{10\ldots k0}$$

Mixed Model

$$Y_{ij} = \beta_{0j} + Y_{10}(X_{ij}) + \cdots Y_{k0}(X_{ij})$$

$$+u_{0j}$$

Where k is the number of intersectional group variables in the model

Model 3: MAIDHA (Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy)

Figure 5

MAIDHA Unconditional Model.

Level 1

$$Y_{i(j1,j2)} = \beta_{0(j1,j2)} + r_{i(j1,j2)}$$

Level 2

$$\beta_{0(j1,j2)} = \gamma_{000} + \mu_{0j1} + \nu_{0j2}$$

Mixed Model

$$Y_{i(jk)} = \gamma_{000} + +\mu_{0j1} + \nu_{0j2} + r_{i(j1,j2)}$$

MAIDHA presents a cross-classified model where schools and intersectional strata were both modeled at level two. The intersectional strata were an additional level two clustering variable used to indicate the unique intersection of identities. An additional subscript is used where j_1 and j_2 identify the cross-classified factors (j_1 intersectional strata and j_2 school) to account for the cross-classification of two level two clusters. The parentheses around the subscripts are separated by a comma (j_1,j_2) to indicate that these two factors operate conceptually at the same level (Hox, 1998).

In the cross-classified model, γ_{000} is the grand mean for the outcome variable $y_{i(j1,j2)}$ which represents the outcome for student *i* in intersectional strata j_1 and school j_2. $\beta_{0(j1,j2)}$ represents the intercept for the predicted outcome for students from the specific combination of intersectional social strata j and school (predicted cell means); we assume that the intercept varies randomly across j_1 and j_2. $r_{i(j1,j2)}$ represents the individual residual, the deviation of a student's score from the students' strata, and the school-predicted intercept value.

μ_{0j1} is the residual error term for intersectional strata, and ν_{0j2} is the residual error term for schools. $\mu_{0(j1,j2)}$ is the random intercept effect and is the residual beyond that predicted by the grand mean, and the two main effects.

The following equations represent the full MAIDHA model. In this model, the main effects at level 1 are represented by individual additive identities (i.e., race/ ethnicity, gender, and class), and the level two intercepts are each intersectional social strata, where *k* at level one is the number of demographic covariates and *k* at level two is the number of intersectional strata. The models had either 7 demographic identity indicator variables at level one clustered into 14 intersectional social strata, or 9 demographic identity indicator variables at level one clustered into 42 intersectional social strata.

Figure 6

MAIDHA Model

Level 1

$$Y_{i(j1,2)} = \beta_{0(j1,j2)} + \beta_{1(j1,j2)}\left(X_{i(j1,j2)}\right) + \cdots \beta_{k(j1,j2)}\left(X_{i(j1,j2)}\right) + r_{i(j1,j2)}$$

Level 2

$$\beta_{0(j1,j2)} = \gamma_{000} + \mu_{0j1} + \nu_{j2}$$

$\beta_{1(j1,j2)}=\gamma_{100}$
$\beta_{k(j1,j2)}=\gamma_{k00}\cdots$

Mixed Model

$$Y_{i(j1,j2)} = \gamma_{000} + Y_{10}\left(X_{i(j1,j2)}\right) + \cdots + Y_{k0}\left(X_{i(j1,j2)}\right) + \mu_{0(j1,j2)} + \nu_{0(j1,j2)}$$

Procedures for Generating Datasets

This study used a Monte Carlo simulation design. All simulations of datasets and subsequent analyses were conducted using R. To generate datasets, a clustered data structure was built, "true" values were selected for coefficients, and the "true" outcome was generated. Then, independent datasets, replicated 1,000 times, were generated for each scenario. Each dataset was built using the parameters from the true coefficient values for each intersectional group and the true outcome generation formula. A random number generator was used to set the seed when generating each simulated data set.

Clustered Data Structure. Clustered data were simulated to represent a structure similar to students nested in schools. One hundred school clusters were simulated with relatively even distributions of cases within each cluster. School clusters were designed such that the intraclass correlation coefficient was at least 0.10. The random intercept variance was set as 0.25, and the residual variance was 1 in order to obtain an ICC close to 0.20.

Coefficient "Truth" Generation. Intersectional groups were designed to have what I refer to as a "true difference" or "no true difference." A true difference refers to the fact that in the population, *x* intersectional group has a true mean difference on the outcome variable compared to the reference category, after accounting for the variance from all other intersectional groups. Then, for groups with no true differences, the population level difference from the reference group is 0. For the sake of brevity, henceforth, I will refer to the former as simply as "true difference" and the latter as "no true difference."

There were two versions of true coefficient values generated, one for scenarios that use two categories and another for scenarios that use three categories. These true values were generated in comparison to the reference category (A0B0 or A0B0C0). To obtain true coefficient values, I used a random number generator to select a starting seed for each intersectional group designated to have a true difference. Then, I determined the true coefficient value for each of these intersectional groups by randomly selecting a value from a distribution with a range of + or – [0.20 to 2.00]. These true coefficient values were then used to generate the true outcome in the population, as described below.

These computed values represent the level 1 main effect truths for the intersectional coefficients in the categorical and interaction models. Since the members represented by an interaction term are identical to those represented by a categorical term, the interaction and categorical terms share the same true coefficient value. For example, data that is coded so those who are A1 = 1 and those who are B2 = 1 are classified as A1B2, which is equivalent to the A1 and B1 interaction where A1*B2 = 1. The interaction model does not estimate coefficients for all intersectional groupings; intersectional groupings with "0" for any letter were excluded from the model.

In MAIDHA, the intersections examined were the level two intercepts. The true values for these intercepts were computed after outcome generation for the true dataset. The outcome was summarized across intersectional strata averages, where the average is the true value for the level two intercept. Tables 6 and 7 specify the designed difference for each coefficient, the randomly generated true coefficient values for the interaction and categorical model level 1 main effects, and the computed level 2 intercept truths for MAIDHA.

Table 6

Coefficient Truths for Scenarios with Two Categories

Intersectional Group	Designed Difference	Truth – Level 1 Main Effects	Truth — Level 2 Intercepts for MAIDHA
A0B0	No difference	0	0.827
A1B0	Positive	0.779	1.686
A2B0	Negative	-0.899	0.029
A3B0	Negative	-0.639	0.503
A4B0	No difference	0	0.990
A5B0	No difference	0	1.062
A6B0	No difference	0	0.937
A0B1	No difference	0	0.792
A1B1	Negative	-1.159	-0.097
A2B1	Positive	0.936	2.044
A3B1	No difference	0	1.106
A4B1	No difference	0	1.012
A5B1	No difference	0	1.064
A6B1	No difference	0	0.998

Table 7

Coefficient Truths for Scenarios with Three Categories

Intersectional Group	Designed Difference	True Values – Level 1 Intersectional Group Coefficients	True Values — Level 2 Intercepts for MAIDHA
A0B0C0	No difference	0	0.291
A1B0C0	Positive	1.639	1.559
A2B0C0	Positive	1.314	0.981
A3B0C0	Negative	-0.956	-1.623
A4B0C0	No difference	0	-0.029
A5B0C0	No difference	0	-0.016
A6B0C0	No difference	0	0.182
A0B1C0	No difference	0	-0.036
A1B1C0	No difference	0	-0.128
A2B1C0	Positive	1.844	1.702
A3B1C0	No difference	0	-0.161
A4B1C0	No difference	0	-0.375
A5B1C0	No difference	0	--0.161
A6B1C0	No difference	0	-0.114
A0B0C1	No difference	0	-0.259
A1B0C1	No difference	0	0.011
A2B0C1	Positive	0.953	0.950
A3B0C1	Negative	-0.957	-1.026

A4B0C1	No difference	0	-0.176
A5B0C1	No difference	0	0.029
A6B0C1	No difference	0	-0.070
A0B0C2	No difference	0	-0.168
A1B0C2	No difference	0	-0.111
A2B0C2	No difference	0	-0.136
A3B0C2	No difference	0	-0.141
A4B0C2	No difference	0	-0.158
A5B0C2	Negative	-0.521	-0.546
A6B0C2	No difference	0	0.028
A0B1C1	No difference	0	-0.085
A0B1C2	No difference	0	-0.063
A1B1C1	No difference	0	-0.249
A2B1C1	Positive	1.666	1.544
A3B1C1	No difference	0	-0.023
A4B1C1	No difference	0	-0.087
A5B1C1	No difference	0	0.051
A6B1C1	Negative	-1.724	-1.924
A1B1C2	Negative	-1.344	-1.251
A2B1C2	No difference	0	0.006
A3B1C2	Negative	-0.610	-0.729
A4B1C2	No difference	0	-0.126

A5B1C2	No difference	0	0.051
A6B1C2	Negative	-1.650	-1.702

True Outcome Generation. Two formulas provided the underlying process of generating the outcome variable based on the number of demographic categories included. A researcher should choose the number of categories they include according to what they theorize to be salient intersectional groups given the outcome of interest and the research context. Thus, I chose to have two different formulas in order to represent the variations in truth for models with two categories and models with three categories.

The two models of outcome generation were built from the categorical model. I chose to use this model to generate the true outcome variable as it does not include separate components of identity, only intersectional terms. I believe the categorical representation of intersectional social locations is best aligned with intersectional theory, compared to the other methods. Because the outcome variable is generated from a categorical model, it is likely that the categorical model analyses yield greater accuracy in estimated coefficients since there is a direct alignment between the truth simulation outcome generation method and the statistical modeling method. I attempted to mitigate this conflation by examining multiple outcomes to make conclusions about the statistical constraints of each model under various conditions. I chose to use a consistent formula for generating the true outcome variable for all three methods of modeling intersectional analyses so that I could directly compare results across modeling methods.

The outcome variable was generated using a two-level model with observations nested in groups. At the first level, the outcome variable was generated as a linear function of 54

intersectional groups. The second level nested the groups into schools, but no additional predictors were included. An error term was generated to be randomly distributed with a mean of 0 and a standard deviation of 1. The outcome variable was assumed to be normally distributed, and the random effects were assumed to be independent. The R code for the truth generation is available online (https://github.com/oszendey/Intersectional-Analyses).

Figure 7

Model 1: Two Demographic Categories

$$\begin{aligned} Y_{ij} = \beta_{0j} &+ Y_{10}(A1B0_{ij}) + Y_{20}(A2B0_{ij}) + Y_{30}(A3B0_{ij}) + Y_{40}(A4B0_{ij}) + Y_{50}(A5B0_{ij}) \\ &+ Y_{60}(A6B0_{ij}) + Y_{70}(A0B1_{ij}) + Y_{80}(A1B1_{ij}) + Y_{90}(A2B1_{ij}) + Y_{100}(A3B1_{ij}) \\ &+ Y_{110}(A4B1_{ij}) + Y_{120}(A5B1_{ij}) + Y_{130}(A6B1_{ij}) + \mu_{0j} \end{aligned}$$

Figure 8

Model 2: Three Demographic Categories

$$\begin{aligned} Y_{ij} = \beta_{0j} &+ Y_{10}(A1B0C0_{ij}) + Y_{20}(A2B0C0_{ij}) + Y_{30}(A3B0C0_{ij}) + Y_{40}(A4B0C0_{ij}) \\ &+ Y_{50}(A5B0C0_{ij}) + Y_{60}(A6B0C0_{ij}) + Y_{70}(A0B1C0_{ij}) + Y_{80}(A1B1C0_{ij}) \\ &+ Y_{90}(A2B1C0_{ij}) + Y_{100}(A3B1C0_{ij}) + Y_{110}(A4B1C0_{ij}) + Y_{120}(A5B1C0_{ij}) \\ &+ Y_{130}(A6B1C0_{ij}) + Y_{140}(A0B0C0_{ij}) + +Y_{150}(A1B0C1_{ij}) + Y_{160}(A2B0C1_{ij}) \\ &+ Y_{170}(A3B0C1_{ij}) + Y_{180}(A4B0C1_{ij}) + Y_{190}(A5B0C1_{ij}) + Y_{200}(A6B0C1_{ij}) \\ &+ Y_{210}(A0B1C1_{ij}) + Y_{220}(A1B1C1_{ij}) + Y_{230}(A2B1C1_{ij}) + Y_{240}(A3B0C1_{ij}) \\ &+ Y_{250}(A4B0C1_{ij}) + Y_{260}(A5B0C1_{ij}) + Y_{270}(A6B0C1_{ij}) + Y_{280}(A0B0C2_{ij}) \\ &+ Y_{290}(A1B0C2_{ij}) + Y_{300}(A2B0C2_{ij}) + Y_{310}(A3B0C2_{ij}) + Y_{320}(A4B0C2_{ij}) \\ &+ Y_{330}(A5B0C2_{ij}) + Y_{340}(A6B0C2_{ij}) + Y_{350}(A0B1C2_{ij}) + Y_{360}(A1B1C2_{ij}) \\ &+ Y_{370}(A2B1C2_{ij}) + Y_{380}(A3B1C2_{ij}) + Y_{390}(A4B1C2_{ij}) + Y_{400}(A5B1C2_{ij}) \\ &+ Y_{410}(A6B1C2_{ij}) + \mu_{0j} \end{aligned}$$

Procedures for Fitting Models and Generating Outcomes

Every dataset was analyzed using each of the three methods of modeling intersectional analyses. This approach allowed the three methods to be compared with respect to how they perform for each scenario because they all use the same dataset for a given scenario and

iteration. Each model used restricted maximum likelihood estimation (REML) (Mason et al., 1983; Raudenbush & Bryk, 1986). REML is appropriate for complex data structures because the restricted maximum likelihood estimates are adjusted for the fixed effects' uncertainty (Raudenbush et al., 1991). While MAIDHA was first designed with Bayes estimates, Mahendran et al. (2022b) demonstrated that it produces similar estimates as frequentist approaches. Therefore, this study used REML instead of Bayes for both computational efficiency and consistency across methods of modeling intersectional analyses.

Each model was only set to run if the ICC was above .100. Therefore, a null model was first built for every iteration to determine the ICC. The dataset was discarded and regenerated if the null model had an ICC of less than .100. Following this, a "full" model was built with all demographic variables entered into each model's final form. Because coding does not change the overall model parameters, dummy coding was used for all models. When fitting the models for each scenario, some intersectional groups were dropped from the within one or more iterations due to non-representation as is the case in some of the more extreme proportion scenarios. In those cases, that specific scenario was recorded and removed from analysis.

Estimated Intersectional Group Coefficients. From each model, the estimates of each intersectional group’s coefficient and their associated p-values were stored. For the interaction model, the overall estimate of an intersectional group is the composite of the two additive terms and their interaction. Therefore, the true coefficient value for A1B1 was compared to the estimate of coefficients A1 + B1 + A1*B1 in the interaction model. The categorical model only contained intersectional groups, entered in as categorical variables. Thus, every coefficient estimate from each intersectional categorical variable was stored and evaluated. Finally, MAIDHA enters the intersectional groups as a level-two clustering variable; these level-two

intercept estimates were stored and compared to true level-two intercept values. In addition to intersectional groups and p-values, model fit information from AIC and BIC were stored for each iteration.

Simulation Outcomes. Bias, accuracy, coverage, power, and type 1 error were the primary outcomes used in this study to evaluate and compare the performance of each model. There were two categories of outcomes: a) those that compare an estimated coefficient value to the true coefficient value (bias and accuracy) and; b) those that examine alignment of statistical significance with the presence or absence of a true coefficient difference (power, type 1 error, and coverage). Below, I describe the process of comparing stored estimates from each model to the true coefficient value and true difference. In addition, I describe the criteria for each outcome that was applied to flag intersectional group coefficient estimates as potentially problematic for misrepresenting the true coefficient value or true difference.

Comparison to True Coefficient of Intercept Values. Bias and accuracy were used to examine the deviation of each model's estimated intersectional group coefficients from the true coefficient value. These true coefficient values were generated under coefficient truth generation and were presented in Tables 6 and 7.

Bias. Bias examines the average difference between the estimated intersectional group coefficient and the true coefficient value across all repetitions: $\boldsymbol{\delta = \bar{\hat{\beta}} - \beta}$ ***where*** $\boldsymbol{\beta}$ was the true coefficient value for each intersectional group and $\bar{\hat{\boldsymbol{\beta}}}$ was the estimated intersectional group coefficient, averaged across 1000 repetitions ($\bar{\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{B} \hat{\boldsymbol{\beta}}_i / \mathbf{1000}$). Thus, bias was calculated for each intersectional group's coefficient, averaged across 1000 replications.

To flag for extreme bias, the standard error around the estimate, $SE(\hat{B})$, was calculated; Bias that that exceeded $2SE(\hat{B})$ was flagged as high. The standard error around the estimate was

calculated from the standard deviation of the intersectional group coefficient estimate averaged across all replications: $SE(\hat{B}) = \sqrt{[\frac{1}{B-1}]\sum_{i=1}^{B}(\hat{\beta}_i - \bar{\hat{\beta}}_i)^2}$.

Accuracy. Accuracy was measured through mean square error (MSE), a metric that incorporates measures of bias and variability. The formula used to calculate MSE was: $(\bar{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta})^2 + (\boldsymbol{SE}(\hat{\boldsymbol{\beta}}))^2$. Similar to bias, MSE was computed for the average of each intersectional group's estimate across all 1,000 repetitions. A highly accurate model would have an MSE close to 0; MSE that exceeded 0.5 was flagged.

Alignment of Statistical Significance with the Presence or Absence of a True Coefficient Difference. Three outcomes-- power, type 1 error, and coverage-- were used to examine the extent to which the p-value estimate for each intersectional group coefficient, and thus the hypothetical decision to reject or retain the null hypothesis, was representative of the true difference for that intersectional group. A model that accurately reflects the relationships simulated into the data would estimate the intersectional groups with true differences to have low p-values and intersectional groups with no true differences to have p-values above the alpha level. For conditions with two demographic categories, there were 5 intersectional groups designed to have a true difference. For conditions with three demographic categories, there were 12 intersectional groups designed to have a true difference.

Using a .05 alpha level, statistical significance was used to evaluate the extent to which estimated intersectional group coefficients reflected the existence of a true difference (or no difference) modeled into the datasets. In the interaction model, the statistical significance of each interaction term was evaluated. In the categorical model, the statistical significance of each categorical term was evaluated. The number of times each intersectional group's p-value was equal to or below .05 was examined to determine power and type 1 error rates. For each

intersectional group's coefficient estimate, across the 1000 repetitions, the number of times the p-value was equal to or less than the alpha of .05 was summed.

Finally, in the MAIDHA model, the difference of each level two intersectional social strata's intercept from 0 was evaluated. In MAIDHA, the intersectional groups of interest were modeled as "random effects," which are not subject to traditional hypothesis testing as fixed effects. Therefore, it is not possible to analyze p-values to determine statistical significance. Instead, I calculated a 95% confidence interval around the random variance of each level two intersectional group intercept. If the interval contained zero, I concluded that the intercept was not significantly different from zero. If the interval did not contain zero, I concluded that the intercept was significantly different from zero. The number of times this interval contained 0 was summarized across 1000 replications.

Coverage. Coverage was determined by calculating the proportion of times the confidence interval surrounding an estimated intersectional group's coefficient contained the true value. Thus, a 95% confidence interval was formed around each intersectional coefficient estimate to determine coverage. Each confidence interval was calculated as $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} \pm \boldsymbol{Z}_{\mathbf{1}-\boldsymbol{\alpha}/\mathbf{2}}\boldsymbol{SE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})$. This interval was then examined to determine the percentage of times, across 1000 repetitions, that the confidence interval contained the true coefficient value for each intersection. A model that correctly represents the true difference would have a coverage estimate close to 95%. Flagging for coverage followed the lower bound of Bradley's (1978) "liberal" criterion for robustness, where the resulting percentage of coverage was flagged if it is less than 92.5%.

Power. For intersections designed to have a true difference, the power of detecting that difference was determined by examining the overall proportion of times the p-value was correctly rejected. For example, if 700 of the 1000 repetitions estimated a p-value equal to or less

than .05 for a given intersectional group's coefficient, then the overall power for detecting that intersectional group's difference was .70. The study aims for power to be at .80 or higher. Thus, model estimates with a power lower than .80 were flagged.

Type 1 Error. For intersections designed to have no true difference, the type 1 error rate was calculated by examining the overall proportion of times the p-value was rejected when there was no true difference. Using a .05 alpha level, it was expected that less than 50 of the 1000 replications contained p-values less than .05. Type 1 error was flagged when this calculated rate was larger than .05.

Model Fit. In addition to each of these data simulation outcomes, I also investigate reported model fit through the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) where AIC = $D(\beta, \sigma 2) + 2k$ and BIC = $D(\beta, \sigma 2) + k \log n$, and k is the number of free parameters in the model (Akaike, 1974; Schwarz, 1978). Unlike primary simulation study outcomes of interest, AIC and BIC do not have flagged cut-off values where "good fit" is defined. Instead, these statistics are used to compare models, where the models with lower AIC and BIC values are preferred. Under "Reporting of Results and Proposed Comparative Analysis," I discuss how I will use AIC and BIC values to evaluate and compare model fit and parsimony across models and conditions.

Chapter 4: Results

Terminology

Before discussing the results, I present the terminology used throughout this chapter. The 27 scenarios have a nomenclature I use as shorthand to represent the combination of conditions used to compare each of the three methods of conducting a statistical analysis through the lens of intersectionality. The nomenclature contains three elements describing each scenario's components: a) sample size; b) proportion representation category; and c) the standard deviation category. The first element represents sample size and is expressed as n_5000, n_10000, or n_20000. The second element represents the proportion of representation and is expressed as p1 for even representation, p2 for uneven representation, and p3 for extremely uneven representation. Finally, the standard deviation is expressed as std_small for small or unchanged standard deviation, std_large for large standard deviation near 5.0, and std_mixed for scenarios where 10% of the intersectional groups in schools were assigned to a large standard deviation condition, and the remainder were unchanged. For example, the nomenclature of n_5000_p1_std_small refers to a scenario with an n of 5000, an even proportional representation, and a small standard deviation. One additional component of the scenarios is the number of identity categories analyzed; a separate data generation process was required to produce data sets containing intersectional groupings based on two or three “identity” categories. These two categories of data were analyzed separately and, thus, this component is not included in the nomenclature.

In addition to scenario nomenclature, there are differences in how I refer to the results of each intersectional group. For interaction and categorical models, the results of each intersectional group are referred to as estimated main effect coefficients, or “coefficients” for

short. For the MAIDHA model, the results are the estimated level two intersectional group intercepts, or "intercept" for short. Finally, I refer to intersectional coefficients/intercepts designed to have an effect as "true effect" and those designed to have no effect as "no effect." When discussing specific coefficients or intercepts, I use an asterisk to designate if it has a true effect; for example, A1B1* has a true effect, but A5B1 does not.

Procedures for Obtaining Results

This simulation study was run in two waves of data generation. First, datasets were generated to represent two-category conditions. Then, datasets were generated to represent three-category conditions. For the two-category condition, 27 different scenario combinations were explored: varied proportion of representation (even, unbalanced, and extremely unbalanced), within-group standard deviation (small, mixed, and large), and sample size (5000, 10,000, and 20,000). Within each of these scenarios, 1000 replications of datasets were generated. Each model was run on each of the generated datasets, yielding 27,000 datasets and, thus 81,000 combinations of scenarios and models. For the three-category condition, 21 different scenario combinations were explored, which varied in proportion of representation (even, unbalanced, and extremely unbalanced), within-group standard deviation (small, mixed, and large), and sample size (5,000, 10,000, and 20,000). Six scenario combinations fit in the two-category models could not be fit in some the three-category model iterations due to a lack of representation in certain intersectional groups. For scenarios with n = 5000 and n = 10,000 in the extreme proportion setting, there were some iterations in which not all groups were represented. Thus, those results were not included in this study. The dropped scenarios were:

- n_5000_p3_std_small
- n_5000_p3_std_large

- n_5000_p3_std_mixed
- n_10000_p3_std_small
- n_10000_p3_std_mixed
- n_10000_p3_std_large.

As described in Chapter 3, 1000 replicates were run for each scenario and the outcome was summarized over these replications. The outcomes examined included accuracy, bias, coverage, type 1 error, and power. Accuracy was calculated as the average deviation, across 1000 replications, of the observed value from the true value. An accurate value is one where the deviation from its true value is 0; values above .50 points away from the true value were flagged.

For bias, the accuracy value was multiplied by the standard error of beta (SE($\hat{B}$)) to incorporate error in the deviation estimate. Bias was flagged as moderate for 0.5* SE($\hat{B}$) or extreme as 2* SE($\hat{B}$).

Coverage was determined by examining the percent of replications for which the true value was included in the 95% confidence interval for each coefficient or intercept. Coverage was flagged if the percentage of coverage was below 92.5%.

Type 1 error was examined for no effect coefficients or intercepts by calculating the percent of replications for which the p-value was below the alpha level of $\alpha = .05$, thus incorrectly rejecting the null. Type 1 error was flagged if the percentage of time the null was incorrectly rejected was greater than the 5% alpha level.

Finally, power was examined for true effect coefficients as the percent of times the null hypothesis was correctly rejected; the percent of replications for which the p-value was less than the alpha level of .05. Power was flagged when the false negative percentage was less than 80%.

Results were generated for each outcome in tables that summarize the outcome for each coefficient or intercept and scenario, with shading to represent flagged instances. The number of tables is extensive and stored in an online repository. To discuss the results, I start with an analysis of each model, where I present summarized results of the online repository tables in the narrative text in this chapter. This analysis is written individually for each outcome and model. For each table in the online repository, I discuss the overall patterns across scenarios and intercepts or coefficients within a given model. In the appendix I compiled summary tables. Then, to compare across models the average outcome values for true effect and no effect coefficients, which I discuss in this chapter.

Finally, I summarize results across outcomes by presenting the overall percentage of flags for each outcome and model. Since the number of flags possible for each model and outcome varied, the number of observed flags was transformed into percentages representing the number of observed flags divided by the number of possible flags. For example, if there were 12 intersectional groups in the model, the number of flags present for bias, accuracy, and coverage was divided by 12. If five of those outcomes were true effects, then the number of flags for power was divided by 5, and the number of flags on a scenario for type 1 error was divided by 7. Finally, I summarize the average of each outcome across all coefficients or intercepts and AIC and BIC values for model fit.

Two-category Results

Accuracy

Model Analysis. Tables 2.1a-c in the repository present the average accuracy for each intersectional main effect coefficient or level two intercept.

Interaction. The two true effect coefficients were flagged for high accuracy in scenarios with large standard deviation. The average accuracy values for the true effect coefficients in the high standard deviation scenarios were A1B1* (accuracy = 22.70) and A2B1* (accuracy =14.94). No other coefficients or scenarios were flagged for this model.

Categorical. The categorical model followed a similar pattern, where the true effect coefficients in large standard deviation scenarios were flagged as inaccurate. The scenario n_5000_p3_std_large presented an exception to this pattern where three no effect coefficients were flagged as inaccurate: A4B0 (accuracy = 0.57), A5B0 (accuracy = 0.57), and A6B0 (accuracy = 0.61). While these values deviated from their true values, that deviation was minimal compared to the average true effect coefficient (accuracy = 13.52) flagged in large standard deviation scenarios.

MAIDHA. The MAIDHA model also demonstrated greater inaccuracies for true effect coefficients, but the pattern deviated from what was observed in the previous two models. Accuracy was always flagged for two true effect intercepts, A1B0* and A2B0*. While the accuracy values were always above .50 for those coefficients, the average value in large standard deviation scenarios (accuracy = 22.94) was greater than the average value in small and mixed standard deviation scenarios (accuracy = 1.26). The two other true effect intercepts, A1B1*, A2B1* were flagged for accuracy when the standard deviation was large. Finally, there was one no effect intercept, A3B1, that was flagged when the standard deviation was large.

Model Comparison. Table 8 presents the average accuracy value for each true and no effect coefficient or intercept in each scenario, by model. The table presents shading when the average accuracy value was larger than the flagging cut-off of .50. The shading shows a clear pattern of what was described in the individual model analysis, where scenarios with large

deviations were less accurate for true effects. For MAIDHA, the accuracy values on average were above the flagging criteria in the mixed scenarios as well. All three models had better accuracy for no effect coefficients or intercepts. This means that each model represents a value of "0" with greater accuracy than a value different from 0. Figure 9 demonstrates the distribution of accuracy values with a dashed line used to indicate the .50 flagging threshold.

Bias

Model Analysis. Tables 2.2a-c contain bias estimates for individual coefficients or intercepts. The first set of tables, 2.2a-c.1, presents the mean bias values. The second set of tables, 2.2a-c.2, reports the average standard error of beta values, which were used to both generate bias values and flag for both moderate and extreme bias.

Interaction. The two true effect coefficients, A1B1* and A2B1*, were always flagged in mixed and large standard deviation scenarios. In addition, the two true effect coefficients were the only ones to receive any flagging for extreme bias: A1B1* had extreme bias on all mixed and large standard deviation scenarios, and A2B1* had extreme bias on all large standard deviation scenarios. The two true effect coefficients contained the highest bias values of all coefficients, where the highest accuracy values were in large standard deviation conditions, with absolute values ranging from 3.66 to 4.80. All no effect coefficients were also flagged for moderate bias on the mixed standard deviation scenarios. Some no effect coefficients also received flagging on some of the large standard deviation scenarios. There were no instances of high bias in any of the small standard deviation scenarios.

Categorical. The categorical model also did not present any flagging for bias when the standard deviation was small. The true effect coefficients were always flagged for bias in the mixed and large standard deviation scenarios, except A1B0*, which was not flagged for bias on

n_10000_p2_mixed, n_10000_p3_mixed, and n_20000_p3_mixed. Most of the bias flagged for true effect coefficients was extreme. Bias was also regularly flagged for no effect coefficients on mixed and large standard deviation scenarios—but not as frequently as the flagging for true effect coefficients. In addition, most of the bias flagged for the no effect coefficients was moderate.

MAIDHA. All intercepts in the MAIDHA model were flagged for moderate or extreme bias in six or more scenarios. The effect of the intercept did not appear to determine whether bias was flagged; three of the five true effect intercepts (A1B0*, A2B0*, and A2B1*) were flagged across all scenarios for bias. In addition, five of the eight no effect intercepts (A4B0, A6B0, A0B1, A4B1, and A6B1) were flagged across all scenarios for bias. There was slightly more bias flagged in large standard deviation scenarios compared to small and mixed.

Model Comparison. Table 9 presents the average absolute bias value for each true effect and no effect coefficient/ intercept, by scenario and model. The table presents shading for moderate and extreme bias based on the average SE($\widehat{\boldsymbol{B}}$) value. The shading displays the patterns seen in the individual model analysis: both categorical and interaction models did not flag any bias on small standard deviation scenarios. For both of those models, true effect coefficients showed greater amounts of bias than no effect coefficients. The average bias values on MAIDHA intercepts were flagged for extreme bias in every scenario.

The distribution of bias values is depicted in Figure 10 flagged for moderate bias, and Figure 11 flagged for extreme bias. As can be seen, most values that fall close to 0 are flagged for moderate bias. Extreme bias captures the values in the tails, or the biggest deviations.

Coverage

Model Analysis. Tables 2.3a-c contain information on coverage for each model and coefficient/ intercept. Tables 2.2a-c.1 present the percent of times, out of 1000 replications, for which the 95% confidence interval contained the true value for the given coefficient/ intercept. The set of tables, 2.3a-c.2 presents average 95% confidence intervals.

Interaction. In every scenario, the two true effect coefficients, A1B1* and A2B1*, were flagged for lack of coverage, along with one no effect coefficient, A3B1. The flagged coefficients had little to no coverage, at or near 0%. There was little to no change in coverage across scenario scenarios.

Categorical. Coverage was strong, above 92.5%, for all coefficients in the small standard deviation scenarios. For mixed and large standard deviation scenarios, every true effect coefficient was flagged for poor coverage. For true effects, coverage was often much higher in mixed standard deviation scenarios compared to large. For example, for A1B0*, coverage in mixed scenarios ranged from 47.60-91.40%, and in large standard deviation scenarios coverage ranged 0 to 1.90%. There was flagging for no effect coefficients across many mixed and large standard deviation scenarios as well.

MAIDHA. The MAIDHA model had perfect (100%) coverage on every single intercept, both true effect and no effect.

Model Comparison. Table 10 presents the average percent of coverage for each true effect and no effect coefficient or intercept by scenario and model. Table 10 presents the patterns of high percent of flagging for coverage in interaction and categorical models. For flagged coefficients in the interaction and categorical models, true effect coefficients always had worse

coverage than the no effect coefficients in the same scenario. As discussed in the individual model analysis, there was no flagging in the MAIDHA model.

Figure 12 presents the distributions of coverage percentages, with a line at 92.50% to indicate flagging. All three distributions of flagged coverage were different, meaning coverage was highly dependent on the model. The distribution analysis depicts a disparity of coverage in the interaction model because most no effect coefficients had perfect coverage, and most true effect coefficients had no coverage. For the categorical distribution there was less of a disparity and more coverage percentages that fell at or near the 92.5% threshold. Finally, the MAIDHA distribution shows perfect coverage.

Power

Tables 2.4a-c contain information on power. Tables 2.4a-b.1 for interaction and categorical models contains the percentage of times the p-value was less than the alpha level of .05, and the Tables 2.4a-b.2 present the average p-value. For the MAIDHA model, Table 3.4c.1 is the percentage of times the calculated interval did not contain zero, and 3.4c.2 is the average interval mean.

Model Analysis.

Interaction. The two true effect coefficients modeled, A1B1* and A2B1*, were detected correctly 100% of the time in all scenarios. Thus, there were no flags for power in the interaction model.

Categorical. There were also no coefficients flagged for power in the categorical model. The true effect coefficients had high coverage and thus were correctly detected nearly all the time. The minimum percentage of true effect correctly detected being 87.80% for A1B0* for the scenario n_5000_p3_std_mixed.

MAIDHA. There were four instances of flagging in the MAIDHA model for power. All flags were on the A3B0* intercept. The flagged scenarios for A3B0* were for n_5000_p3_std_small = 44.60%, n_5000_p3_mixed = 20.50%, n_10000_p3_std_mixed = 55.60%, and n_5000_p3_std_large = 46.50%. All the scenarios with low power had extreme proportional representation, and three of the four were for the smallest sample size, n= 5000.

Model Comparison. Table 11 presents the average rate of correctly detecting true effects. When averaged across all true coefficients, all percentages for detecting true effects were above the threshold of flagging of 80%, with the lowest percentage of 84.10% for the MAIDHA model on n_5000_p3_std_mixed. Overall, the interaction model had the highest percentage of true effects correctly detected, where both true effects were always detected. Power across all three models was extremely strong. The distribution of flags in Figure 13 further demonstrates the strength of power.

Type 1 Error

Model Analysis.

Interaction. In the interaction model, the null hypothesis for no effect coefficients was frequently rejected above the 5% intended error rate. A3B1 was detected as having an effect 90-100% of the time, whereas the other coefficients generally had an error rate below 10%.

Categorical. The categorical model also had a high false positive rate, and the type 1 error flagging was related to standard deviation scenarios. The percentage of incorrectly rejected p-values was lowest for small standard deviation scenarios, for which only about half of the coefficients and scenarios were flagged. Meanwhile, type 1 error was the highest for mixed standard deviations, where all the coefficients were flagged across each of the mixed standard deviation scenarios. The lowest false positive rate for mixed standard deviation scenarios was

A3B1 in n_5000_p3_std_mixed at 5.20% and the highest was for A4B0 in n_20000_p2_std_mixed at 58.30%. In the large standard deviation scenarios, false positive rates ranged from 4.40% for A4B0 in n_5000_p1_std_large to 32.70% for A6B0 in n_20000_p1_std_large. In addition to standard deviation, it appeared that sample size had a relationship to the false positive rate where larger sample sizes had higher false positive rates than smaller sample sizes.

MAIDHA. The MAIDHA model had a low amount of flagging for type 1 error and thus a minimal false positive rate. The main exception to the low rate was A3B1, which was detected as incorrectly having an effect 80-100% of the time. The other intercepts had minimal flagging. Specifically, across the 27 scenarios, A0B0 had four false positives, A0B1 had eight false positives, A4B1 had three false positives, and A5B1 had three false positives.

Model Comparison. When averaged across all coefficients/intercepts, the percentage of false positives for each model was frequently above the threshold of 5%, as seen in Table 12. The distribution analysis, Figure 14, shows how many of the detection rates were at or above the 5% threshold. While MAIDHA did have minimal flagging for type 1 error, the near 100% false positive rate of A3B1 skewed the overall average. Therefore, on average, all models suffered from high type 1 error rates. MAIDHA's type 1 error rates were similar across scenarios. For interaction and categorical type 1, error rates were lowest for small standard deviation scenarios and higher for mixed scenarios.

Two-category Summary/ Conclusion

To summarize the two-category scenarios, I first present summary Tables 13-15 which detail the average percentage of flags, per outcome, by scenario. Next, I explore a cross-outcome analysis based on Table 16 which reports the percentage of flags averaged across all outcomes,

by model and scenario. I explore the nuance of standard deviations using Figure 15. Then, table 17 and Figure 16 present the overall percentage of flagged coefficients/ intercepts by model and outcome. To explore whether there is a relationship between model and outcome, I conducted a chi square test of independence on the percent of flags. Finally, I supplement the scenario and outcomes analysis with information on model fit in Table 18. For Tables 13-17, a five-category shading system is utilized to provide a description of the overall percentage of flagging. "Excellent" is used to describe flagged percentages from 0-19% where there is no shading. A light orange shades flags 20-39% where they are considered "
"moderate" flagging. A medium orange is used to shade flag percentages 40-59% and is considered a "fair" amount of flagging. "Poor" flagging is indicated by a dark orange, which shades flag percentages from 60-79%. Finally, anything flagged about 80% is shaded bright red, and considered "extremely poor".

Interaction. In the interaction model, Table 13, power was never flagged. Accuracy was flagged one-third of the time in each of the large standard deviation scenarios but not at all in the mixed or small standard deviation scenarios. Bias was extremely poor in mixed standard deviations and ranged from fair to poor in large standard deviation scenarios. Coverage was flagged half the time, across all scenarios. Finally, type 1 error performance was moderate to extremely poor, with the greatest percentage of flags occurring in mixed standard deviation scenarios.

Categorical. In the categorical model, Table 14, there were no flagged instances for bias, accuracy, coverage, or power in small standard deviation scenarios. Type 1 error was flagged between 25-75% of the time in small standard deviation scenarios but in mixed and large standard deviation scenarios there was always poor or extremely poor performance. Power had

excellent performance across all scenarios. Accuracy had excellent performance in small and mixed standard deviation scenarios. Accuracy had a moderate performance in most large standard deviation scenarios. Performance was poor, however, when the sample size was small (n = 5000), and the proportional representation was extremely uneven (p3). While bias had excellent performance for small standard deviation scenarios, it was quite problematic across mixed and large standard deviation scenarios, ranging from being flagged 46.15-100% of the time. Similarly, coverage was problematic across mixed and large standard deviation scenarios, while the percentage of flags ranged from 52.85 to 100% of the time.

MAIDHA. In the MAIDHA model, table 15, coverage was never flagged and thus was considered excellent across all scenarios. Power was also minimally flagged, with just four scenarios resulting in moderate flagging. Type 1 error performance was excellent in small standard deviation scenarios. Still, flagging increased to either moderate or fair for mixed standard deviation scenarios with uneven and extremely uneven proportion representation, as well as large standard deviations with extremely uneven proportion representation. Accuracy had excellent performance in small and mixed scenarios but increased to fair performance in large standard deviation scenarios. Finally, bias for MAIDHA had extremely poor performance, where every scenario was heavily flagged, ranging from a low of 71.43% to a high of 92.86%.

Cross-Outcome Analysis. Table 16 presents the overall percentage of flags, averaged across outcomes. No model outperformed all others on every single outcome, indicating that each model has comparative strengths and weaknesses. For example, the interaction and the categorical models had a high average percentage of flags on the n_10000_p3_std_large scenario at 50.00% and 61.54%, respectively, the MAIDHA model only flagged at 38.52%. Whereas, on n_5000_p3_std_small, MAIDHA's average of 23.37% of flagged instances were worse than that

of interaction at 20.00% and MAIDHA at 5.00%. Across all models, small standard deviation scenarios were related to less flagging. To better understand the patterns of sample size and proportional representation, Figure 15 presents the overall percentage of flags for small standard deviation scenarios. In this visualization, the interaction model had the highest flags in all scenarios = except for n_5000_p3 where the MAIDHA model had a higher percentage. For the interaction model, the highest percentage of average flags was observed for the n_20000_p1 scenario. The categorical model consistently performed best, with the lowest percentage of flags occurring for n_5000_p2, n_5000_p3, n_10000_p3, and n_20000_p3. There were no clear patterns between the scenario and the percentage of outcomes. However, the categorical model surprisingly performed worse on even proportional representation scenarios than extreme proportional representation.

Table 17 presents, accompanied by Figure 16, the overall percentage of flags by each outcome and model. This table again demonstrates the variation in model performance across outcomes. The MAIDHA model had the smallest percentage of flags on coverage, power, and error. The interaction model had the smallest percentage of flags on accuracy, and the interaction and categorical models had the smallest percentage on power. Overall, the most concerning outcome was the type 1 error rate, which was poor to extremely poor across categorical and interaction models and moderate for the MAIDHA model.

To determine if there were statistically significant differences in the distributions of bias, accuracy, coverage, power, and type 1 error across interaction, categorical, and MAIDHA models, a Pearson's Chi-squared test was conducted. The test yielded a Chi-squared statistic of 223.63 with 8 degrees of freedom. The p-value of < .001 was below the alpha level of 0.05. This

indicates that we can reject the null hypothesis and conclude that there were statistically significant differences in the distributions of the five outcomes across the three models.

AIC and BIC, as seen in Table 18, were consistent across models in each scenario. Therefore, no evidence exists that model fit changes due to the selected model. However, the overall fit did change based on the scenario. The scenarios that had the best overall fit (in order from best) were N_5000_p1_std_small, N_5000_p2_std_small, and _5000_p3_std_small. The scenarios that had the worst overall fit, in order from worst, were N_20000_p3_std_large, n_20000_p2_std_large, and n_20000_p1_std_large.

Three-category Scenarios

Accuracy

Model Analysis. Tables 3.1a-c in the repository present three-category accuracy results for each model.

Interaction. Accuracy values were below the flagging threshold of .50 in all small standard deviation scenarios. Four coefficient estimates in the mixed standard deviation scenarios received three or more flags: A2B1C1*, A1B1C2*, A2B1C2, and A6B1C2. Accuracy values were the highest overall in large standard deviation scenarios; every true effect coefficient was flagged in all seven scenarios. In addition, each of the no effect coefficients had at least three flags on the large standard deviation scenarios.

Categorical. There were no coefficients flagged for accuracy on any of the small standard deviation scenarios. Five true effect coefficients (A1B0C0*, A2B0C0*, A2B1C0*, A2B1C1*, A1B1C2*) and three no effect coefficients (A6B0C0, A2B0C2, and A2B1C2) received at least one flag on the mixed standard deviation scenarios. The large standard deviation scenarios had

the highest number of flags; all coefficients except for seven no effect coefficients had at least one flag.

MAIDHA. There were eight intercepts flagged on every single scenario: A2B0C0*, A3B0C0*, A2B0C1*, A1B0C2, A3B0C2, A1B1C2*, A2B1C2, and A6B1C2*. Four scenarios were flagged on all mixed and large standard deviation scenarios: A2B1C0*, A3B1C1, A0B1C2, and A4B0C2. Finally, all remaining intercepts received flagging on at least one large standard deviation scenario except for A2B0C0*, A4B0C1, A0B1C1, A2B1C1*, and A5B1C1.

Model Comparison. Table 19 presents the overall pattern for the individual analysis. Across analyses, small standard deviation scenarios were the most accurate, followed by mixed scenarios, and the large scenarios were least accurate. While both true effect and no effect coefficients or intercepts were flagged, the average deviation from the true score for true effect coefficients/ intercepts was always higher compared to no effect coefficients/ intercepts. In the distribution of accuracy values displayed in Figure 17, the majority of coefficient/intercept and scenario combinations had minimal deviation and were thus accurate. The values in the tails of each distribution show that while on a whole the models were relatively accurate, each had model coefficient/ intercepts that were estimated extremely inaccurately.

Bias

Model Analysis. Tables 3.2a-c contain bias estimates for individual coefficients. The first set of Tables, 3.2a-c.1, report mean bias values. The second set of Tables, 3.2a-c.2 report the average standard error of beta values, which were used to generate bias values, which in turn were used to flag both moderate and extreme bias.

Interaction. Every coefficient produced by the interaction model presented bias in more than half of the scenarios. In the small standard deviation scenarios, all bias was considered

moderate. In mixed and large standard deviation scenarios, extreme bias was observed. The bias appeared to be relatively balanced between true effect and no effect coefficients.

Categorical. There was no bias flagged in small standard deviation scenarios for any coefficient. Overall, there was more extreme amounts of bias present in large standard deviation scenarios. However, the following eight no effect coefficients were flagged for bias in the mixed scenario but not for the large scenarios: A4B1C0, A0B0C1, A4B0C1, A0B0C2, A4B0C2, A0B1C1, A4B1C1, and A4B1C2.

MAIDHA. In the MAIDHA model, 24 of the 42 intercepts MAIDHA= were flagged as extremely biased for most or all scenarios. Another 12 intercepts were flagged on most or all scenarios for moderate bias. There was no noticeable relationship between scenario and flagging.

Model Comparison. As seen in Table 20 there was a clear pattern between flagging for bias and the standard deviation modeled into the data for both the categorical and interaction models. The small standard deviation scenarios were most accurate, followed by mixed standard and large standard deviation scenarios. The pattern for the MAIDHA model was less clear, but higher bias values were observed for the large standard deviation scenarios. Overall, the categorical model yielded the lowest average bias, and the MAIDHA model yielded the highest average bias.

The distributions of flags are shown in Figures 18 and 19. Overall, the categorical model produced a clear divide between flagging for moderate bias compared to the absence of flagging for bias, while the interaction and MAIDHA models had many instances on the threshold between no bias and flagging for moderate bias This can be seen in the Figures 18 and 19 where the bias values are further from the threshold line in the categorical model compared to the spread across the threshold line for the interaction and MAIDHA models. Most of the extreme

bias flags for categorical and interaction models were only observed in the tails, while the MAIDHA model had many extreme bias values close to the center.

Coverage

Model Analysis. Coverage is reported in two sets of tables in the online repository. Tables 3.3a-c.1 report the overall percentage of times, out of 1000 replications, that the true value was located within the 95% confidence interval. These tables are shaded to represent the flagging of values below 92.5%. The second set of Tables 3.3a-c.2 present the average upper and lower limits of the confidence intervals across 1000 replications.

Interaction. In the interaction model, coverage was strong for no effect coefficients, where the average coverage was just under the cut-off at 92.39%. Meanwhile, true effect coefficients demonstrated substantially worse coverage, where the average coverage was 67.32%. Every true effect coefficient had some scenarios in which it was flagged for poor coverage. However, only two no-effect coefficients scenarios were flagged: the first occurrence occurred for A1B1C1, which was flagged for under coverage for the n_20000_p1_std_large scenario (coverage = 66.30%). The second occurrence was for A2B1C2, which was flagged on most mixed and large standard deviation scenarios where the average coverage on flagged scenarios was (coverage = 21.00%). The combination of true effects and large standard deviation scenarios produced extremely poor coverage results, where the average was 22.62% coverage. Models in small standard deviation scenarios produced the highest overall coverage.

Categorical. Similarly, in the categorical model, coverage was weak for true effect coefficients (average coverage = 37.67%) and slightly stronger for no effect coefficients (average coverage = 78.23%). There were eight coefficients, a mix of true effects and no effects, that were flagged for poor coverage across all scenarios: A2B1C1*, A6B1C1*, A2B0C2, A5B0C2,

A1B1C2*, A2B1C, A3B1C2*, and A6B1C2*. All other coefficients had strong coverage in small standard deviation scenarios but worse coverage in mixed and/or large standard deviation scenarios.

MAIDHA. The pattern for the MAIDHA model did not resemble the coverage results for the two other models. Three true effect intercepts were flagged across most, if not all, scenarios: A1B0C0* (average coverage = 59.59%, A6B1C1* (average coverage = 62.27%), and A1B1C2* (average coverage = 67.97%). Three other true effect intercepts had flagging on only one scenario: A3B0C0* (n_20000_p3_std_mixed = 89.30%), A2B1C0* (n_5000_p2_std_mixed = 91.00%), and A2B1C1* (n_5000_p2_std_mixed = 91.00%). Only one no effect intercept received any flagging, A2B0C0, which received flags on all n_5000_std_mixed and n_10000_std_mixed scenarios across all proportion representations as well as n_20000_p3_std_mixed. The average coverage for the flagged values of the A2B0C0 intercept was 88.12%.

Model Comparison. Table 21 presents the average percentage of observations each true value fell in the 95% confidence interval, with values under 92.5% shaded to indicate under coverage. Overall, MAIDHA presented the strongest coverage across true effects and no effects. The interaction model demonstrated strong coverage for small standard deviation scenarios but was considerably weaker for large standard deviation scenarios. The categorical model, on average, displays the weakest coverage across scenarios.

Figure 20 displays the overall distribution of coverage percentages. A dashed red line represents a flagging criterion of 92.5% or below. The distribution shows that the categorical model suffers from under coverage for many coefficients and replications. Meanwhile, the MAIDHA model has minimal dispersion of under coverage.

Power

Model Analysis. Tables 3.4a-c present findings for power. Tables 3.4a-b.1 report results for the interaction and categorical models and show the percentage of times the p-value was less than the alpha level of .05, and Tables 3.4a-b.2 presents the average p-value. For the MAIDHA model, Table 3.4c.1 reports the percentage of observations for which the calculated interval did not contain zero, and 3.4c.2 presents the average interval mean.

Interaction. Only one coefficient, A3B1C2*, was flagged for minimal power. For A3B1C2*, in most scenarios, the null hypothesis was retained more than 80% of the time. Specifically, A3B1C2* had low power in all but four scenarios; those four had uneven proportion representation as a common denominator. All other coefficients had nearly perfect detection across scenarios.

Categorical. Most coefficients had near-perfect power across scenarios. However, A3B1C2* and A6B1C2* always had lower detection percentages and were always flagged. A5B0C2* had low power on small and mixed scenarios, and A1B1C2* had low power on small and some mixed scenarios. Finally, A2B0C0* and A2B0C1* were flagged for several large standard deviation scenarios.

MAIDHA. Four intercepts were flagged for over half of the scenarios: A2B0C0*, A2B0C1*, A5B0C2*, and A3B1C2*. There was no distinct pattern for those four intercepts as to which scenarios had a higher error. Two other intercepts were flagged on n_p2_std_mixed: A1B0C0* and A2B1C0*.

Model Comparison. Table 22 presents the percentage of flags across scenarios. On average, the interaction model had the strongest power, and the average of all coefficients was above the flagging criteria of 80%. Both the categorical and MAIDHA models' average power

was below the 80% threshold for most scenarios. These patterns are evident in the distribution of percentages in Figure 21. Figure 21 shows that the percentages in all three models were close to 100. MAIDHA had many cases with low percentages in the tail, indicating a high number of intercepts that were rarely detected.

Type 1 Error

Model Analysis.

Interaction. Most coefficients in the interaction model were flagged for type 1 error most of the time. Some coefficients, such as A4B1C1, had detection rates slightly over the alpha level (average = 10.50%). Whereas others, such as A2B1C2, had p-values that were nearly always above the alpha level (average = 99.84%).

Categorical. Every coefficient in the categorical model received flagging in at least some scenarios. Most coefficients received flagging on over half of the scenarios. Oftentimes, scenarios with large standard deviations had higher type 1 error rates compared to those in mixed or small standard deviations on the same coefficient —but this pattern did not hold across all coefficients.

MAIDHA. The MAIDHA model had a high type 1 error rate. All but one of the intercepts, A5B0C1, were flagged in almost all scenarios. The percentage of false detection was often well above the 5% alpha threshold, frequently between 90-100%.

Model Comparison. Table 23 presents the average false positive percentage across all coefficients/intercepts for each model. Every model had an average type 1 error rate above the 5% alpha threshold for all scenarios. The highest average percentage of false positives was 89.41% for the interaction model in n_20000_p1_std_large. The lowest rate was for the categorical model in n_5000_p1_mixed at 17.44%. Overall, the categorical model had the lowest

type 1 error rates. The distribution in Figure 22 shows the wide variety of type 1 error rate percentages across intercept/coefficients and scenarios. In each model, there was a minority of instances that were below the 5% alpha threshold.

Model Summary/Conclusion

To summarize the two-category scenarios, I first present summary Tables 24-26 which detail the average percentage of flags, per outcome, by scenario, with each table representing a different model. Then, I explore a cross-outcome analysis based on Table 27 which reports the percentage of flags averaged across all outcomes, by model and scenario. I explore the nuance of standard deviation scenarios using Figure 23. Then, Table 28 presents the overall percentage of flagged coefficients/ intercepts by model and outcome, with an accompanying visualization shown in Figure 24. To explore if there was a relationship between model and outcome, I conducted a chi square test of independence on the percent of flags. Finally, I supplement the scenario and outcomes analysis with information on model fit in Table 29. For Tables 24-28, a five-category shading system is utilized to provide a description of the overall percentage of flagging. “Excellent” is used to describe flagged percentages from 0-19% where there is no shading. A light orange shades flags 20-39% where they are considered moderate flagging. A medium orange is used to shade flag percentages 40-59% and is considered a fair amount of flagging. Poor flagging is indicated by a dark orange, which shades flag percentages from 60-79%. Finally, anything flagged about 80% is shaded bright red, and considered extremely poor.

Interaction

Table 24 presents the percentages of flags for each outcome and scenario in the interaction model. The interaction model had extremely poor performance on type 1 error, where most coefficients received flags nearly all the time across scenarios. Conversely, the interaction

model performed moderately well to excellent for power, where the percentage of coefficients flagged in each scenario was always 20% or less. For accuracy and coverage, there was a relationship between the standard deviation scenario and the percentage of flags. For accuracy and coverage, large standard deviation scenarios were related to a higher level of flagging. Bias had a relationship with proportional representation; there were always more flags in even representation (p1) compared to uneven representation (p2) and frequently more than in extremely uneven representation (p3).

Categorical

Table 25 presents the percentages of flags for each outcome and scenario in the categorical model. Type 1 error performed worst across all scenarios, where more than 20% of the coefficients were flagged for most scenarios. The categorical model performed best on power, with only one scenario flagged more than 20% of the time. All outcomes in the interaction model appeared to have a relationship with large and mixed standard deviations related to higher flagging compared to small standard deviations. In addition, a higher sample size was often related to increased type 1 error. Only one outcome/ scenario, n_20000_p3_std_mixed, had extremely poor performance for type 1 error.

MAIDHA

Table 26 presents the percentage of flags for each scenario and outcome in the MAIDHA model. The MAIDHA model had extremely poor performance in all scenarios for both bias and type 1 error. The MAIDHA model performed best on coverage, where it had excellent performance across all scenarios. In small and mixed scenarios, there was increased flagging for smaller sample sizes and n = 20,000 when the proportional representation was extreme. Finally, the percentage of flags for accuracy increased for large standard deviations scenarios.

Cross Model Analysis

Table 27 presents the average percentage of flags across all outcomes for each model and scenario. The categorical model presented the only instances with outcomes flagged, on average, less than 20% of the time. The low percentage of flagging for the categorical model occurred most often for the small standard deviation scenarios. For all three models, flagging was the highest for large standard deviation scenarios. For the interaction model, flagging was higher for even representation than uneven (although it was also high for extremely uneven).

Since all models performed stronger for small standard deviation scenarios, it was useful to isolate the small standard deviations to better understand the impact of proportional representation and sample size; this is displayed in Figure 23. The interaction model had the fewest flags on n_20000_p2_std_small and the highest number on n_10000_p1_small. The categorical model has the most flags on n_5000_p2_std_small and the least number on n_10000_p2_small. Finally, MAIDHA had the most flags on n_20000_p3_small and the least flags on n_20000_p1_small. Overall, there was not a noticeable relationship between the averaged outcomes sample size and proportional representation.

Table 28 and Figure 16 present the total percentage of flags each outcome received across models. Of all the models, the categorical model performed best for four of the five outcomes: accuracy, bias, power, and type 1 error. The MAIDHA model performed best on coverage. Despite the better performance, the overall percentages for the categorical model were still concerningly high, ranging from 3.97 to 43.90%. Across all models, the worst outcome was type 1 error, which was flagged in the interaction model 93.20% of the time, in the categorical model 47.25% of the time, and in the MAIDHA model 91.11% of the time.

To determine if there were statistically significant differences in the distributions of bias, accuracy, coverage, power, and type 1 error across interaction, categorical, and MAIDHA models, a Pearson's Chi-squared test was conducted. The test yielded a Chi-squared statistic of 219.01 with 8 degrees of freedom. The p-value of < .001 was below the alpha level of 0.05. This indicates that we can reject the null hypothesis and conclude that there were statistically significant differences in the distributions of the five outcomes across the three models.

Table 29 presents the AIC and BIC, an indication of model fit. The categorical and MAIDHA models often had very similar model fit indices, but the interaction model differed substantially. This makes sense since the AIC and BIC penalize models with more parameters, and the interaction model had about half as many parameters as the other two models. Therefore, there were no consistent three scenarios across models that were best and worst. For the interaction model, the three scenarios with the best fit were n_5000_p2_std_small, n_5000_p1_std_small, and n_5000_p2_std_mixed. The three scenarios with the worst fit were n_20000_p3_std_large, n_20000_p2_std_large, and n_20000_p1_std_large. For the categorical and MAIDHA models the three scenarios with the best fit were n_20000_p1_std_mixed, n 20000_p1_std_large, and n_20000_p3_std_mixed. The three scenarios with the worst fit were n_10000_p2_std_small, n_5000_p1_std_large, and n_5000_p1_std_small.

Comparison of Two and Three-Category Results

Tables 30 and 31 present previously explored results reconfigured to explore two and three-category percentages of flags side by side. Table 30 presents the average percentage of flagged outcomes per scenario by model. Table 31 presents the percentage of flagged coefficients/intercepts by outcome.

Similarities

The two and three-category scenarios had consistent themes across other scenarios and models. First, in the individual model analysis of coefficients/intercepts across outcomes and models, a relationship between a given outcome and larger amounts of variation was observed frequently.

In Table 30, it is evident that the categorical model had a consistent relationship with standard deviation scenarios and always had a low percentage of flagging on the small standard deviation scenarios. In the interaction model, there was a consistent relationship with the mixed standard deviation scenarios where in both the two and three-category scenarios, flagging occurred approximately 50% of the time.

Table 31 helps portray the consistency in performance across outcomes for each model. For the interaction and categorical models, power was the best-performing outcome across two and three-category scenarios, and type 1 error was the worst-performing outcome. Coverage was the best performing outcome for the MAIDHA model for both two and three-category scenarios.

Differences

While the individual model analysis often centered around true vs. no true coefficients being flagged at different rates, that pattern was more apparent for two-category models than three.

Table 30 shows that the interaction model had a higher percentage of flags for small and large standard deviation conditions in the three-category scenarios. In addition, for the three-category scenarios, the interaction model flagged even proportional representation at higher rates compared to unbalanced and extremely unbalanced representation scenarios. For the categorical models, mixed standard deviation scenarios were flagged substantially higher in two-category

conditions compared to three-category conditions. Finally, the MAIDHA model always had a higher percentage of flagging in three-category conditions across scenarios.

In Table 31, it is useful to look at differences in the overall percentage of an outcome flagged. Specifically, I point to instances in each model where the percentage of an outcome flagged differed by 15 or more percentage points. For the interaction model, bias (18.25 percentage point difference) and coverage (25.40 percentage point difference) were substantially worse in the two-category scenarios than the three-category scenarios. Still, in the interaction model, power was never flagged for two-category scenarios but was flagged 16.19% of the time for three-category scenarios. In the categorical model, bias was flagged less in three-category scenarios (31.97 percentage point difference) and accuracy was flagged more in three-=category scenarios (29.87 percentage point difference). The MAIDHA model had substantial differences between two and three-category scenarios: bias was flagged at a higher rate in two-category scenarios (34.34 percentage point difference). Whereas a higher percentage of three-category scenarios were flagged for accuracy (66.10 percentage point difference), power (25.61 percentage point difference), and type 1 error (69.21 percentage point difference). Particularly striking is the change in type 1 error; for two-category scenarios, the type 1 error rate was substantially lower than in the other models, whereas, for the three-category scenarios, it was about the same as the other models.

Chapter 5: Discussion

In this dissertation, I sought to understand the implications of an intersectional lens when applied to a quantitative research design. While the specific focus of this work was on the statistical implications resulting from the methods of modeling an intersectional analysis, Chapter 2 also addresses conceptual implications because the two cannot be separated in a research project. For Chapter 5, I aim to widen the lens and re-integrate conceptual considerations in light of the statistical results. The aim of this chapter is to engage in a discussion on how and in what ways it is appropriate to use the results of this dissertation and to provide actionable steps for researchers.

This chapter begins with a discussion of the research questions with a focus on surprising statistical findings and why they might have occurred. Then, I discuss how researchers can apply the findings of this study to their own quantitative intersectional projects. In the discussion, I explore the limitations researchers will face when applying methods of intersectional analyses. In Next, I provide suggestions for future simulation studies that explore specific issues not addressed in this dissertation study. Before concluding, I reintegrate the discussion of model use, and how these results can be used in conjunction with statistical theory.

Summary of Findings

Revisit Research Questions

To answer the research questions, I reference the flagging classifications from Chapter 4: 0-19% “excellent” performance on outcomes, 20-39% “moderate” performance on outcomes, 40-59% “fair” performance on outcomes, 60-79% “poor” performance on outcomes, and 80-100% “extremely poor” performance on outcomes. Tables 13-17, 24-28, and 30-31 are a useful reference for this discussion.

Research Question 1: What are the statistical advantages and disadvantages of each model under different demographic data characteristics?

There was no one model or scenario in which performance on the outcomes observed was uniformly strong. Instead, for each scenario, every model had both statistical advantages and disadvantages. To answer this question, I investigated the overall percentage of coefficient and outcomes flagged for each model, by scenario conditions. For each model, I start with a presentation of what worked well*:* scenarios where performance on each outcome was relatively strong. Then, I discuss what went poorly: scenarios where performance on each outcome was relatively poor. I then summarize the findings from chapter 4 based on differences between true and no effect coefficient or intercepts. These results, for each model, are presented in table 32. In the section labeled *Across Outcomes,* I explore the average percentage of flags across all outcomes and discuss overall best and worst performances.

Interaction.

Individual Outcomes. Accuracy had excellent performance for small and most mixed standard deviation scenarios. For the two-category scenarios, bias had excellent performance in small standard deviation scenarios. Power always had excellent performance in the two-category scenarios and moderate performance in the three-category scenarios.

Accuracy had poor or extremely poor performance in most three-category large standard deviation scenarios. Bias had extremely poor performance in all two-category mixed standard deviation scenarios, and poor performance in scenarios with extreme representation, two-categories, and large standard deviation. Bias had extremely poor performance for three-category small and mixed standard scenarios. Across scenarios, type 1 error was always poor or extremely poor.

Differences in accuracy and bias were observed for outcomes of no effect coefficients compared to those with true effects, in which case true effect coefficients often performed worse. Coverage was also worse for no effect coefficients for the three-category = large standard deviation scenarios. In two-category scenarios, the interaction model always had moderate performance on coverage. However, for the three-category scenarios, coverage had excellent performance for all small standard deviation scenarios and fair performance in large standard deviation scenarios.

Across Outcomes. The lowest overall percentage of flagging across outcomes occurred when the standard deviation was small and only two demographic categories were included. For these scenarios, the performance of the interaction model was moderate. There were no scenario combinations in which performance across outcomes was excellent. The percentage of flags in the two-category small standard deviation scenarios was consistent across sample size and proportion representation. While there was a greater percentage of flags compared to the two-category scenario, the three-category scenario also had moderate performance on small standard deviation scenarios. Within the three-category scenarios, the lowest percentage of flags occurred for small standard deviations with uneven proportion representation.

There were several instances of poor performance for the interaction model, but none of these scenarios resulted in extremely poor performance. The worst performance for the interaction model occurred for three-categories in large standard deviation scenarios when the proportion of representation was even or extremely uneven.

Categorical.

Individual Outcomes. Accuracy, bias, coverage, and power all had scenarios within the categorical model results in which there was excellent performance. Accuracy had excellent

performance on small and mixed standard deviation scenarios. Bias and coverage had excellent performance on small standard deviation scenarios. Power had excellent performance in almost all scenarios.

There were also instances of poor or extremely poor performance for the categorical model across individual outcomes. Accuracy had one instance of poor performance in the two-category scenarios: in large standard deviations, uneven proportion representation, and small sample (n = 5,000). Accuracy had poor performance for all uneven or extremely uneven proportion representation within the three-category large standard deviation scenarios. Bias and coverage both had either poor or extremely poor performance on most two-category mixed and large standard deviation scenarios. Bias and coverage ranged from fair to poor performance in the three-category mixed and large standard deviation scenarios. For two categories, type 1 error had poor or extremely poor performance for all two-category mixed and large standard deviation scenarios, as well as several scenarios with small standard deviations. In the three-category scenarios, there were several instances of poor performance in type 1 error throughout mixed and large standard deviation scenarios, but only one instance of extremely poor performance, with a large sample size (n = 20,000), extremely uneven proportion, and a mixed standard deviation.

There were also differences observed for true compared to no effect coefficients. Accuracy and bias were often higher in mixed and large standard deviation scenarios for true effect compared to no effect coefficients. In scenarios with two categories mixed and large standard deviations, as well as in three-category scenarios with small standard deviations, coverage was often higher for no-effect coefficients than for true effect coefficients.

Across Outcomes. The categorical model had excellent performance in all small standard deviation scenarios. In addition, it had moderate performance in three-category mixed standard deviation scenarios for small (n = 5,000) and moderate (n = 10,000) sample sizes.

There was poor performance on several two-category scenarios: moderate (n = 10,000) and large (n = 20,000) sample sizes in the mixed standard deviation scenario with even or uneven representation, large (n = 20,000) sample sizes in the large standard deviation scenario, and moderate sample size (n = 10,000) in the large standard deviation scenario when the proportion of representation was extremely uneven.

MAIDHA.

Individual Outcomes

The MAIDHA model always had excellent coverage across all scenarios. In addition, it had excellent accuracy in two-category small and mixed standard deviation scenarios and three-category small standard deviation scenarios. There was excellent performance for power in most two-category scenarios, and moderate performance in most three-category scenarios. For two-categories, there was excellent performance for type 1 error in all scenarios with small standard deviations, mixed standard deviations with even proportion representation, large standard deviations and either even or uneven proportion representation.

Accuracy had no instances of poor or extremely poor performance in two-category scenarios but did have fair performance on two-category large standard deviation scenarios. Accuracy had extremely poor performance on three-category large standard deviation scenarios. There was poor or extremely poor performance for bias on all scenarios in both two and three-category models. Finally, type 1 error was always extremely poor in the three-category scenarios.

There was also a relationship between the true effects and no effects and how each intercept effect type performed on each outcome. True effect intercepts always performed worse than no effects on accuracy and bias. True effect intercepts often had worse coverage than no effect intercepts for three-category scenarios.

Across Outcomes. The MAIDHA model had excellent performance for all two-category scenarios with small standard deviations except the small sample size (n = 5,000) with uneven representation, where it had fair performance. In addition, MAIDHA had fair performance across all other scenario combinations in the two-category scenario except the small sample size (n =5,000) with uneven proportion representation in large standard deviations. The MAIDHA model had poor performance in three-category scenarios with large standard deviations for all moderate sample sizes (n = 10,000) as well as large sample sizes (n = 20,000) with uneven or extremely uneven proportion representation.

Research Question 2: In what ways does each model perform differently from one another under each demographic data characteristic scenario?

Looking at specific outcomes across demographic data characteristics, the MAIDHA model performed better than the interaction and categorial models for type 1 error within in two-category scenarios, where it performed moderately. The categorical model outperformed the other models in the three-category scenarios, where it had a fair overall performance on type 1 error, while MAIDHA and the interaction model had extremely poor performance on type 1 error in the three-category scenarios. For power, the MAIDHA and interaction models saw an increase in the percentage of flags with the three-category scenarios. In contrast, the categorical model had excellent power performance in two and three-category scenarios. MAIDHA always had excellent coverage, while the interaction and categorical models had fair coverage. Finally,

MAIDHA had a higher percentage of flags on accuracy and bias than the interaction and categorical models.

Scenarios. Despite overall different performances on individual outcomes, models had similar performance on average percent of total flags for each scenario. In total, the greatest differences in performance were associated with the magnitude of the within-intersectional group standard deviation simulated in the data. All models performed better for most outcomes in small standard deviation scenarios compared to mixed and large standard deviation scenarios. This suggests that the amount of standard deviation within intersectional groups is related to how well models estimate the true value and detect a true effect if it is simulated in the data. In addition, there were differences for each model based on the number of categories included in the analysis. For most models, better performance was observed across outcomes for two categories compared to three. However, the categorical model had several outcomes where that pattern was reversed: type 1 error was always worse for two-category scenarios; coverage and bias were worse for two-category scenarios in mixed and small standard deviations. While there were some instances where outcomes seemed to differ regarding proportion representation or sample size, those findings were minimal compared to the changes associated with standard deviation and the number of categories included.

Comparison of Two and Three-Category Findings. The number of categories included in each model had a strong impact on statistical outcomes. In this study, I expected scenarios with two-categories to have better outcomes than those with three. However, that relationship was sometimes flipped, such that three-category scenarios showed less overall flagging. Models were not directly comparable between two and three-category scenarios because of the two separate data generation processes. However, due to the similarity of the data generation

processes and a desire to understand the impact of the number of intersectional groups, it is useful to discuss differences in the performance of models when there were two versus three-categories. In this section, I compare the two and three-category models on each outcome, specifically focusing on instances where the percentage of flagging was greater than ten-percentage points (see Tables 17 and 28).

` For accuracy, the percentage of flagging across all models was higher for three-category scenarios versus two-category scenarios such that models with more categories were less accurate. Bias followed a similar pattern for the interaction and MAIDHA models, where two-category scenarios had a lower overall percentage of flagging for bias. However, in the categorical model, bias was flagged less often for each model for three-category scenarios compared to the two-category scenarios. This pattern of observations may have occurred because the additional fixed effects accounted for different sources of variability in the data.

There were also surprising results for coverage. Coverage flagging was reduced by over ten percentage points for two versus three-category scenarios for the interaction and categorical models. This indicates that the three-category models had better coverage and that the true value appeared in the confidence interval more frequently. This may be due to various factors, such as: a) the fixed effects accounting for a greater proportion of the variance; or b) the amount of error surrounding the estimated coefficient/ intercepts resulted in wider confidence intervals, thus providing a larger interval for the true value to fall into. Power was flagged at substantially higher rates for three-category scenarios compared to two-category scenarios for the interaction and categorical models. This observation suggests that as more terms, and thus more true effect terms, are introduced, the models are less likely to detect all the true effects; a pattern which was expected for power.

Due to multiple comparisons, I expected type 1 error to increase when more terms were included in each model. The interaction and the MAIDHA models followed this pattern: there was greater type 1 error in the three compared to the two-category scenarios. However, for the categorical model, there was a surprising 31.72 percentage point decrease in flagging for type 1 error for the three compared to the two-category scenarios. It is unclear why this pattern may have occurred, but I present several hypotheses related to the A) relationship with power, B) amount of random variation, and C) overall sample size.

A) Relationship with power: this pattern may be due to the relationship between power and type 1 error; overall power was lower in the categorical model compared to the other two models. There were no substantial changes in power from the two and three-category scenarios in the categorical model, while there were in MAIDHA and interaction. The tradeoff of lower power may have led to a more conservative control of type 1 error for the categorical model, but this is unlikely to tell the full story.

B) Amount of random variation: type 1 error can be influenced by random variation in the data. Therefore, it's possible that the two different data generation strategies impacted the type 1 error for the categorical model in different ways. Therefore, it is possible that the data generation process favored the three-category categorical model compared to the two-category model.

C) Overall sample size: type 1 errors can be influenced by sample size. Thus, it is possible that with the three-category model, the sample size was not large enough to support the increased complexity, leading to less detectable differences.

Next steps for Applied Researchers

Given the complexity of the findings summarized above, this section aims to explore how the results may be used to advance methods of modeling intersectional analyses. Specifically, this section explores how researchers interested in using a method of intersectional analyses can consider and counter the limitations of within-group standard deviation, number of categories, the high type 1 error rate, and—specific to MAIDHA— the random effect statistical significance.

Within Intersectional Group Standard Deviation

Each model demonstrated better performance across outcomes when the within-group standard deviation was small. However, a small within-group standard deviation is unrealistic in most education contexts. In addition, the amount of variance in real contexts has more irregularity than the fixed values set in my study. For example, I used a mixed condition to help demonstrate how the overall mix of variances may influence parameters, where 10% of the groups had a high standard deviation. In a real-life context, within-group standard deviation will likely be more variable in mixed contexts. While the standard deviation scenarios in this study helped to demonstrate that the amount of within-group variability is influential, what was modeled is unlikely to be a scenario that matches the variability of lived experiences within intersectional groups.

Before beginning analysis, researchers must consider how the overall within-group standard deviation influences model parameters. Suppose within-group variability is large. Researchers can engage in *a priori* simulation studies designed to examine how the variance influences statistical outcomes. This can provide confidence in the results if the simulation study shows that the variance structure of their intersectional groups does not lead to inflated flagging

of outcomes. On the contrary, if the variance structure of intersectional groups negatively impacts modeling outcomes, they may wish to explore alternative methods of analysis.

Number of Categories

Findings from this study can be used by researchers to consider the implications the number of categories they plan to include in their analysis has on their model. This simulation study examined two and three-category scenarios. Within those categories, this study examined varying number of intersectional groups such that for two-category scenarios, there were a total of 14 intersectional groups and for three-category scenarios there were a total of 42 intersectional groups. In other contexts, the number of categories and number of groups within each category will differ. Additional research is needed to determine how these types of expected changes in research design influence model outcomes.

These results showed one version of incorporating intersectional groups. However, there are other ways that it is possible to explore three-categories of intersections without incorporating them all directly into a model. An alternative approach would be to take a subset of the data of just category A and perform an intersectional analysis of B and C. For example, if, based on theory, a researcher is interested in the experience of gender identity in academia, they may already recognize that the female gender identity has unique intersections with racialized identity and income status. So, instead of running an intersectional analysis on gender*race*income, they could opt to run an intersectional analysis of race*income within the female gender identity. While this will answer different research questions compared to the former approach, it would likely be more informative and yield better model outputs.

Type 1 Error Rate

Of all the outcomes, type 1 error had the most consistent flagging across models and scenarios. When there was sizeable within-group variation, type 1 error was almost always flagged. While this may be discouraging, there are approaches that can help mitigate type 1 error inflation. Relationships of intersectional groups that were previously hidden can be considered a new discovery and would benefit from the researcher selecting a lower alpha threshold. In a commentary authored by over 20 statisticians, Benjamin et al. (2018) recommend using an alpha threshold of .005 in instances of new discoveries to reduce the rate of false positives. In addition, researchers should utilize a p-value adjustment method to account for multiple comparisons, such as the B-H method of a family-wise procedure (Benjamini & Hochberg, 1995; Shaffer, 1995).

Significance in MAIDHA

The MAIDHA model presents challenges for determining significant differences between intersectional groups. In the MAIDHA model, intersectional groups were explored as random effects. In hierarchical linear modeling, it is unusual to develop a hypothesis and report a p-value for random effects. In fact, no such p-value is estimated by most software outputs. Instead of relying on p-values and test statistics, I created confidence intervals to examine whether there was a "significant" difference between intersectional groups. In this work around, I used the confidence intervals around the intercepts to consider whether the interval excluded "0" which would suggest an effect associated with that intersectional group's experience on the outcome. This approach of using estimated confidence intervals was imprecise. There are other alternatives researchers may choose to explore, such as leaning on high dimensional fixed effects to make decisions. Instead of relying on the level two random effects to inform a decision,

researchers could explore the fixed effects within each intercept by examining point estimates and the 95% confidence intervals surrounding them for each additive identity (within a given intersectional group).

Proportion Representation

The representation of intersectional groups is a challenge for researchers interested in exploring intersectionality. As I discussed in chapter two, it is common in education research to have small sample sizes in some demographic categories, and large sample sizes in others. This is due to the overall representation of all intersections in the population of this country and the systems of oppression that have caused disproportionate access to educational opportunities and attainment.

Researchers often cannot represent the universe of all possible intersections within their sample. When I started this work, I initially had smaller proportions suggested for some of the identity indicators in the extreme representation scenarios. However, when combined for two-way and three-way intersections, the odds that those identities would appear in the sample for even my largest size of n = 20,000 became quite small. For example, I initially had the extreme condition set that A1 = .005, B0 = .200, and C0 = .100. Therefore, the probability that someone of those three intersecting identities, A1*B0*C0, could be selected for the sample was p = .0001. Because this was a simulation study, I could manufacture the proportions. I raised them to understand how full representation would influence data properties. However, in real-life conditions, a researcher cannot increase the proportion when an intersectional group is absent in their study.

While I could achieve representation in many cases, that representation was often small. For example, I had one student who was A6*B1*C2 in school ID # 62. I allowed one student

from that school to be used to make inferences about the population of all students in that intersection represented in that school. Researchers must examine not only the statistical properties but also the implications for interpreting the results. It may not be appropriate, depending on the context, to include cases where a single student represents an intersectional group's experience in large-scale quantitative studies due to the risk of misrepresenting an entire intersectional group's experiences. Yet, a student's experiences should not be overlooked just because they exist at a unique intersection within a system of oppression. Researchers must tread carefully with their modeling and interpretation choices when making inferences about the population.

Researchers interested in applying an intersectional approach should consider the overall representation in their sample (or ideal sample) before selecting research questions and determining a research approach. Braun (2021) coined the term "carrying capacity of data" to provide a framework for examining the extent to which a data source can answer a given research question, is appropriate for a particular hypothesis, and can handle specific statistical approaches. This framework, or similar techniques to investigating a data source, should be applied before estimating a model. By scrutinizing the data in the planning stages, researchers can ensure they choose an appropriate model and avoid situations where they may jump to inappropriate conclusions.

Next Steps in Studying Methods of Modeling Intersectional Analyses

In this research I explored three methods of modeling intersectional analyses: interaction models, categorical models, and MAIDHA. These three methods were selected because they were appropriate for analyzing large-scale data when examining a single continuous outcome variable. When I ran each of these models, I varied four demographic characteristics to create

unique scenarios. The demographic data characteristics I chose to focus on were the number of demographic categories, the within-intersectional group standard deviation, the proportion of representation of identity indicators within demographic categories, and the overall sample size. Finally, I evaluated the performance of the models using five outcomes: accuracy, bias, coverage, power, and type 1 error.

In this study, I therefore made choices about models, demographic data characteristics, and outcomes. Other researchers may choose to study methods of modeling intersectional analyses and select other conditions or different demographic data characteristics to design a different study with the same goals as this dissertation. At this point, understanding how to best study intersectional experiences with oppression is novel and the ways to study methods of modeling intersectional analyses therefore feels infinite. In the following sections I recommend a few specific areas that would be of value for future research to focus on: model selection, clustered context, true value selection, type 1 error, and within-group standard deviation.

Model Selection

As I discussed in Chapter 2, the interaction model is limited in its theoretical alignment with intersectionality theory. However, I opted to examine it as a model in this study due to its widespread use. As I discuss under the upcoming Model Use section, a categorical model can apply in nearly any situation for which a researcher would select an interaction model. In addition to its limited theoretical alignment, the interaction model does not add much statistical value. Within the two-category scenarios, the interaction and categorical models performed similarly. In the three-category scenarios, the categorical model outperformed the interaction model in all outcomes except for coverage. Unless coverage is of the utmost concern for a given research design, I would suggest future studies drop the interaction model.

Outcome Selection and Criteria

In future studies, the flagging criteria for accuracy should be a smaller threshold or it could even be removed as an outcome. With a flagging criterion of .50 for accuracy, a bit of a contradiction was presented in the results: in many scenarios the results were deemed fairly accurate, but often times the confidence interval did not contain the true value. This is because the confidence interval in my simulated design was often narrower than the width of the accuracy threshold. So, if the flagging threshold were lowered for accuracy, accuracy and coverage would be more in line. However, the benefits of having accuracy as an outcome when both bias and coverage capture many of the facets that accuracy aims to is unclear to me. To narrow the focus of the study, and to make the interpretation of the outcomes simpler, accuracy could be removed as an outcome from future studies.

Clustered Context

I chose to explore the models specifically in a clustered data context. Previous simulations on intersectionality in quantitative research had only used non-clustered contexts. It would be useful to explore how models statistically compare when used in the context of a hierarchical linear model versus a single level in a study. For example, a study could explore the accuracy of intersectional estimates in a model clustered in school IDs (as I have done) but then explore single-level models within individual schools. While the two different models could not be used to address same research questions, knowing if one model version has better statistical properties would be useful in designing research questions specific to the capacity of the data.

I held my clustered data context constant throughout the simulations. The data generation process maintained 100 schools with intersectional groups distributed evenly within them. In addition, the ICC was held constant, though the standard deviation manipulations naturally

caused variability. Future studies might manipulate the number of schools, representation within schools, and the ICC. While I explored the models within a clustered context, they may not have represented the complexities and differences of, say, districts in a state. Rarely do we see schools in education where identities are consistently distributed across a state. Therefore, I suggest future research explores how different characteristics at the cluster level influence model parameters.

This study only examined identity-based fixed effects, and it did not include additional predictors outside of identity. In education contexts, there are many predictors that can influence the relationship between intersectional identity and a given outcome in education. It would be useful to study the impacts of the introduction of other predictors, specifically designed to "interact" with given intersectional identities. Specifically, this may be useful to explore in the MAIDHA model, as this would mean the model may have cross-level interactions. Because the random effect structure would have a greater interpretable value, I believe that contexts with cross-level interactions may increase the utility of MAIDHA in the education context. A researcher could determine how different intersectional group memberships influence the relationship between a given predictor variable and the outcome variable. Thus, exploring predictors may open up additional uses of this model in education research.

True Values

In this study, I did not intentionally set an effect size but rather randomly selected coefficient values from a range of |.2 to 2.0|. However, I believe that this is an additional avenue that can be explored; it would be useful to know how a larger range of effect sizes influences model parameters under a variety of other demographic data characteristics.

Similarly, I selected a set of true effects that were consistent across each scenario. Researchers could also change which intersectional groups have a true effect and how many groups have a true effect. This, combined with an exploration of representation, would be valuable. For example, a true effect for a group with a higher representation may be easier to detect than a true effect for a group with a lower representation. This would help better inform the conversation around representation and help researchers understand what representation is needed if they hope to see the effect of a particular group.

Type 1 Error

Type 1 error proved to be a consistent issue in each method of modeling intersectional analysis. In this chapter, I suggested ways to mitigate type 1 errors through p-value adjustments and raising the alpha level. Future simulation studies should examine the extent to which those methods are effective at improving type 1 error rates. Further, type 1 error had a surprising reversal in percent flagged from the two-category to the three-category scenarios for the categorical model. Research should attempt to replicate these findings and, if they hold, seek to determine possible reasons for this surprising trend in type 1 error flagging.

Within Group Standard Deviation

As I discussed earlier in this chapter, the way I designed my standard deviation scenarios is unlikely to match other lived-experience contexts. Given the influence of standard deviation, it would be useful for researchers to design additional standard deviation conditions to further understand the impact of large and mixed variability settings. It would be beneficial to design a simulation based on existing relationships between and within intersectional groups in an existing educational data context.

Model Use

This work specifically investigated statistical properties, but it did not explore interpretations of intersectional group's coefficients/ intercepts within the context of advantages and disadvantages with oppression. The overall purpose of this research was to inform decisions about the statistical utility of the models by identifying the statistical advantages and disadvantages of each method of modeling intersectional analysis under different demographic data characteristics. However, a decision about model selection should never be based on statistical properties alone; the results of this dissertation cannot be used absent the context and theory.

Alignment with Research Questions and Theory

Determining how to incorporate intersectionality into a quantitative research project should be based on multiple factors and driven by theory. While this dissertation focuses on the statistical components of three models, these results should not be considered outside of intersectional theory, research goals, or the data's carrying capacity. The model selection is not a cut-and-dry "Which performs best?" but rather, "Which performs best in the context of my research goals?" Each of the models supports slightly different research questions.

The interaction model supports research questions that explore the extent to which given additive identities interact. As discussed in Chapter 2, I do not believe this model aligns with intersectional theory due to its reliance on single axis categories of identities. In addition, it is not possible to explore all intersections of identity due to reference category exclusion, as discussed in Chapter 2. The categorical model improves the research questions that the interaction model may be asking since it avoids single-axis categories of identities while allowing all intersections of identity to be represented in an analysis. Both the categorical and interaction models place

intersectional groups in a point of comparison. Coming from "traditional" quantitative training, this is likely the direction researchers may lean towards. However, intersectionality requires pushing outside of traditional boundaries, and bodes well with a critical quantitative approach. Although it is important to recognize that simply comparing the magnitude of estimated coefficients across intersectionally formed groups does not, in itself, constitute intersectionality research, there are times when such comparisons are useful for informing understanding of how oppression operates different among intersectional groups (Bowleg et al., 2008; Lopez et al., 2018). Researchers must question the extent to which comparison-based questions are useful in their context, and how their research leads towards dismantling systems of oppression.

In contrast, the MAIDHA model answers very different research questions than the categorical and interaction models. The level two intercepts do provide an opportunity to assess magnitude of an intersectional group. While it's possible to examine how much above or below 0 a group falls, testing that difference is a challenge. It is less suitable for research questions in which identity groups are being directly compared based on significance since there is no direct way to test for the difference between random effects. However, suppose the difference between groups is not the focus of the main research question. In that case, the MAIDHA model may be suitable for intersectional approaches, as it does provide a unique intercept for each intersectional group. MAIDHA therefore deviates from many of the traditional quantitative approaches and offers opportunities to consider alternative ways of theorizing quantitative methods within an intersectional framework.

Recommended Model Use

The results of this study were complex, but there were some scenarios in which researchers with similar data structures may apply one or more of the methods of modeling

intersectional analyses with less limitations. In general, most models performed best with small within-group standard deviation and a lower number of demographic categories. In practice researchers will have different configurations of the number of demographic categories and the number of identity indicators within each demographic category. Therefore, it may be helpful to consider this with respect to the overall number of intersectional groups. In my two-category scenarios there were 14 intersectional groups, and in my three-category scenarios there were 42 intersectional groups. The interaction model is feasible to use with small standard deviations when there are up to 14 intersectional groups. However, researchers need to be weary of the risks of type 1 error. The categorical model is feasible for use with up to 42 intersectional groups with small within-group standard deviation. However, the categorical model also presented inflated type 1 error in those conditions. Finally, MAIDHA is recommended for use in up to 14 intersectional groups with small within-group standard deviation, but researchers need to be wary of bias.

Integration of Qualitative Data

Based on my results, any effort to examine intersectionality using the quantitative models explored in this study will likely produce substantial error. The small standard deviation scenarios led to promising results, but education contexts often see large amounts of within-group variation in demographic categories. Therefore, based on the findings from the simulation analyses presented above, I caution against a purely quantitative study. At this same time, I do not think quantitative results need to be avoided; they still hold promise to advance understanding of the impacts of oppression for intersectional groups of interest and support incorporating more voices in social science research. Specifically, I think quantitative intersectional analysis offers promise for exploratory-based designs where a researcher may use

quantitative intersectional analysis to identify potential differences that are then explored in greater detail through a qualitative study. Because of the limitations I found in my study, I suggest that quantitative results should be triangulated and backed by literature and theory.

Conclusion

In this dissertation, I was interested in how well each model estimated each intersectional group's coefficient or intercept. This approach differs from previous intersectional simulation studies (Mahendran et al., 2022b, 2022a), where the researchers investigated the model overall, instead of the coefficients/intercepts estimated by the model. In education, researchers are often interested in understanding demographic-group differences in a given outcome. Thus, knowing if the estimates from intersectional groups were accurate and unbiased is important for further transforming the capabilities of intersectionality in quantitative methods. Thus, this study expands the understanding of how methods of modeling intersectional analysis may be applied in educational contexts.

Specifically, this study expanded researchers' understanding of the advantages and disadvantages of methods of modeling intersectional analyses under different demographic data contexts. This study is the first to date to explore, in any capacity, the impact of within-group standard deviation, number of demographic categories, and proportion of representation within categories in methods of modeling intersectional analyses. Coupled with the inclusion of sample size, this exploration revealed the depth of impact that the number of categories and within-group standard deviation have on the statistical properties of estimated models.

The findings of this study expanded our current understanding of how demographic data characteristics influence the statistical parameters of method of modeling intersectional analyses. Until now, there has been no knowledge base regarding how the technical properties of estimated

coefficients are impacted by and vary across methods of developing an intersectional model. This dissertation opens more doors for questions than answers it provides, and it serves as a starting point for the continued study of intersectionality.

Works Cited

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review. *Journal of Applied Psychology*, *90*(1), 94–107. https://doi.org/10.1037/0021-9010.90.1.94

Akaike, H. (1974). *Factor analysis and AIC*. 16.

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*(2), 308–314. https://doi.org/10.1037/0033-2909.115.2.308

Anzaldúa, G. (1987). *Borderlands = la frontera: The new mestiza*. Spinsters/Aunt Lute.

Bambara, T. C. (1970). *The Black woman: An anthology*. Signet.

Bauer, G. R., Churchill, S. M., Mahendran, M., Walwyn, C., Lizotte, D., & Villa-Rueda, A. A. (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM - Population Health*, *14*, 100798. https://doi.org/10.1016/j.ssmph.2021.100798

Beal, F. (1970). Double Jeopardy: To be Black & female. In *Words of fire: An Anthology of African-American Feminist Thought* (pp. 146–155). The New Press.

Bell, A., Holman, D., & Jones, K. (2019). Using Shrinkage in Multilevel Models to Understand Intersectionality: A Simulation Study and a Guide for Best Practice. *Methodology*, *15*(2), 88–96. https://doi.org/10.1027/1614-2241/a000167

Bell, L. A. (2016). Theoretical foundations for social justice education. In *Teaching for diversity and social justice, 2nd ed.* (pp. 1–14). Routledge/Taylor & Francis Group.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. JSTOR.

Benjamin., (2018).Redefine Statistical Significance. *Nature Human Behavior, 2, 6-10,*

Bonilla-Silva, E. (1997). Rethinking Racism: Toward a Structural Interpretation. *American Sociological Review*, *62*(3), 465–480. https://doi.org/10.2307/2657316

Bowleg, L. (2008). When Black + lesbian + woman ≠ Black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles: A Journal of Research*, *59*(5–6), 312–325. https://doi.org/10.1007/s11199-008-9400-z

Bowleg, L. (2012). The problem with the phrase women and minorities: Intersectionality-an important theoretical framework for public health. *American Journal of Public Health*, *102*(7), 1267–1273. https://doi.org/10.2105/AJPH.2012.300750

Braun, H., (2021). Data in the Educational and Social Sciences: It's Time for Some Respect. *International Journal of Educational Methodology, 7(*3*),* 447-463. https://doi.org/10.12973/ijem.7.3.447

Braumoeller, B. F. (2004). Hypothesis Testing and Multiplicative Interaction Terms. *International Organization*, *58*(4), 807–820.

Bronfenbrenner, U., & Evans, G. W. (2000). Developmental Science in the 21st Century: Emerging Questions, Theoretical Models, Research Designs and Empirical Findings. *Social Development*, *9*(1), 115–125. https://doi.org/10.1111/1467-9507.00114

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*(3), 549–562. https://doi.org/10.1037/0033-2909.93.3.549

Carastathis, A., 1981- author. (2016). *Intersectionality: Origins, contestations, horizons*. University of Nebraska Press.

Cho, S., Crenshaw, K. W., & McCall, L. (2013). Toward a Field of Intersectionality Studies: Theory, Applications, and Praxis. *Signs: Journal of Women in Culture and Society*, *38*(4), 785–810. https://doi.org/10.1086/669608

Choo, H. Y., & Ferree, M. M. (2010). Practicing Intersectionality in Sociological Research: A Critical Analysis of Inclusions, Interactions, and Institutions in the Study of Inequalities. *Sociological Theory*, *28*(2), 129–149. https://doi.org/10.1111/j.1467-9558.2010.01370.x

Clarke, P., & Wheaton, B. (2007). Addressing Data Sparseness in Contextual Population Research: Using Cluster Analysis to Create Synthetic Neighborhoods. *Sociological Methods & Research*, *35*(3), 311–351. https://doi.org/10.1177/0049124106292362

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist*, *64*(3), 170–180. https://doi.org/10.1037/a0014564

Collins, P. H. (1986). Learning from the Outsider Within: The Sociological Significance of Black Feminist Thought. *Social Problems*, *33*(6), S14–S32. JSTOR. https://doi.org/10.2307/800672

Collins, P. H. (1991). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Collins, P. H. (2007). Pushing the Boundaries or Business as Usual? Race, Class, and Gender Studies and Sociological Inquiry. In *Sociology in America* (pp. 572–604). University of Chicago Press.

Collins, P. H. (2014). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment* (Revised tenth anniversary edition..). Routledge.

Collins, P. H. (2019). *Intersectionality as critical social theory*. Duke University Press.

Collins, P. H., & Bilge, S. (2016). *Intersectionality*. Polity Press.

Combahee River Collective,. (1986). *The Combahee River Collective statement: Black Feminist organizing in the seventies and eighties.*

Cooper, A. J. (Anna J., 1858-1964. (1982). *A voice from the South: By a Black woman of the South*. University of North Carolina at Chapel Hill Library.

Covarrubias, A., Nava, P. E., Lara, A., Burciaga, R., Vélez, V. N., & Solorzano, D. G. (2018). Critical race quantitative intersections: A *testimonio* analysis. *Race Ethnicity and Education*, *21*(2), 253–273. https://doi.org/10.1080/13613324.2017.1377412

CRC (Combahee River Collective). (1983). A Black Feminist Statement [first published in 1977]. In C. Moraga & G. Anzaldùa (Eds.), *This Bridge Called My Back: Wrtings by Radical Women of Color* (pp. 210–218). New York Kitchen Table: Women of Color Press.

Crenshaw, K. (1989a). *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*. 31.

Crenshaw, K. (1989b). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, *1989*, 139–168.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Daly, A., Dekker, T., & Hess, S. (2016). Dummy coding vs effects coding for categorical variables: Clarifications and extensions. *Standalone Technical Contributions in Choice Modelling*, *21*, 36–41. https://doi.org/10.1016/j.jocm.2016.09.005

Dillway, H., & Broman, C. (2001). Race, Class, and Gender Differences in Marital Satisfaction and Divisions of Household Labor Among Dual-Earner Couples: A Case for Intersectional Analysis. *Journal of Family Issues*, *22*(3), 309–327. https://doi.org/10.1177/019251301022003003

Evans, C. R. (2019a). Adding interactions to models of intersectional health inequalities: Comparing multilevel and conventional methods. *Social Science & Medicine*, *221*, 95–105. https://doi.org/10.1016/j.socscimed.2018.11.036

Evans, C. R. (2019b). Modeling the intersectionality of processes in the social production of health inequalities. *Social Science & Medicine*, *226*, 249–253. https://doi.org/10.1016/j.socscimed.2019.01.017

Evans, C. R. (2019c). Reintegrating contexts into quantitative intersectional analyses of health inequalities. *Health & Place*, *60*, 102214. https://doi.org/10.1016/j.healthplace.2019.102214

Evans, C. R., Leckie, G., & Merlo, J. (2020). Multilevel versus single-level regression for the analysis of multilevel information: The case of quantitative intersectional analysis. *Social Science & Medicine*, *245*, 112499. https://doi.org/10.1016/j.socscimed.2019.112499

Evans, C. R., Williams, D. R., Onnela, J.-P., & Subramanian, S. V. (2018). A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science & Medicine*, *203*, 64–73. https://doi.org/10.1016/j.socscimed.2017.11.011

Garry, A. (2011). Intersectionality, Metaphors, and the Multiplicity of Gender. *Hypatia*, *26*(4), 826–850.

Gideon Schwarz. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Hancock, A.-M. (2007). Intersectionality as a Normative and Empirical Paradigm. *Politics & Gender*, *3*(02). https://doi.org/10.1017/S1743923X07000062

Hancock, A.-M. (2016). *Intersectionality: An intellectual history*. Oxford University Press.

Hancock, A.-M. (2019). Empirical Intersectionality: A Tale of Two Approaches. In O. Hankivsky & J. S. Jordan-Zachery (Eds.), *The Palgrave Handbook of Intersectionality in Public Policy* (pp. 95–132). Springer International Publishing. https://doi.org/10.1007/978-3-319-98473-5_5

Hardy, M. A. (1993). *Regression with dummy variables.* (pp. vi, 90). Sage Publications, Inc.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hinze, S. W., Lin, J., & Andersson, T. E. (2012). Can we capture the intersections? Older Black women, education, and health. *Women's Health Issues: Official Publication of the Jacobs Institute of Women's Health*, *22*(1), e91-98. https://doi.org/10.1016/j.whi.2011.08.002

Hoffman, L., & Walters, R. W. (2022). Catching Up on Multilevel Modeling. *Annual Review of Psychology*, *73*(1), 659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Hooks, B. (2015). *Ain't I a woman: Black women and feminism*. Routledge.

Hox, J. (1998). Multilevel Modeling: When and Why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, Data Analysis, and Data Highways* (pp. 147–154). Springer Berlin Heidelberg.

Hurtado, S., Clayton-Pedersen, A. R., Allen, W. R., & Milem, J. F. (1998). Enhancing Campus Climates for Racial/Ethnic Diversity: Educational Policy and Practice. *The Review of Higher Education*, *21*(3), 279–302. https://doi.org/10.1353/rhe.1998.0003

Jaccard, J., & Turrisi, R. (2003). *Interaction Effects in Multiple Regression*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984522

James, A. (2008). Making Sense of Race and Racial Classification. In T. Zuberi & E. Bonilla-Silva (Eds.), *White Logic White Methods*. Rowman & Littlefield Publishers.

Jang, S. T. (2019). Schooling Experiences and Educational Outcomes of Latinx Secondary School Students Living at the Intersections of Multiple Social Constructs. *Urban Education*, 0042085919857793. https://doi.org/10.1177/0042085919857793

Johnson, O., & Jabbari, J. (2022). Suspended While Black in Majority White Schools: Implications for Math Efficacy and Equity. *The Educational Forum*, *86*(1), 26–50. https://doi.org/10.1080/00131725.2022.1997312

Jones, K., Johnston, R., & Manley, D. (2016). Uncovering interactions in multivariate contingency tables: A multi-level modelling exploratory approach. *Methodological Innovations*, *9*, 2059799116672874. https://doi.org/10.1177/2059799116672874

Kaplan, J. B. (2014). The Quality of Data on “Race” and “Ethnicity”: Implications for Health Researchers, Policy Makers, and Practitioners. *Race and Social Problems*, *6*(3), 214–236. https://doi.org/10.1007/s12552-014-9121-6

Kochhar, Rakesh. (2023). *The Enduring Grip of the Gender Pay Gap*. Pew Research Center. https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/

Lizotte, D. J., Mahendran, M., Churchill, S. M., & Bauer, G. R. (2020). Math versus meaning in MAIHDA: A commentary on multilevel statistical models for quantitative intersectionality. *Social Science & Medicine*, *245*, 112500. https://doi.org/10.1016/j.socscimed.2019.112500

López, N., Erwin, C., Binder, M., & Chavez, M. J. (2018). Making the invisible visible: Advancing quantitative methods in higher education using critical race theory and intersectionality. *Race Ethnicity and Education*, *21*(2), 180–207. https://doi.org/10.1080/13613324.2017.1375185

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). *Sufficient Sample Sizes for Multilevel Modeling*. *1*, 7.

Mahendran, M., Lizotte, D., & Bauer, G. R. (2022a). Describing Intersectional Health Outcomes: An Evaluation of Data Analysis Methods. *Epidemiology (Cambridge, Mass.)*, *33*(3), 395–405. https://doi.org/10.1097/EDE.0000000000001466

Mahendran, M., Lizotte, D., & Bauer, G. R. (2022b). Quantitative methods for descriptive intersectional analysis with binary health outcomes. *SSM - Population Health*, *17*, 101032. https://doi.org/10.1016/j.ssmph.2022.101032

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual Analysis through the Multilevel Linear Model. *Sociological Methodology*, *14*, 72–103. https://doi.org/10.2307/270903

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181–190. https://doi.org/10.1037/0033-2909.113.1.181

May, V. M. (2015). *Pursuing intersectionality, unsettling dominant imaginaries*. Routledge.

McCall, L. (2005). The Complexity of Intersectionality. *Signs*, *30*(3), 1771–1800. https://doi.org/10.1086/426800

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376–390. https://doi.org/10.1037/0033-2909.114.2.376

McLeroy, K. R., Bibeau, D., Steckler, A., & Glanz, K. (1988). An ecological perspective on health promotion programs. *Health Education Quarterly*, *15*(4), 351–377. https://doi.org/10.1177/109019818801500401

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Merlo, J. (2018). Multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) within an intersectional framework. *Social Science & Medicine*, *203*, 74–80. https://doi.org/10.1016/j.socscimed.2017.12.026

Misra, J., Curington, C. V., & Green, V. M. (2021). Methods of intersectional research. *Sociological Spectrum*, *41*(1), 9–28. https://doi.org/10.1080/02732173.2020.1791772

Nash, J. C., 1980- author. (2019). *Black feminism reimagined: After intersectionality*. Duke University Press.

Nissen, J. M., Her Many Horses, I., & Van Dusen, B. (2021). Investigating society's educational debts due to racism and sexism in student attitudes about physics using quantitative critical race theory. *Physical Review Physics Education Research*, *17*(1), 010116. https://doi.org/10.1103/PhysRevPhysEducRes.17.010116

Pearson, M. I., Castle, S. D., Matz, R. L., Koester, B. P., & Byrd, W. C. (2022). Integrating Critical Approaches into Quantitative STEM Equity Work. *CBE—Life Sciences Education*, *21*(1), es1. https://doi.org/10.1187/cbe.21-06-0158

Raudenbush, S. W. (1989). The analysis of longitudinal, multilevel data. *International Journal of Educational Research*, *13*(7), 721–740. https://doi.org/10.1016/0883-0355(89)90024-4

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17. https://doi.org/10.2307/2112482

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A Multilevel, Multivariate Model for Studying School Climate With Estimation Via the EM Algorithm and Application to U.S. High-School Data. *Journal of Educational Statistics*, *16*(4), 295–330. https://doi.org/10.3102/10769986016004295

Rhodes, W. (2010). Heterogeneous Treatment Effects: What Does a Regression Estimate? *Evaluation Review*, *34*(4), 334–361. https://doi.org/10.1177/0193841X10372890

Russell, M., Szendey, O., & Kaplan, L. (2021). An Intersectional Approach to DIF: Do Initial Findings Hold across Tests? *Educational Assessment*, *26*(4), 284–298. https://doi.org/10.1080/10627197.2021.1965473

Schudde, L. (2018). Heterogeneous Effects in Education: The Promise and Challenge of Incorporating Intersectionality Into Quantitative Methodological Approaches. *Review of Research in Education*, *42*(1), 72–92. https://doi.org/10.3102/0091732X18759040

Schulman, K. A., Berlin, J. A., Harless, W., Kerner, J. F., Sistrunk, S., Gersh, B. J., Dubé, R., Taleghani, C. K., Burke, J. E., Williams, S., Eisenberg, J. M., Ayers, W., & Escarce, J. J. (1999). The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization. *New England Journal of Medicine*, *340*(8), 618–626. https://doi.org/10.1056/NEJM199902253400806

Scott, N. A., & Siltanen, J. (2017). Intersectionality and quantitative methods: Assessing regression from a feminist perspective. *International Journal of Social Research Methodology*, *20*(4), 373–385. https://doi.org/10.1080/13645579.2016.1201328

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584. https://doi.org/10.1146/annurev.ps.46.020195.003021

Shields, S. A. (2008). Gender: An Intersectionality Perspective. *Sex Roles*, *59*(5–6), 301–311. https://doi.org/10.1007/s11199-008-9501-8

Spector, P. E., & Brannick, M. T. (2011). Methodological Urban Legends: The Misuse of Statistical Control Variables. *Organizational Research Methods*, *14*(2), 287–305. https://doi.org/10.1177/1094428110369842

Sweeney, R. E., & Ulveling, E. F. (1972). A Transformation for Simplifying the Interpretation of Coefficients of Binary Variables in Regression Analysis. *The American Statistician*, *26*(5), 30–32. https://doi.org/10.1080/00031305.1972.10478949

te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). When size matters: Advantages of weighted effect coding in observational

studies. *International Journal of Public Health*, *62*(1), 163–167. https://doi.org/10.1007/s00038-016-0901-1

U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), [2022]. Retrieved on [5/15/23]

Van Dusen, B., & Nissen, J. (2020). Associations between learning assistants, passing introductory physics, and equity: A quantitative critical race theory investigation. *Physical Review Physics Education Research*, *16*(1), 010117. https://doi.org/10.1103/PhysRevPhysEducRes.16.010117

Van Dusen, B., Nissen, J., Talbot, R. M., Huvard, H., & Shultz, M. (2022). A QuantCrit Investigation of Society's Educational Debts Due to Racism and Sexism in Chemistry Student Learning. *Journal of Chemical Education*, *99*(1), 25–34. https://doi.org/10.1021/acs.jchemed.1c00352

Viano, S., & Baker, D. J. (2020). How Administrative Data Collection and Analysis Can Better Reflect Racial and Ethnic Identities. *Review of Research in Education*, *44*(1), 301–331. https://doi.org/10.3102/0091732X20903321

Wilson, A. S. P., & Urick, A. (2022). An intersectional examination of the opportunity gap in science: A critical quantitative approach to latent class analysis. *Social Science Research*, *102*, 102645. https://doi.org/10.1016/j.ssresearch.2021.102645

Zinn, M. B., & Dill, B. T. (1996). Theorizing Difference from Multiracial Feminism. *Feminist Studies*, *22*(2), 321–331.

Zuberi, T. (2001). Thicker than Blood. In *Thicker than Blood How Racial Statistics Lie* (pp. xv–xxii). University of Minnesota Press. https://www.jstor.org/stable/10.5749/j.cttttcnc.5

Zuberi, T., & Bonilla-Silva, E. (2008). *White Logic, White Methods: Racism and Methodology*. Rowman & Littlefield Publishers. http://ebookcentral.proquest.com/lib/bostoncollege-ebooks/detail.action?docID=1343788

Appendix

Online Repository [https://github.com/oszendey/Intersectional-Analyses/blob/main/Appendix_Repository.xlsx]

Table 8

Two Categories: Average Accuracy Values for True Effect and No Effect Intersectional Groups, by Model

	Interaction		Categorical		MAIDHA	
	True Effect	*No Effect*	*True Effect*	*No Effect*	*True Effect*	*No Effect*
n_5000_p1_std_small	0.03	0.01	0.01	0.01	0.40	0.02
n_5000_p2_std_small	0.03	0.01	0.01	0.01	0.39	0.02
n_5000_p3_std_small	0.03	0.01	0.01	0.01	0.39	0.02
n_10000_p1_std_small	0.03	0.01	0.00	0.00	0.40	0.02
n_10000_p2_std_small	0.03	0.01	0.00	0.00	0.40	0.02
n_10000_p3_std_small	0.03	0.01	0.01	0.00	0.39	0.02
n_20000_p1_std_small	0.03	0.01	0.00	0.00	0.40	0.02
n_20000_p2_std_small	0.03	0.01	0.00	0.00	0.40	0.02
n_20000_p3_std_small	0.03	0.01	0.00	0.00	0.40	0.02
n_5000_p1_std_mixed	0.27	0.04	0.19	0.04	0.77	0.03
n_5000_p2_std_mixed	0.24	0.04	0.18	0.04	0.71	0.03
n_5000_p3_std_mixed	0.23	0.03	0.20	0.05	0.69	0.03
n_10000_p1_std_mixed	0.26	0.04	0.17	0.03	0.76	0.03
n_10000_p2_std_mixed	0.23	0.04	0.15	0.02	0.71	0.03
n_10000_p3_std_mixed	0.22	0.03	0.16	0.03	0.69	0.03
n_20000_p1_std_mixed	0.25	0.04	0.16	0.02	0.76	0.03
n_20000_p2_std_mixed	0.22	0.04	0.14	0.02	0.71	0.03
n_20000_p3_std_mixed	0.21	0.03	0.14	0.02	0.70	0.03
n_5000_p1_std_large	19.53	0.30	13.88	0.18	15.78	0.33
n_5000_p2_std_large	18.90	0.26	13.40	0.18	14.93	0.33
n_5000_p3_std_large	18.17	0.30	13.67	0.34	14.67	0.33
n_10000_p1_std_large	19.69	0.30	13.90	0.11	15.97	0.33
n_10000_p2_std_large	18.78	0.27	13.24	0.10	15.06	0.33
n_10000_p3_std_large	18.09	0.30	13.39	0.20	14.84	0.32
n_20000_p1_std_large	19.67	0.28	13.88	0.07	16.03	0.33
n_20000_p2_std_large	18.60	0.26	13.12	0.06	15.09	0.33
n_20000_p3_std_large	18.01	0.30	13.23	0.12	14.88	0.33

Note. Accuracy is shaded red if the average value is greater than 0.50 to represent flagged coefficients.

Table 9

Two Categories: Average Bias Values for True Effect and No Effect Intersectional Groups, by Model

	Interaction		Categorical		MAIDHA	
	True Effect	*No Effect*	*True Effect*	*No Effect*	*True Effect*	*No Effect*
n_5000_p1_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_5000_p2_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_5000_p3_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_10000_p1_std_small	0.00	0.00	0.00	0.00	0.26	0.23
n_10000_p2_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_10000_p3_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_20000_p1_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_20000_p2_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_20000_p3_std_small	0.00	0.00	0.00	0.00	0.27	0.23
n_5000_p1_std_mixed	0.42	0.14	0.37	0.13	0.42	0.29
n_5000_p2_std_mixed	0.39	0.12	0.34	0.12	0.40	0.28
n_5000_p3_std_mixed	0.38	0.11	0.34	0.11	0.39	0.28
n_10000_p1_std_mixed	0.41	0.12	0.35	0.12	0.42	0.28
n_10000_p2_std_mixed	0.38	0.12	0.33	0.12	0.40	0.28
n_10000_p3_std_mixed	0.36	0.11	0.32	0.11	0.39	0.28
n_20000_p1_std_mixed	0.40	0.12	0.35	0.12	0.42	0.28
n_20000_p2_std_mixed	0.37	0.12	0.33	0.12	0.40	0.28
n_20000_p3_std_mixed	0.36	0.11	0.32	0.11	0.39	0.28
n_5000_p1_std_large	4.32	0.18	3.63	0.17	2.00	1.29
n_5000_p2_std_large	4.26	0.15	3.57	0.14	1.95	1.27
n_5000_p3_std_large	4.16	0.25	3.59	0.21	1.93	1.27
n_10000_p1_std_large	4.34	0.17	3.65	0.17	2.01	1.29
n_10000_p2_std_large	4.24	0.14	3.56	0.14	1.96	1.27
n_10000_p3_std_large	4.15	0.25	3.57	0.22	1.94	1.27
n_20000_p1_std_large	4.34	0.17	3.65	0.16	2.02	1.29
n_20000_p2_std_large	4.22	0.14	3.55	0.14	1.96	1.27
n_20000_p3_std_large	4.14	0.25	3.57	0.22	1.94	1.28

Note. Bias that exceeded 1/2 SE(B^) but was less than 2SE(B^) was flagged as moderate (light red) and bias that exceeded 2SE(B^) was flagged as high (dark red).

Table 10

Two Categories: Average Coverage Percentages for True Effect and No Effect Intersectional Groups, by Model

	Interaction		Categorical		MAIDHA	
	True Effect	*No Effect*	*True Effect*	*No Effect*	*True Effect*	*No Effect*
n_5000_p1_std_small	0.00%	75.00%	95.94%	94.90%	100.00%	100.00%
n_5000_p2_std_small	0.00%	75.00%	95.34%	95.30%	100.00%	100.00%
n_5000_p3_std_small	0.00%	74.90%	95.82%	95.26%	100.00%	100.00%
n_10000_p1_std_small	0.00%	75.00%	94.88%	94.04%	100.00%	100.00%
n_10000_p2_std_small	0.00%	75.00%	95.22%	94.89%	100.00%	100.00%
n_10000_p3_std_small	0.00%	74.98%	95.16%	95.26%	100.00%	100.00%
n_20000_p1_std_small	0.00%	75.00%	94.98%	94.83%	100.00%	100.00%
n_20000_p2_std_small	0.00%	75.00%	94.82%	95.09%	100.00%	100.00%
n_20000_p3_std_small	0.00%	75.00%	95.06%	95.30%	100.00%	100.00%
n_5000_p1_std_mixed	0.00%	75.58%	39.60%	85.73%	100.00%	100.00%
n_5000_p2_std_mixed	0.00%	75.40%	45.36%	85.58%	100.00%	100.00%
n_5000_p3_std_mixed	1.65%	78.40%	56.20%	89.19%	100.00%	100.00%
n_10000_p1_std_mixed	0.00%	75.00%	25.98%	77.44%	100.00%	100.00%
n_10000_p2_std_mixed	0.00%	75.05%	31.46%	75.19%	100.00%	100.00%
n_10000_p3_std_mixed	0.05%	75.35%	42.52%	81.68%	100.00%	100.00%
n_20000_p1_std_mixed	0.00%	75.00%	14.88%	60.28%	100.00%	100.00%
n_20000_p2_std_mixed	0.00%	75.00%	24.38%	57.65%	100.00%	100.00%
n_20000_p3_std_mixed	0.00%	75.00%	29.32%	69.73%	100.00%	100.00%
n_5000_p1_std_large	0.00%	75.00%	0.00%	92.48%	100.00%	100.00%
n_5000_p2_std_large	0.00%	75.00%	0.00%	93.70%	100.00%	100.00%
n_5000_p3_std_large	1.45%	74.98%	1.36%	92.39%	100.00%	100.00%
n_10000_p1_std_large	0.00%	75.00%	0.00%	88.74%	100.00%	100.00%
n_10000_p2_std_large	0.00%	75.00%	0.00%	92.50%	100.00%	100.00%
n_10000_p3_std_large	0.10%	75.00%	0.04%	89.04%	100.00%	100.00%
n_20000_p1_std_large	0.00%	75.00%	0.00%	82.83%	100.00%	100.00%
n_20000_p2_std_large	0.00%	75.00%	0.00%	89.65%	100.00%	100.00%
n_20000_p3_std_large	0.00%	75.00%	0.00%	82.83%	100.00%	100.00%

Note. Average percentages were shaded light red to represent flagging if the true value was in the interval less than 92.5% of the time.

Table 11

Two Categories Average Power Rate (Percent) for True Effect Intersectional Groups, by Model

	Interaction	Categorical	MAIDHA
n_5000_p1_std_small	100.00%	100.00%	100.00%
n_5000_p2_std_small	100.00%	100.00%	100.00%
n_5000_p3_std_small	100.00%	99.78%	88.92%
n_10000_p1_std_small	100.00%	100.00%	100.00%
n_10000_p2_std_small	100.00%	100.00%	100.00%
n_10000_p3_std_small	100.00%	100.00%	97.88%
n_20000_p1_std_small	100.00%	100.00%	100.00%
n_20000_p2_std_small	100.00%	100.00%	100.00%
n_20000_p3_std_small	100.00%	100.00%	99.96%
n_5000_p1_std_mixed	100.00%	100.00%	99.92%
n_5000_p2_std_mixed	100.00%	99.92%	99.82%
n_5000_p3_std_mixed	100.00%	96.30%	84.10%
n_10000_p1_std_mixed	100.00%	100.00%	100.00%
n_10000_p2_std_mixed	100.00%	100.00%	100.00%
n_10000_p3_std_mixed	100.00%	99.78%	91.12%
n_20000_p1_std_mixed	100.00%	100.00%	100.00%
n_20000_p2_std_mixed	100.00%	100.00%	100.00%
n_20000_p3_std_mixed	100.00%	100.00%	98.92%
n_5000_p1_std_large	100.00%	100.00%	100.00%
n_5000_p2_std_large	100.00%	100.00%	100.00%
n_5000_p3_std_large	100.00%	99.82%	89.30%
n_10000_p1_std_large	100.00%	100.00%	100.00%
n_10000_p2_std_large	100.00%	100.00%	100.00%
n_10000_p3_std_large	100.00%	100.00%	98.14%
n_20000_p1_std_large	100.00%	100.00%	100.00%
n_20000_p2_std_large	100.00%	100.00%	100.00%
n_20000_p3_std_large	100.00%	100.00%	99.98%

Note. No values were shaded as all averages were above the 80% threshold.

Table 12

Two Categories: Average Type 1 Error Rate (Percent) for No Effect Intersectional Groups, by Model

	Interaction	Categorical	MAIDHA
n_5000_p1_std_small	4.50%	5.09%	11.30%
n_5000_p2_std_small	5.17%	4.69%	11.59%
n_5000_p3_std_small	4.93%	4.71%	12.41%
n_10000_p1_std_small	5.40%	5.96%	11.18%
n_10000_p2_std_small	5.27%	5.11%	11.28%
n_10000_p3_std_small	5.20%	4.74%	12.47%
n_20000_p1_std_small	5.57%	5.18%	11.11%
n_20000_p2_std_small	5.20%	4.91%	11.18%
n_20000_p3_std_small	4.83%	4.70%	11.68%
n_5000_p1_std_mixed	8.67%	14.26%	11.70%
n_5000_p2_std_mixed	10.43%	14.40%	13.86%
n_5000_p3_std_mixed	6.53%	10.80%	15.26%
n_10000_p1_std_mixed	11.63%	22.55%	12.02%
n_10000_p2_std_mixed	17.13%	24.81%	14.63%
n_10000_p3_std_mixed	8.63%	18.33%	17.86%
n_20000_p1_std_mixed	18.43%	39.71%	11.81%
n_20000_p2_std_mixed	27.83%	42.35%	14.66%
n_20000_p3_std_mixed	11.47%	30.28%	17.27%
n_5000_p1_std_large	5.20%	7.53%	11.37%
n_5000_p2_std_large	6.17%	6.28%	12.22%
n_5000_p3_std_large	5.87%	7.61%	15.34%
n_10000_p1_std_large	5.60%	11.25%	11.30%
n_10000_p2_std_large	5.87%	7.50%	11.60%
n_10000_p3_std_large	6.10%	10.95%	15.26%
n_20000_p1_std_large	5.70%	17.16%	11.11%
n_20000_p2_std_large	5.70%	10.35%	11.30%
n_20000_p3_std_large	5.47%	17.18%	14.03%

Note. Light red shading was used to represent flagging above the 5% threshold.

Table 13

Two Categories: Interaction Model Percentage of Flags by Outcome

	Accuracy	Bias	Coverage	Power	Type 1 Error
n_5000_p1_std_small	0.00%	0.00%	50.00%	0.00%	50.00%
n_5000_p2_std_small	0.00%	0.00%	50.00%	0.00%	75.00%
n_5000_p3_std_small	0.00%	0.00%	50.00%	0.00%	50.00%
n_10000_p1_std_small	0.00%	0.00%	50.00%	0.00%	75.00%
n_10000_p2_std_small	0.00%	0.00%	50.00%	0.00%	75.00%
n_10000_p3_std_small	0.00%	0.00%	50.00%	0.00%	75.00%
n_20000_p1_std_small	0.00%	0.00%	50.00%	0.00%	100.00%
n_20000_p2_std_small	0.00%	0.00%	50.00%	0.00%	75.00%
n_20000_p3_std_small	0.00%	0.00%	50.00%	0.00%	50.00%
n_5000_p1_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_5000_p2_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_5000_p3_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_10000_p1_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_10000_p2_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_10000_p3_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_20000_p1_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_20000_p2_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_20000_p3_std_mixed	0.00%	100.00%	50.00%	0.00%	100.00%
n_5000_p1_std_large	33.33%	50.00%	50.00%	0.00%	75.00%
n_5000_p2_std_large	33.33%	33.33%	50.00%	0.00%	100.00%
n_5000_p3_std_large	33.33%	66.67%	50.00%	0.00%	75.00%
n_10000_p1_std_large	33.33%	50.00%	50.00%	0.00%	100.00%
n_10000_p2_std_large	33.33%	33.33%	50.00%	0.00%	75.00%
n_10000_p3_std_large	33.33%	66.67%	50.00%	0.00%	100.00%
n_20000_p1_std_large	33.33%	50.00%	50.00%	0.00%	100.00%
n_20000_p2_std_large	33.33%	33.33%	50.00%	0.00%	100.00%
n_20000_p3_std_large	33.33%	66.67%	50.00%	0.00%	50.00%

Note. Percentages were calculated for each cell, out of the total possible flags for that cell.

Table 14

Two Categories: Categorical Model Percentage of Flags by Outcome

	Accuracy	Bias	Coverage	Power	Type 1 Error
n_5000_p1_std_small	0.00%	0.00%	0.00%	0.00%	62.50%
n_5000_p2_std_small	0.00%	0.00%	0.00%	0.00%	25.00%
n_5000_p3_std_small	0.00%	0.00%	0.00%	0.00%	25.00%
n_10000_p1_std_small	0.00%	0.00%	0.00%	0.00%	75.00%
n_10000_p2_std_small	0.00%	0.00%	0.00%	0.00%	37.50%
n_10000_p3_std_small	0.00%	0.00%	0.00%	0.00%	25.00%
n_20000_p1_std_small	0.00%	0.00%	0.00%	0.00%	50.00%
n_20000_p2_std_small	0.00%	0.00%	0.00%	0.00%	62.50%
n_20000_p3_std_small	0.00%	0.00%	0.00%	0.00%	25.00%
n_5000_p1_std_mixed	0.00%	100.00%	100.00%	0.00%	100.00%
n_5000_p2_std_mixed	0.00%	92.31%	92.31%	0.00%	100.00%
n_5000_p3_std_mixed	0.00%	76.92%	69.23%	0.00%	100.00%
n_10000_p1_std_mixed	0.00%	100.00%	100.00%	0.00%	100.00%
n_10000_p2_std_mixed	0.00%	100.00%	100.00%	0.00%	100.00%
n_10000_p3_std_mixed	0.00%	84.62%	92.31%	0.00%	100.00%
n_20000_p1_std_mixed	0.00%	100.00%	100.00%	0.00%	100.00%
n_20000_p2_std_mixed	0.00%	100.00%	100.00%	0.00%	100.00%
n_20000_p3_std_mixed	0.00%	92.31%	92.31%	0.00%	100.00%
n_5000_p1_std_large	38.46%	69.23%	61.54%	0.00%	87.50%
n_5000_p2_std_large	38.46%	46.15%	53.85%	0.00%	75.00%
n_5000_p3_std_large	61.54%	69.23%	61.54%	0.00%	100.00%
n_10000_p1_std_large	38.46%	69.23%	84.62%	0.00%	100.00%
n_10000_p2_std_large	38.46%	69.23%	61.54%	0.00%	87.50%
n_10000_p3_std_large	38.46%	84.62%	84.62%	0.00%	100.00%
n_20000_p1_std_large	38.46%	84.62%	84.62%	0.00%	100.00%
n_20000_p2_std_large	38.46%	84.62%	76.92%	0.00%	100.00%
n_20000_p3_std_large	38.46%	92.31%	92.31%	0.00%	100.00%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 15

Two Categories: MAIDHA Model Percentage of Flags by Outcome

	Accuracy	Bias	Coverage	Power	Type 1 Error
n_5000_p1_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_5000_p2_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_5000_p3_std_small	14.29%	71.43%	0.00%	20.00%	11.11%
n_10000_p1_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_10000_p2_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_10000_p3_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_20000_p1_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_20000_p2_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_20000_p3_std_small	14.29%	71.43%	0.00%	0.00%	11.11%
n_5000_p1_std_mixed	14.29%	78.57%	0.00%	0.00%	11.11%
n_5000_p2_std_mixed	14.29%	71.43%	0.00%	0.00%	33.33%
n_5000_p3_std_mixed	14.29%	78.57%	0.00%	20.00%	44.44%
n_10000_p1_std_mixed	14.29%	78.57%	0.00%	0.00%	11.11%
n_10000_p2_std_mixed	14.29%	71.43%	0.00%	0.00%	33.33%
n_10000_p3_std_mixed	14.29%	78.57%	0.00%	20.00%	44.44%
n_20000_p1_std_mixed	14.29%	78.57%	0.00%	0.00%	11.11%
n_20000_p2_std_mixed	14.29%	78.57%	0.00%	0.00%	33.33%
n_20000_p3_std_mixed	14.29%	78.57%	0.00%	0.00%	44.44%
n_5000_p1_std_large	42.86%	78.57%	0.00%	0.00%	11.11%
n_5000_p2_std_large	42.86%	92.86%	0.00%	0.00%	11.11%
n_5000_p3_std_large	42.86%	92.86%	0.00%	20.00%	55.56%
n_10000_p1_std_large	42.86%	85.71%	0.00%	0.00%	11.11%
n_10000_p2_std_large	42.86%	85.71%	0.00%	0.00%	11.11%
n_10000_p3_std_large	42.86%	92.86%	0.00%	0.00%	55.56%
n_20000_p1_std_large	42.86%	71.43%	0.00%	0.00%	11.11%
n_20000_p2_std_large	42.86%	92.86%	0.00%	0.00%	11.11%
n_20000_p3_std_large	42.86%	92.86%	0.00%	0.00%	44.44%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 16

Two Categories: Percentage of Flags, Averaged Across Outcomes

	Interaction	Categorical	MAIDHA
n_5000_p1_std_small	20.00%	12.50%	19.37%
n_5000_p2_std_small	25.00%	5.00%	19.37%
n_5000_p3_std_small	20.00%	5.00%	23.37%
n_10000_p1_std_small	25.00%	15.00%	19.37%
n_10000_p2_std_small	25.00%	7.50%	19.37%
n_10000_p3_std_small	25.00%	5.00%	19.37%
n_20000_p1_std_small	30.00%	10.00%	19.37%
n_20000_p2_std_small	25.00%	12.50%	19.37%
n_20000_p3_std_small	20.00%	5.00%	19.37%
n_5000_p1_std_mixed	50.00%	60.00%	20.79%
n_5000_p2_std_mixed	50.00%	56.92%	23.81%
n_5000_p3_std_mixed	50.00%	49.23%	31.46%
n_10000_p1_std_mixed	50.00%	60.00%	20.79%
n_10000_p2_std_mixed	50.00%	60.00%	23.81%
n_10000_p3_std_mixed	50.00%	55.39%	31.46%
n_20000_p1_std_mixed	50.00%	60.00%	20.79%
n_20000_p2_std_mixed	50.00%	60.00%	25.24%
n_20000_p3_std_mixed	50.00%	56.92%	27.46%
n_5000_p1_std_large	41.67%	51.35%	26.51%
n_5000_p2_std_large	43.33%	42.69%	29.37%
n_5000_p3_std_large	45.00%	58.46%	42.26%
n_10000_p1_std_large	46.67%	58.46%	27.94%
n_10000_p2_std_large	38.33%	51.35%	27.94%
n_10000_p3_std_large	50.00%	61.54%	38.26%
n_20000_p1_std_large	46.67%	61.54%	25.08%
n_20000_p2_std_large	43.33%	60.00%	29.37%
n_20000_p3_std_large	40.00%	64.62%	36.03%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 17

Two Categories: Percentage of Flagged Coefficients/Intercepts by Model and Outcome

	Interaction	Categorical	MAIDHA
Bias	50.00%	56.13%	78.57%
Accuracy	11.11%	13.68%	23.81%
Coverage	50.00%	55.84%	0.00%
Power	0.00%	0.00%	2.96%
Type1Error	85.19%	79.17%	21.90%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 18

AIC and BIC Values, by Model

	Interaction		Categorical		MAIDHA	
	AIC	BIC	AIC	BIC	AIC	BIC
n_5000_p1_std_small	14381	14485	14381	14485	14391	14463
n_5000_p2_std_small	14376	14480	14376	14480	14386	14458
n_5000_p3_std_small	14370	14474	14370	14474	14380	14452
n_10000_p1_std_small	28778	28893	28778	28893	28788	28867
n_10000_p2_std_small	28773	28888	28773	28888	28783	28862
n_10000_p3_std_small	28767	28882	28767	28882	28777	28857
n_20000_p1_std_small	57550	57676	57550	57676	57560	57647
n_20000_p2_std_small	57545	57672	57545	57672	57555	57642
n_20000_p3_std_small	57539	57666	57539	57666	57550	57637
n_5000_p1_std_mixed	21083	21188	21083	21188	21098	21169
n_5000_p2_std_mixed	20806	20911	20806	20911	20820	20892
n_5000_p3_std_mixed	20728	20833	20728	20833	20742	20814
n_10000_p1_std_mixed	41942	42057	41942	42057	41956	42035
n_10000_p2_std_mixed	41501	41616	41501	41616	41515	41594
n_10000_p3_std_mixed	41358	41473	41358	41473	41371	41451
n_20000_p1_std_mixed	83753	83879	83753	83879	83767	83854
n_20000_p2_std_mixed	82914	83041	82914	83041	82928	83015
n_20000_p3_std_mixed	82602	82728	82602	82728	82616	82702
n_5000_p1_std_large	30630	30735	30630	30735	30660	30732
n_5000_p2_std_large	30552	30657	30552	30657	30582	30654
n_5000_p3_std_large	30473	30577	30473	30577	30502	30574
n_10000_p1_std_large	61105	61220	61105	61220	61135	61214
n_10000_p2_std_large	60992	61108	60992	61108	61022	61101
n_10000_p3_std_large	60852	60968	60852	60968	60882	60961
n_20000_p1_std_large	122023	122150	122023	122150	122053	122140
n_20000_p2_std_large	121850	121977	121850	121977	121880	121967
n_20000_p3_std_large	121593	121720	121593	121720	121623	121710

Table 19

Three Categories: Average Accuracy Values for No Effect and True Effect Intersectional Groups, by Model

	Interaction		Categorical		MAIDHA	
	True Effect	*No Effect*	*True Effect*	*No Effect*	*True Effect*	*No Effect*
n_5000_p1_std_small	0.09	0.04	0.02	0.02	0.43	0.20
n_5000_p2_std_small	0.06	0.02	0.04	0.03	0.37	0.19
n_10000_p1_std_small	0.09	0.04	0.01	0.01	0.45	0.21
n_10000_p2_std_small	0.06	0.02	0.02	0.02	0.41	0.20
n_20000_p1_std_small	0.09	0.04	0.00	0.00	0.45	0.21
n_20000_p2_std_small	0.06	0.02	0.01	0.01	0.43	0.20
n_20000_p3_std_small	0.05	0.01	0.02	0.02	0.39	0.20
n_5000_p1_std_mixed	0.51	0.21	0.29	0.10	0.88	0.42
n_5000_p2_std_mixed	0.51	0.11	0.77	0.20	0.84	0.47
n_10000_p1_std_mixed	0.47	0.21	0.23	0.07	0.88	0.43
n_10000_p2_std_mixed	0.49	0.10	0.55	0.12	0.93	0.50
n_20000_p1_std_mixed	0.44	0.20	0.20	0.06	0.88	0.43
n_20000_p2_std_mixed	0.46	0.10	0.43	0.08	0.96	0.51
n_20000_p3_std_mixed	0.21	0.19	0.25	0.15	0.73	0.43
n_5000_p1_std_large	22.97	3.51	13.76	1.99	10.76	4.55
n_5000_p2_std_large	12.99	1.73	17.02	3.18	8.83	4.25
n_10000_p1_std_large	23.76	3.64	14.02	1.87	11.37	4.84
n_10000_p2_std_large	15.54	1.96	18.26	3.14	11.43	5.24
n_20000_p1_std_large	24.13	3.69	14.13	1.82	11.65	4.99
n_20000_p2_std_large	16.59	2.06	18.29	3.08	12.44	5.67
n_20000_p3_std_large	8.80	4.62	10.77	5.35	9.69	4.77

Note. Accuracy is shaded red if the average value is greater than 0.50 to represent flagged coefficients.

Table 20

Three Categories: Average Bias Values for No Effect and True Effect Intersectional Groups, by Model

	Interaction		Categorical		MAIDHA	
	True Effect	*No Effect*	*True Effect*	*No Effect*	*True Effect*	*No Effect*
n_5000_p1_std_small	0.18	0.18	0.00	0.00	0.65	0.28
n_5000_p2_std_small	0.08	0.08	0.00	0.00	0.61	0.27
n_10000_p1_std_small	0.18	0.18	0.00	0.00	0.65	0.28
n_10000_p2_std_small	0.08	0.08	0.00	0.00	0.63	0.28
n_20000_p1_std_small	0.18	0.18	0.00	0.00	0.66	0.29
n_20000_p2_std_small	0.08	0.08	0.00	0.00	0.65	0.28
n_20000_p3_std_small	0.06	0.06	0.00	0.00	0.63	0.28
n_5000_p1_std_mixed	0.53	0.41	0.41	0.13	0.91	0.41
n_5000_p2_std_mixed	0.51	0.24	0.58	0.15	0.91	0.42
n_10000_p1_std_mixed	0.51	0.41	0.40	0.13	0.91	0.41
n_10000_p2_std_mixed	0.51	0.24	0.56	0.15	0.96	0.44
n_20000_p1_std_mixed	0.50	0.40	0.39	0.13	0.92	0.41
n_20000_p2_std_mixed	0.49	0.23	0.54	0.15	0.97	0.44
n_20000_p3_std_mixed	0.31	0.26	0.35	0.19	0.87	0.41
n_5000_p1_std_large	3.31	1.94	3.10	0.77	3.10	1.36
n_5000_p2_std_large	2.66	1.29	3.35	0.95	2.85	1.29
n_10000_p1_std_large	3.37	1.99	3.17	0.78	3.19	1.41
n_10000_p2_std_large	2.89	1.40	3.60	1.05	3.22	1.45
n_20000_p1_std_large	3.40	2.00	3.19	0.78	3.23	1.43
n_20000_p2_std_large	2.97	1.45	3.67	1.09	3.37	1.52
n_20000_p3_std_large	2.13	1.36	2.64	1.16	3.02	1.39

Note. Bias that exceeded 1/2 SE(Bˆ) but was less than 2SE(Bˆ) was flagged as moderate (light red) and bias that exceeded 2SE(Bˆ) was flagged as high (dark red).

Table 21

Three Categories: Average Percent of coverage for True Effect and No Effect Intersectional Groups, by Model

	Interaction		Categorical		MAIDHA	
	True Effect	No Effect	True Effect	No Effect	True Effect	No Effect
n_5000_p1_std_small	99.96%	99.93%	61.45%	85.36%	86.67%	100.00%
n_5000_p2_std_small	100.00%	100.00%	62.59%	85.68%	90.14%	99.99%
n_10000_p1_std_small	99.52%	99.59%	59.34%	85.34%	84.58%	100.00%
n_10000_p2_std_small	100.00%	100.00%	61.99%	85.52%	88.06%	100.00%
n_20000_p1_std_small	96.02%	95.40%	57.13%	85.46%	82.74%	100.00%
n_20000_p2_std_small	99.98%	100.00%	60.71%	85.29%	86.08%	100.00%
n_20000_p3_std_small	100.00%	100.00%	60.07%	85.75%	87.52%	100.00%
n_5000_p1_std_mixed	91.80%	95.99%	44.13%	87.68%	94.64%	99.57%
n_5000_p2_std_mixed	95.98%	99.41%	46.58%	85.89%	94.28%	99.31%
n_10000_p1_std_mixed	76.00%	88.69%	31.90%	85.17%	95.43%	99.67%
n_10000_p2_std_mixed	86.44%	97.07%	36.52%	82.67%	94.87%	99.57%
n_20000_p1_std_mixed	49.80%	85.49%	21.84%	81.76%	96.82%	99.89%
n_20000_p2_std_mixed	64.58%	90.19%	27.04%	78.04%	96.00%	99.77%
n_20000_p3_std_mixed	95.04%	85.79%	40.13%	80.48%	95.82%	99.37%
n_5000_p1_std_large	35.96%	85.74%	23.01%	71.18%	87.48%	100.00%
n_5000_p2_std_large	54.12%	95.06%	24.09%	74.67%	92.93%	99.93%
n_10000_p1_std_large	17.76%	85.19%	16.53%	64.73%	86.23%	100.00%
n_10000_p2_std_large	18.90%	85.96%	12.54%	64.35%	91.19%	100.00%
n_20000_p1_std_large	1.26%	80.26%	14.94%	61.48%	86.58%	100.00%
n_20000_p2_std_large	2.64%	84.73%	8.18%	55.28%	89.48%	100.00%
n_20000_p3_std_large	27.96%	85.71%	20.44%	71.07%	91.12%	99.99%

Note. Coverage was flagged when the true effect was in the 95% confidence interval less than 92.5% of the time.

Table 22

Three Categories: Average Power Rate (Percent) for True Effect Intersectional Groups, by Model

	Interaction	Categorical	MAIDHA
n_5000_p1_std_small	81.06%	68.23%	76.06%
n_5000_p2_std_small	93.98%	68.22%	73.13%
n_10000_p1_std_small	81.08%	68.44%	80.73%
n_10000_p2_std_small	98.60%	68.14%	79.06%
n_20000_p1_std_small	81.32%	68.27%	84.40%
n_20000_p2_std_small	99.94%	68.18%	82.91%
n_20000_p3_std_small	81.50%	68.10%	73.83%
n_5000_p1_std_mixed	81.72%	72.61%	73.96%
n_5000_p2_std_mixed	86.18%	69.37%	69.33%
n_10000_p1_std_mixed	83.08%	75.98%	78.48%
n_10000_p2_std_mixed	90.46%	76.81%	78.59%
n_20000_p1_std_mixed	86.02%	77.37%	84.13%
n_20000_p2_std_mixed	97.14%	80.81%	81.86%
n_20000_p3_std_mixed	82.04%	76.43%	73.57%
n_5000_p1_std_large	82.24%	78.23%	75.43%
n_5000_p2_std_large	89.74%	77.16%	71.34%
n_10000_p1_std_large	83.90%	82.05%	81.10%
n_10000_p2_std_large	95.86%	83.30%	79.42%
n_20000_p1_std_large	87.04%	83.48%	85.58%
n_20000_p2_std_large	99.60%	83.80%	82.90%
n_20000_p3_std_large	83.64%	73.37%	76.06%

Note. Power was flagged when the percentage of true effects detected was less than 80%, for coefficients/ intercepts designed to have an effect.

Table 23

Three Categories: Average Type 1 Error Rate (Percent) for No Effect Intersectional Groups, by Model

	Interaction	Categorical	MAIDHA
n_5000_p1_std_small	60.50%	17.88%	60.04%
n_5000_p2_std_small	39.23%	17.61%	48.44%
n_10000_p1_std_small	76.03%	17.91%	70.48%
n_10000_p2_std_small	47.66%	17.77%	57.27%
n_20000_p1_std_small	86.27%	17.79%	77.33%
n_20000_p2_std_small	59.04%	18.09%	65.63%
n_20000_p3_std_small	39.43%	17.38%	57.11%
n_5000_p1_std_mixed	42.34%	17.44%	45.69%
n_5000_p2_std_mixed	31.04%	18.77%	41.68%
n_10000_p1_std_mixed	54.84%	21.06%	60.81%
n_10000_p2_std_mixed	38.86%	22.31%	52.09%
n_20000_p1_std_mixed	71.04%	24.27%	70.27%
n_20000_p2_std_mixed	47.76%	26.61%	60.66%
n_20000_p3_std_mixed	35.00%	24.99%	51.28%
n_5000_p1_std_large	58.44%	30.91%	57.85%
n_5000_p2_std_large	37.04%	28.81%	47.70%
n_10000_p1_std_large	77.60%	38.37%	70.87%
n_10000_p2_std_large	47.16%	41.14%	60.26%
n_20000_p1_std_large	89.41%	42.09%	78.28%
n_20000_p2_std_large	58.71%	50.10%	69.44%
n_20000_p3_std_large	40.50%	34.24%	61.79%

Note. Type 1 error was flagged when effects were detected over 5% of the time, for coefficients/ intercepts designed to have no effect.

Table 24

Three Categories: Interaction Model Percentage of Flags by Outcome

	Accuracy	Bias	Coverage	Power	Type1Error
n_5000_p1_std_small	0.00%	100.00%	0.00%	20.00%	100.00%
n_5000_p2_std_small	0.00%	66.67%	0.00%	20.00%	85.71%
n_10000_p1_std_small	0.00%	100.00%	0.00%	20.00%	100.00%
n_10000_p2_std_small	0.00%	66.67%	0.00%	0.00%	100.00%
n_20000_p1_std_small	0.00%	100.00%	0.00%	20.00%	100.00%
n_20000_p2_std_small	0.00%	66.67%	0.00%	0.00%	85.71%
n_20000_p3_std_small	0.00%	66.67%	0.00%	20.00%	100.00%
n_5000_p1_std_mixed	25.00%	100.00%	16.67%	20.00%	85.71%
n_5000_p2_std_mixed	16.67%	75.00%	8.33%	20.00%	85.71%
n_10000_p1_std_mixed	25.00%	100.00%	33.33%	20.00%	100.00%
n_10000_p2_std_mixed	16.67%	83.33%	25.00%	20.00%	100.00%
n_20000_p1_std_mixed	25.00%	100.00%	41.67%	20.00%	100.00%
n_20000_p2_std_mixed	16.67%	83.33%	33.33%	0.00%	100.00%
n_20000_p3_std_mixed	8.33%	83.33%	16.67%	20.00%	100.00%
n_5000_p1_std_large	91.67%	100.00%	41.67%	20.00%	100.00%
n_5000_p2_std_large	66.67%	83.33%	41.67%	20.00%	71.43%
n_10000_p1_std_large	91.67%	100.00%	50.00%	20.00%	100.00%
n_10000_p2_std_large	66.67%	83.33%	50.00%	20.00%	71.43%
n_20000_p1_std_large	91.67%	100.00%	58.33%	20.00%	100.00%
n_20000_p2_std_large	66.67%	83.33%	50.00%	0.00%	71.43%
n_20000_p3_std_large	58.33%	91.67%	50.00%	20.00%	100.00%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 25

Three Categories: Categorical Model Percentage of Flags by Outcome

	Accuracy	Bias	Coverage	Power	Type1Error
n_5000_p1_std_small	0.00%	0.00%	19.51%	0.00%	31.03%
n_5000_p2_std_small	0.00%	0.00%	19.51%	0.00%	48.28%
n_10000_p1_std_small	0.00%	0.00%	19.51%	0.00%	27.59%
n_10000_p2_std_small	0.00%	0.00%	19.51%	0.00%	17.24%
n_20000_p1_std_small	0.00%	0.00%	19.51%	0.00%	34.48%
n_20000_p2_std_small	0.00%	0.00%	19.51%	0.00%	34.48%
n_20000_p3_std_small	0.00%	0.00%	19.51%	0.00%	37.93%
n_5000_p1_std_mixed	9.76%	48.78%	31.71%	0.00%	13.79%
n_5000_p2_std_mixed	19.51%	53.66%	46.34%	8.33%	55.17%
n_10000_p1_std_mixed	9.76%	63.41%	46.34%	0.00%	41.38%
n_10000_p2_std_mixed	12.20%	63.41%	58.54%	0.00%	62.07%
n_20000_p1_std_mixed	7.32%	78.05%	58.54%	0.00%	58.62%
n_20000_p2_std_mixed	12.20%	75.61%	68.29%	0.00%	72.41%
n_20000_p3_std_mixed	9.76%	70.73%	78.05%	8.33%	82.76%
n_5000_p1_std_large	51.22%	58.54%	48.78%	16.67%	37.93%
n_5000_p2_std_large	78.05%	60.98%	60.98%	25.00%	48.28%
n_10000_p1_std_large	48.78%	73.17%	58.54%	8.33%	48.28%
n_10000_p2_std_large	68.29%	65.85%	60.98%	0.00%	62.07%
n_20000_p1_std_large	48.78%	75.61%	63.41%	0.00%	68.97%
n_20000_p2_std_large	65.85%	68.29%	68.29%	0.00%	62.07%
n_20000_p3_std_large	65.85%	58.54%	60.98%	16.67%	51.72%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 26

Three Categories: MAIDHA Model Percentage of Flags by Outcome

	Accuracy	Bias	Coverage	Power	Type1Error
n_5000_p1_std_small	19.05%	90.48%	7.14%	33.33%	93.33%
n_5000_p2_std_small	19.05%	90.48%	7.14%	33.33%	80.00%
n_10000_p1_std_small	19.05%	92.86%	7.14%	25.00%	93.33%
n_10000_p2_std_small	19.05%	92.86%	7.14%	33.33%	90.00%
n_20000_p1_std_small	19.05%	95.24%	7.14%	16.67%	93.33%
n_20000_p2_std_small	19.05%	92.86%	7.14%	16.67%	93.33%
n_20000_p3_std_small	19.05%	92.86%	7.14%	33.33%	96.67%
n_5000_p1_std_mixed	28.57%	88.10%	7.14%	33.33%	80.00%
n_5000_p2_std_mixed	28.57%	85.71%	14.29%	50.00%	83.33%
n_10000_p1_std_mixed	28.57%	88.10%	7.14%	33.33%	86.67%
n_10000_p2_std_mixed	28.57%	85.71%	9.52%	25.00%	93.33%
n_20000_p1_std_mixed	28.57%	88.10%	4.76%	25.00%	90.00%
n_20000_p2_std_mixed	28.57%	90.48%	7.14%	25.00%	93.33%
n_20000_p3_std_mixed	28.57%	90.48%	11.90%	33.33%	100.00%
n_5000_p1_std_large	83.33%	90.48%	7.14%	33.33%	80.00%
n_5000_p2_std_large	80.95%	88.10%	7.14%	33.33%	90.00%
n_10000_p1_std_large	85.71%	90.48%	7.14%	25.00%	93.33%
n_10000_p2_std_large	88.10%	88.10%	7.14%	25.00%	96.67%
n_20000_p1_std_large	85.71%	90.48%	7.14%	16.67%	93.33%
n_20000_p2_std_large	88.10%	88.10%	7.14%	25.00%	96.67%
n_20000_p3_std_large	83.33%	88.10%	7.14%	25.00%	96.67%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 27

Three Categories: Percentage of Flags, Averaged Across Outcomes

	Interaction	Categorical	MAIDHA
n_5000_p1_std_small	44.00%	10.11%	48.67%
n_5000_p2_std_small	34.48%	13.56%	46.00%
n_10000_p1_std_small	44.00%	9.42%	47.48%
n_10000_p2_std_small	33.33%	7.35%	48.48%
n_20000_p1_std_small	44.00%	10.80%	46.29%
n_20000_p2_std_small	30.48%	10.80%	45.81%
n_20000_p3_std_small	37.33%	11.49%	49.81%
n_5000_p1_std_mixed	49.48%	20.81%	47.43%
n_5000_p2_std_mixed	41.14%	36.60%	52.38%
n_10000_p1_std_mixed	55.67%	32.18%	48.76%
n_10000_p2_std_mixed	49.00%	39.24%	48.43%
n_20000_p1_std_mixed	57.33%	40.50%	47.29%
n_20000_p2_std_mixed	46.67%	45.70%	48.90%
n_20000_p3_std_mixed	45.67%	49.93%	52.86%
n_5000_p1_std_large	70.67%	42.63%	58.86%
n_5000_p2_std_large	56.62%	54.66%	59.90%
n_10000_p1_std_large	72.33%	47.42%	60.33%
n_10000_p2_std_large	58.29%	51.44%	61.00%
n_20000_p1_std_large	74.00%	51.35%	58.67%
n_20000_p2_std_large	54.29%	52.90%	61.00%
n_20000_p3_std_large	64.00%	50.75%	60.05%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 28

Three Categories: Percentage of Flagged Coefficients/Intercepts by Model and Outcome

	Interaction	Categorical	MAIDHA
Accuracy	31.75%	24.16%	44.22%
Bias	87.30%	43.55%	89.91%
Coverage	24.60%	43.90%	7.71%
Power	16.19%	3.97%	28.57%
Type1Error	93.20%	47.45%	91.11%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 29

Three Categories: AIC and BIC

	Interaction		**Categorical**		**MAIDHA**	
	AIC	*BIC*	*AIC*	*BIC*	*AIC*	*BIC*
n_5000_p1_std_small	14897	15053	119013	119361	119130	119232
n_5000_p2_std_small	14580	14737	81532	81880	81571	81673
n_10000_p1_std_small	29774	29947	57661	58009	57674	57776
n_10000_p2_std_small	29154	29327	121757	122105	121878	121980
n_20000_p1_std_small	59492	59681	42751	43068	42788	42882
n_20000_p2_std_small	58269	58459	28870	29187	28882	28976
n_20000_p3_std_small	57873	58063	61364	61681	61484	61577
n_5000_p1_std_mixed	21951	22107	84319	84667	84361	84464
n_5000_p2_std_mixed	21697	21853	57638	57986	57651	57753
n_10000_p1_std_mixed	43264	43437	42646	42963	42690	42784
n_10000_p2_std_mixed	42865	43038	57684	58032	57696	57799
n_20000_p1_std_mixed	85693	85883	14453	14740	14467	14551
n_20000_p2_std_mixed	84774	84963	30942	31229	31060	31145
n_20000_p3_std_mixed	81741	81931	21586	21873	21633	21718
n_5000_p1_std_large	31361	31518	120999	121347	121116	121218
n_5000_p2_std_large	31131	31288	84618	84966	84654	84757
n_10000_p1_std_large	61953	62126	28893	29210	28905	28999
n_10000_p2_std_large	61744	61917	60982	61299	61098	61191
n_20000_p1_std_large	123039	123229	14476	14763	14489	14573
n_20000_p2_std_large	122536	122725	30917	31203	31032	31117
n_20000_p3_std_large	119354	119544	21715	22002	21754	21839

Table 30

Average Percentage of Flagging Across Two and Three-category Scenarios

	Interaction		Categorical		MAIDHA	
	2 cat.	3 cat	2 cat	3 cat	2 cat	3 cat
n_5000_p1_std_small	20.00%	44.00%	12.50%	10.11%	19.37%	48.67%
n_5000_p2_std_small	25.00%	34.48%	5.00%	13.56%	19.37%	46.00%
n_10000_p1_std_small	25.00%	44.00%	15.00%	9.42%	19.37%	47.48%
n_10000_p2_std_small	25.00%	33.33%	7.50%	7.35%	19.37%	48.48%
n_20000_p1_std_small	30.00%	44.00%	10.00%	10.80%	19.37%	46.29%
n_20000_p2_std_small	25.00%	30.48%	12.50%	10.80%	19.37%	45.81%
n_20000_p3_std_small	20.00%	37.33%	5.00%	11.49%	19.37%	49.81%
n_5000_p1_std_mixed	50.00%	49.48%	60.00%	20.81%	20.79%	47.43%
n_5000_p2_std_mixed	50.00%	41.14%	56.92%	36.60%	23.81%	52.38%
n_10000_p1_std_mixed	50.00%	55.67%	60.00%	32.18%	20.79%	48.76%
n_10000_p2_std_mixed	50.00%	49.00%	60.00%	39.24%	23.81%	48.43%
n_20000_p1_std_mixed	50.00%	57.33%	60.00%	40.50%	20.79%	47.29%
n_20000_p2_std_mixed	50.00%	46.67%	60.00%	45.70%	25.24%	48.90%
n_20000_p3_std_mixed	50.00%	45.67%	56.92%	49.93%	27.46%	52.86%
n_5000_p1_std_large	41.67%	70.67%	51.35%	42.63%	26.51%	58.86%
n_5000_p2_std_large	43.33%	56.62%	42.69%	54.66%	29.37%	59.90%
n_10000_p1_std_large	46.67%	72.33%	58.46%	47.42%	27.94%	60.33%
n_10000_p2_std_large	38.33%	58.29%	51.35%	51.44%	27.94%	61.00%
n_20000_p1_std_large	46.67%	74.00%	61.54%	51.35%	25.08%	58.67%
n_20000_p2_std_large	43.33%	54.29%	60.00%	52.90%	29.37%	61.00%

Note. The six scenario combinations dropped from the three-category model are not included. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 31

Percentage of Flagged Instances Two versus Three Categories

	Interaction		Categorical		MAIDHA	
	2 cat	3 cat	2 cat	3 cat	2 cat	3 cat
Bias	50.00%	31.75%	56.13%	24.16%	78.57%	44.22%
Accuracy	11.11%	87.30%	13.68%	43.55%	23.81%	89.91%
Coverage	50.00%	24.60%	55.84%	43.90%	0.00%	7.71%
Power	0.00%	16.19%	0.00%	3.97%	2.96%	28.57%
Type1Error	85.19%	93.20%	79.17%	47.45%	21.90%	91.11%

Note. No shading represents excellent performance, light orange shading (20-39%) represents moderate performance, medium orange shading (40-59%) represents fair performance, dark orange shading (60-79%) represents poor performance, and bright red shading (80-100%) represents extremely poor performance.

Table 32

Chapter 5 Results

	Interaction		Categorical		MAIDHA	
Categories	2	3	2	3	2	3
Accuracy	Excellent performance for small and mixed deviation scenarios.	Excellent performance for small and mixed deviation scenarios	Excellent in mixed or small std dev One instance of poor performance (N = 5000, std= large, p = 3).	Excellent in mixed and small standard deviation scenarios. Poor performance for uneven or extremely uneven proportion representation in the large standard deviation scenarios.	Excellent in small and mixed standard deviation scenarios. Fair performance on large standard deviation scenarios.	Excellent performance for small standard deviation scenarios. Poor or extremely poor for large standard deviation scenarios.
Bias	Extremely poor in mixed standard deviation scenarios. Excellent in small standard deviation scenarios.	Extremely poor in mixed standard deviation scenarios. Excellent performance in small standard deviation scenarios.	Excellent for small standard deviation scenarios. Poor or extremely poor on mixed and large standard deviation scenarios.	Excellent performance for small standard deviation scenarios Fair to poor performance on mixed and large standard deviation scenarios	Poor or extremely poor performance.	Poor or extremely poor performance.
Coverage	Always fair performance.	Excellent in mixed standard deviation scenarios; fair in large standard deviation scenarios.	Excellent for small standard deviation scenarios Poor or extremely poor for mixed and large std dev	Excellent performance for small standard deviation scenarios Fair to poor on mixed and large standard deviation scenarios	Excellent across all scenarios	Excellent performance across all scenarios.
Power	Always excellent performance	Always moderate performance.	Almost always excellent performance.	Almost always excellent performance	Almost always excellent performance.	Moderate performance in most scenarios.
Type 1 error	Always poor of extremely poor performance.	Always poor or extremely poor performance.	Poor or extremely poor performance on almost all mixed and large standard deviation scenarios.	Several instances of poor performance in mixed and large standard deviation scenarios. One instance of extremely poor performance.	Excellent for small standard deviation scenarios excellent for mixed standard deviation scenarios with uneven representation, excellent for large standard deviation scenarios with even or uneven representation.	Extremely poor performance.

Figure 9

Two Categories: Distribution of Accuracy Flags

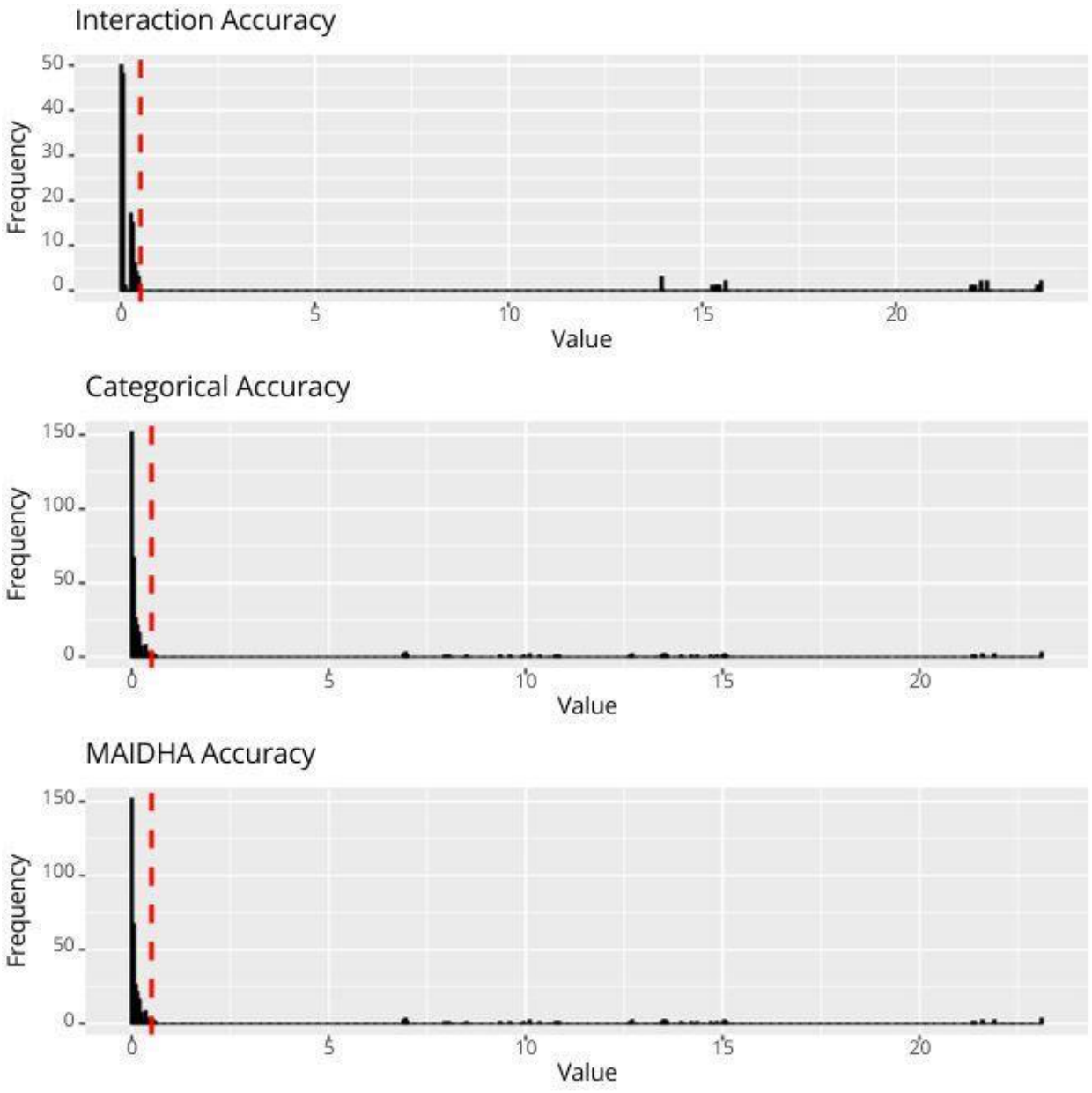


Figure 10

Two Categories: Distribution of Moderate Bias Flags

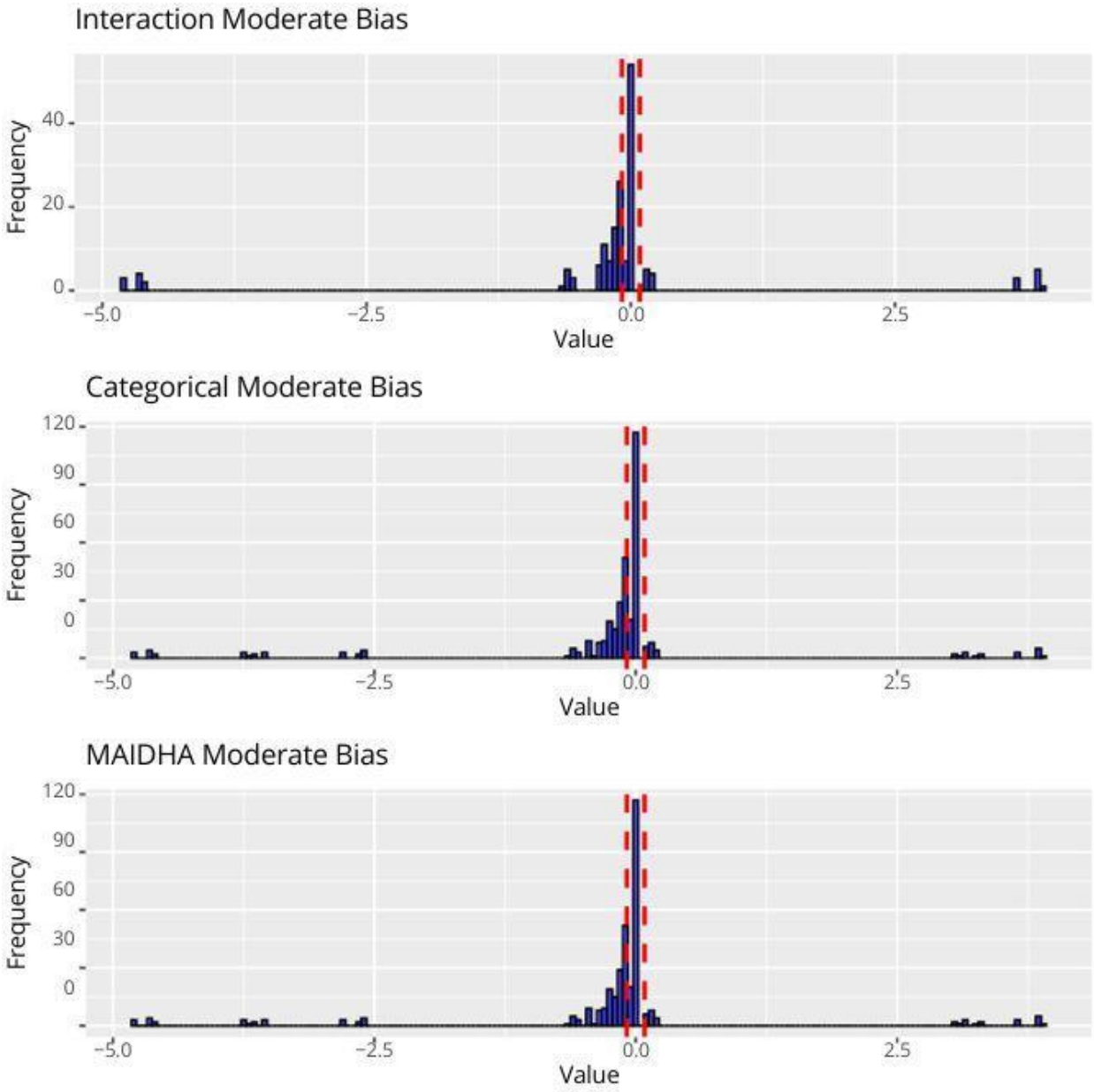


Figure 11

Two Categories: Distribution of Extreme Bias Flags

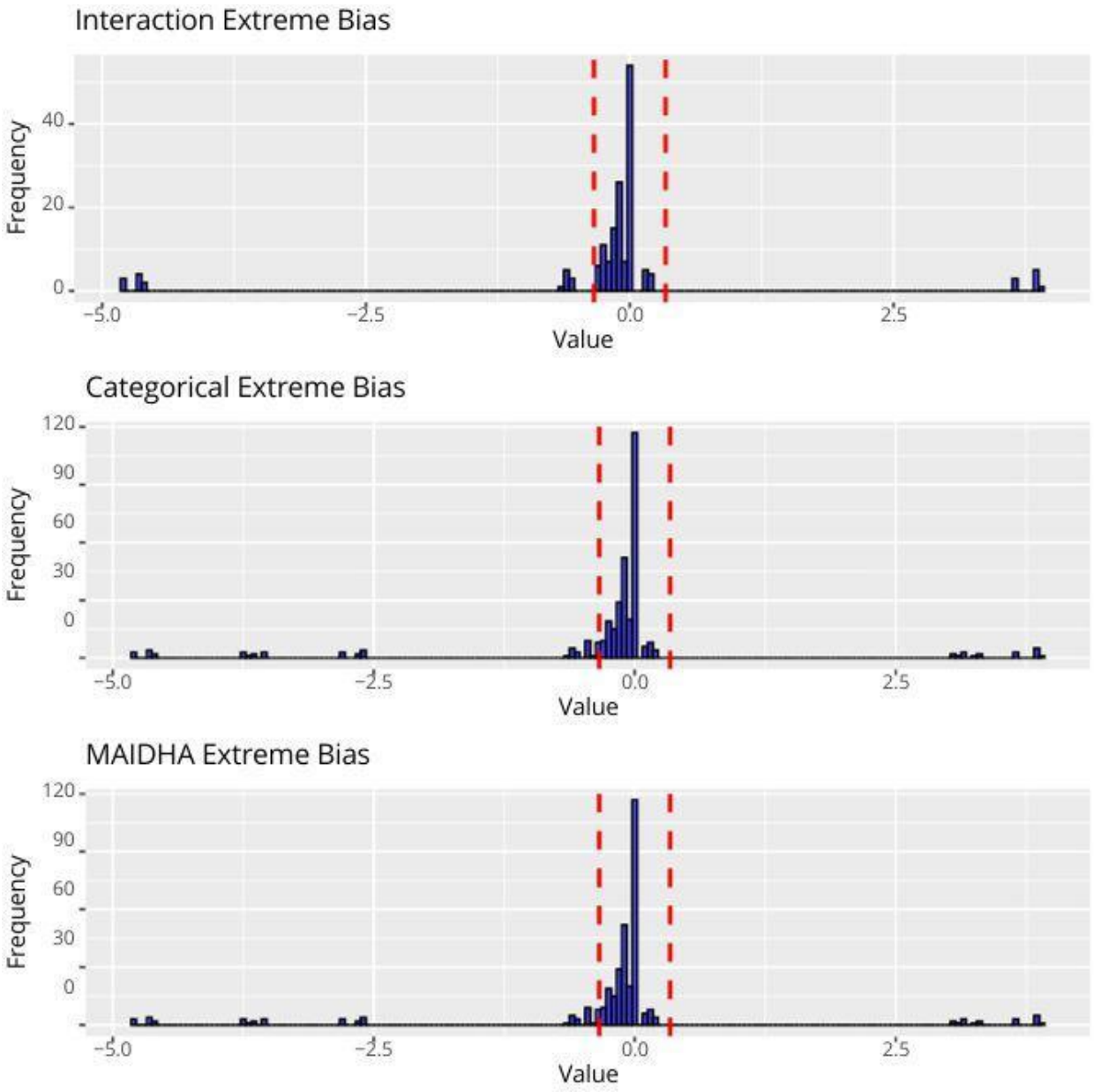


Figure 12

Two Categories: Distribution of Coverage Flags

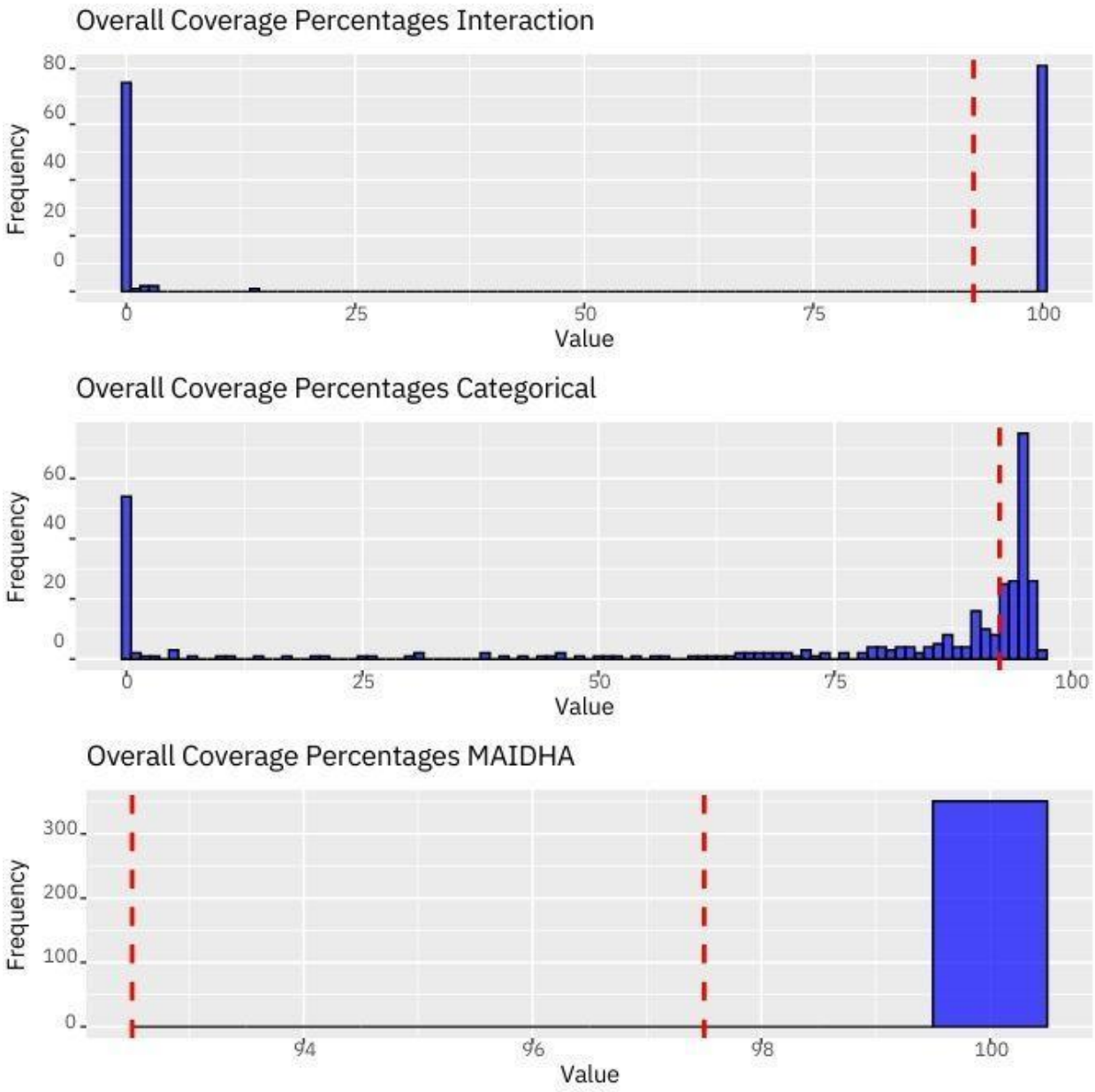


Figure 13

Two Categories: Distribution of Power Flags

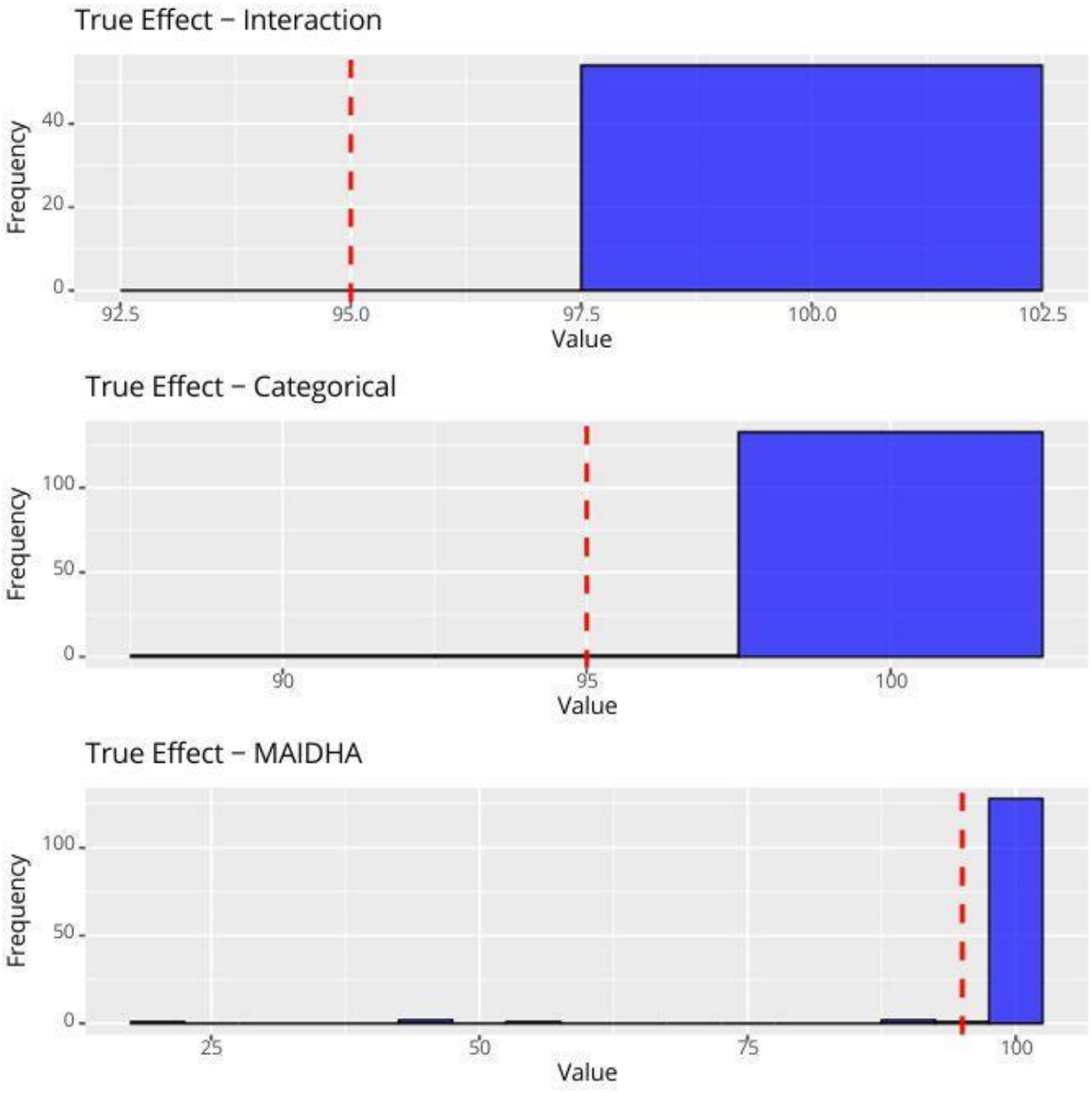


Figure 14

Two Categories: Distribution of Type 1 Error Flags

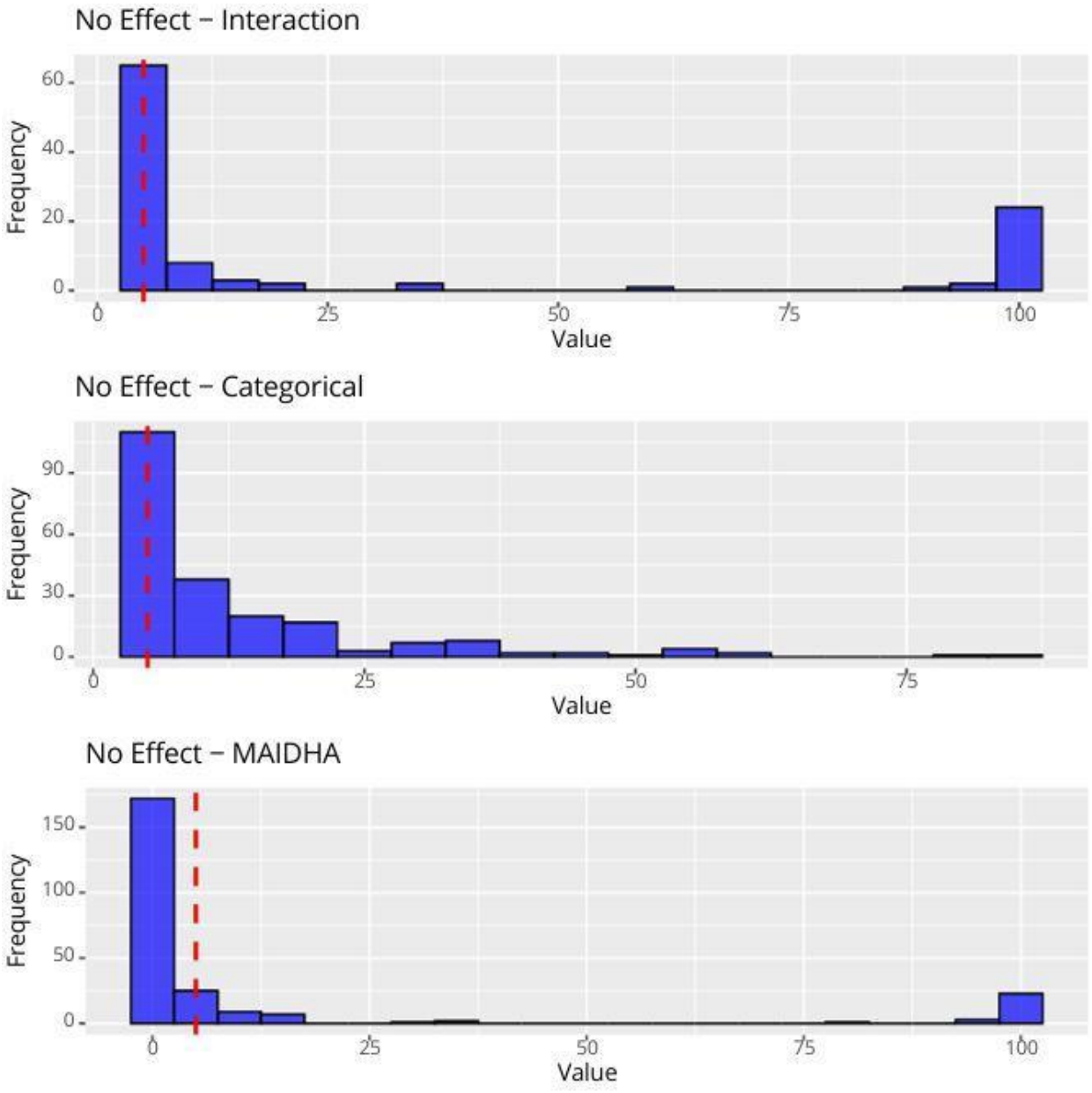


Figure 15

Percentage of Small Standard Deviation Flagged Instances

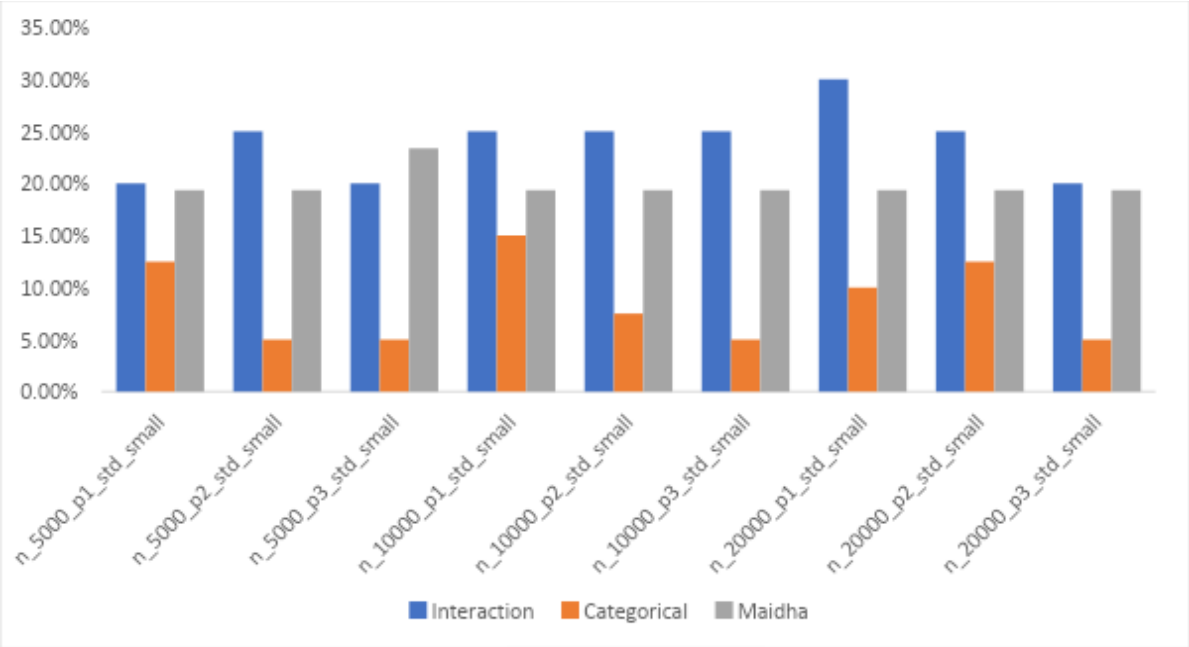


Figure 16

Distribution of Outcomes Across Models

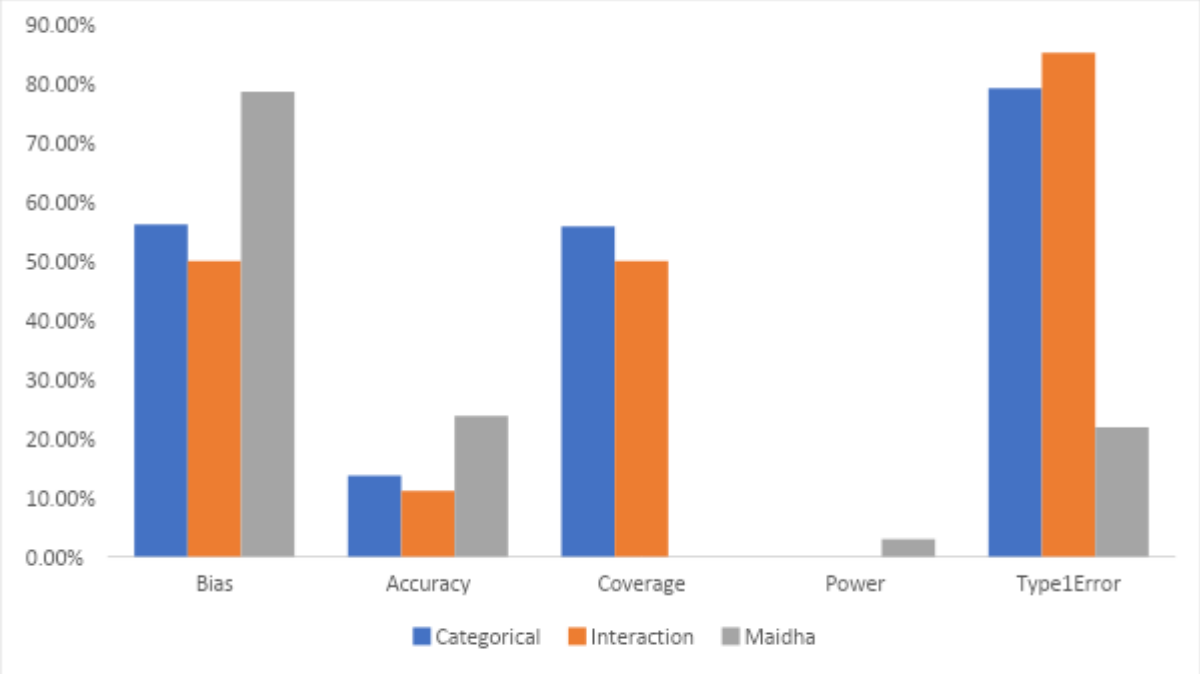


Figure 17

Three Categories: Distribution of Accuracy Flags

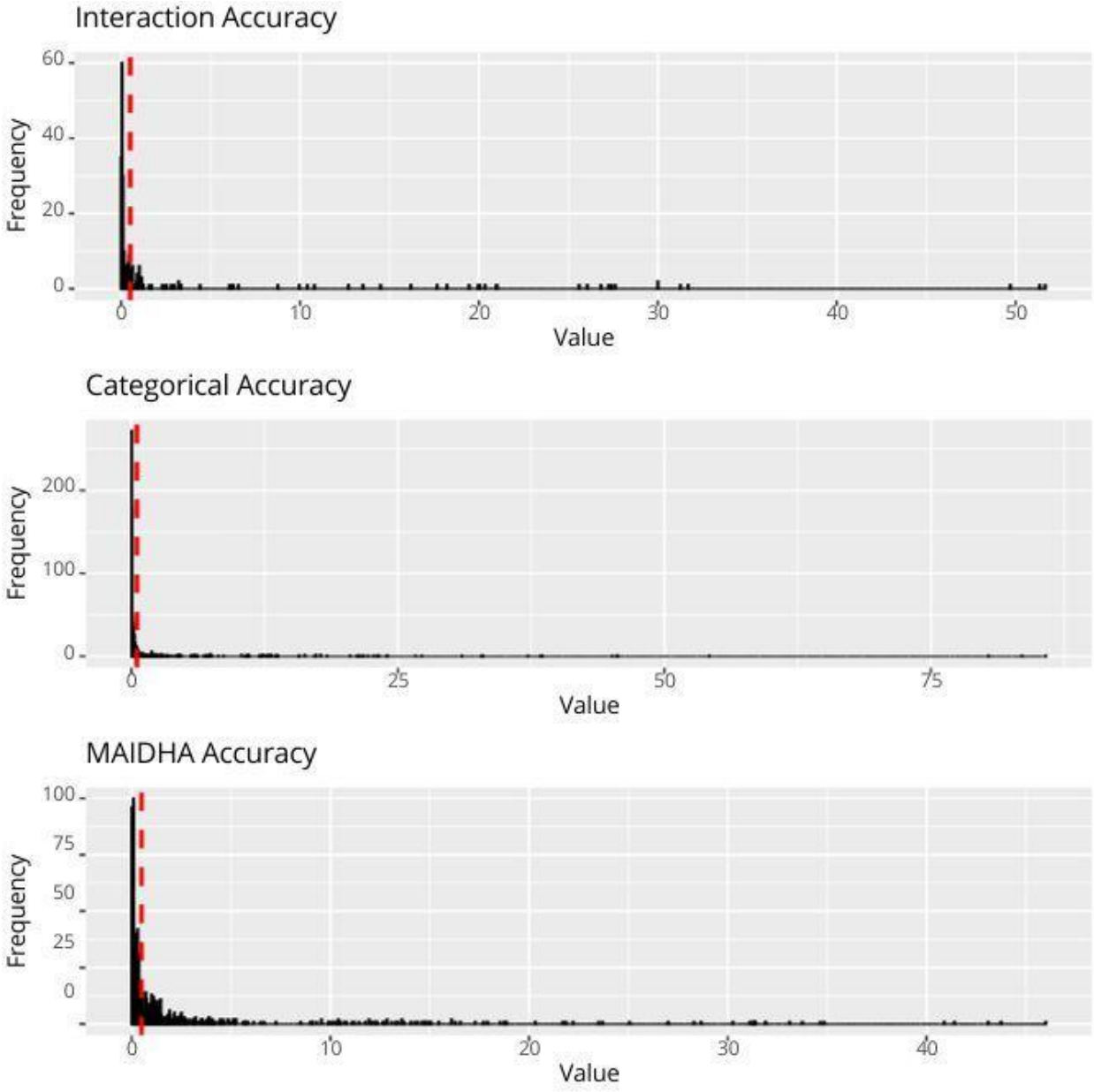


Figure 18

Three Categories: Distribution of Moderate Bias Flags

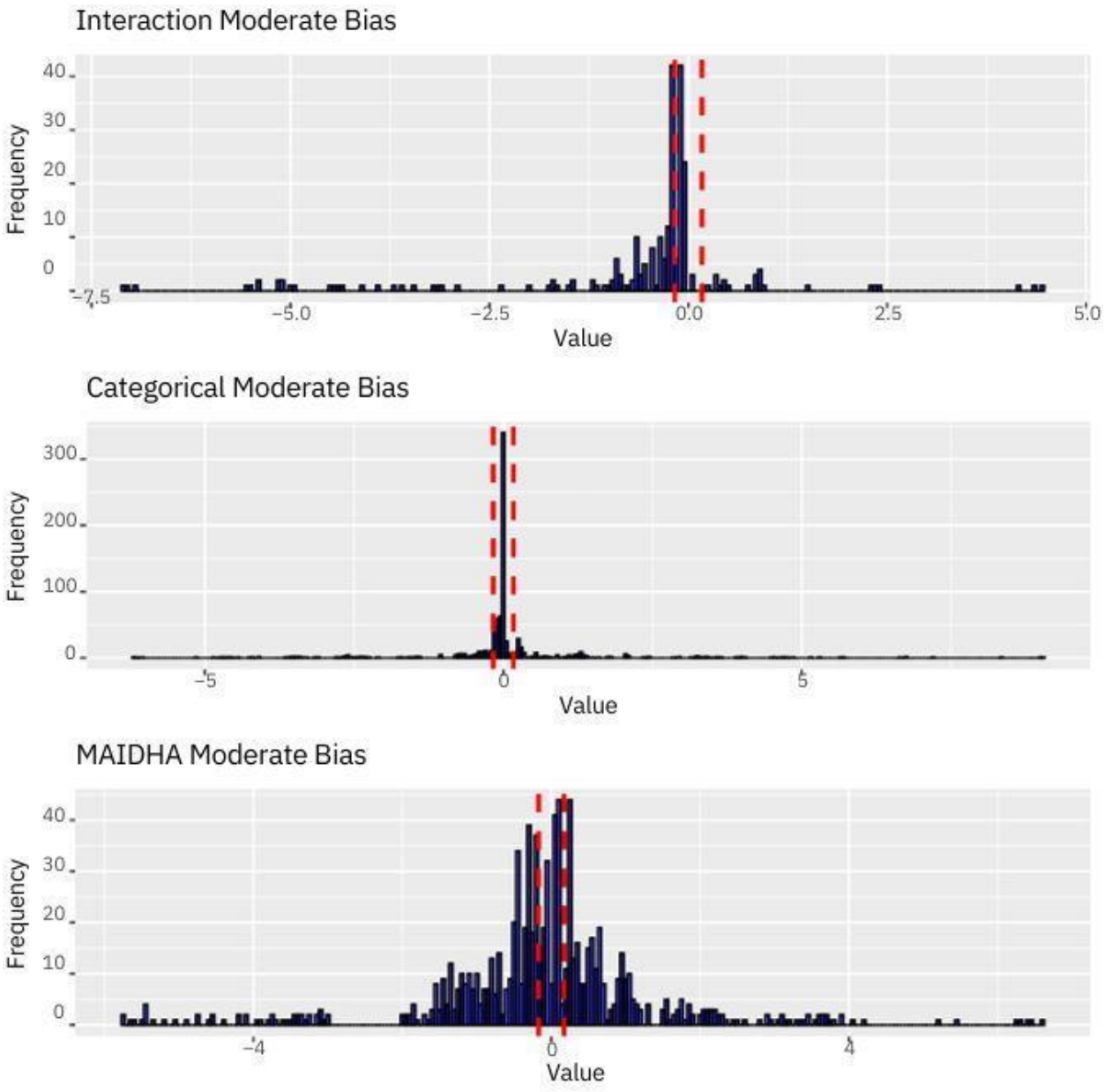


Figure 19

Three Categories: Distribution of Extreme Bias Flags

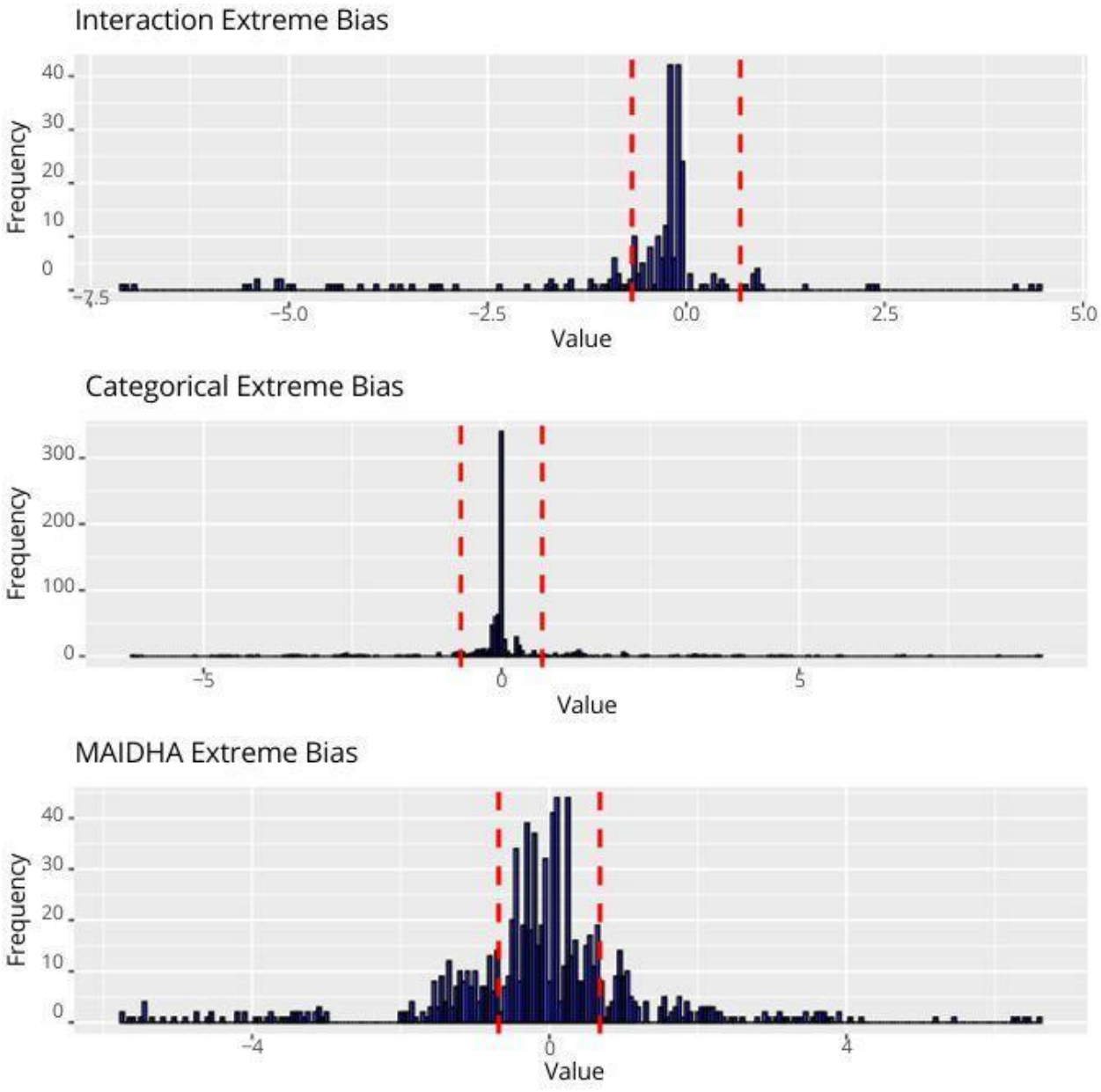


Figure 20

Three Categories: Distribution of Coverage Percent

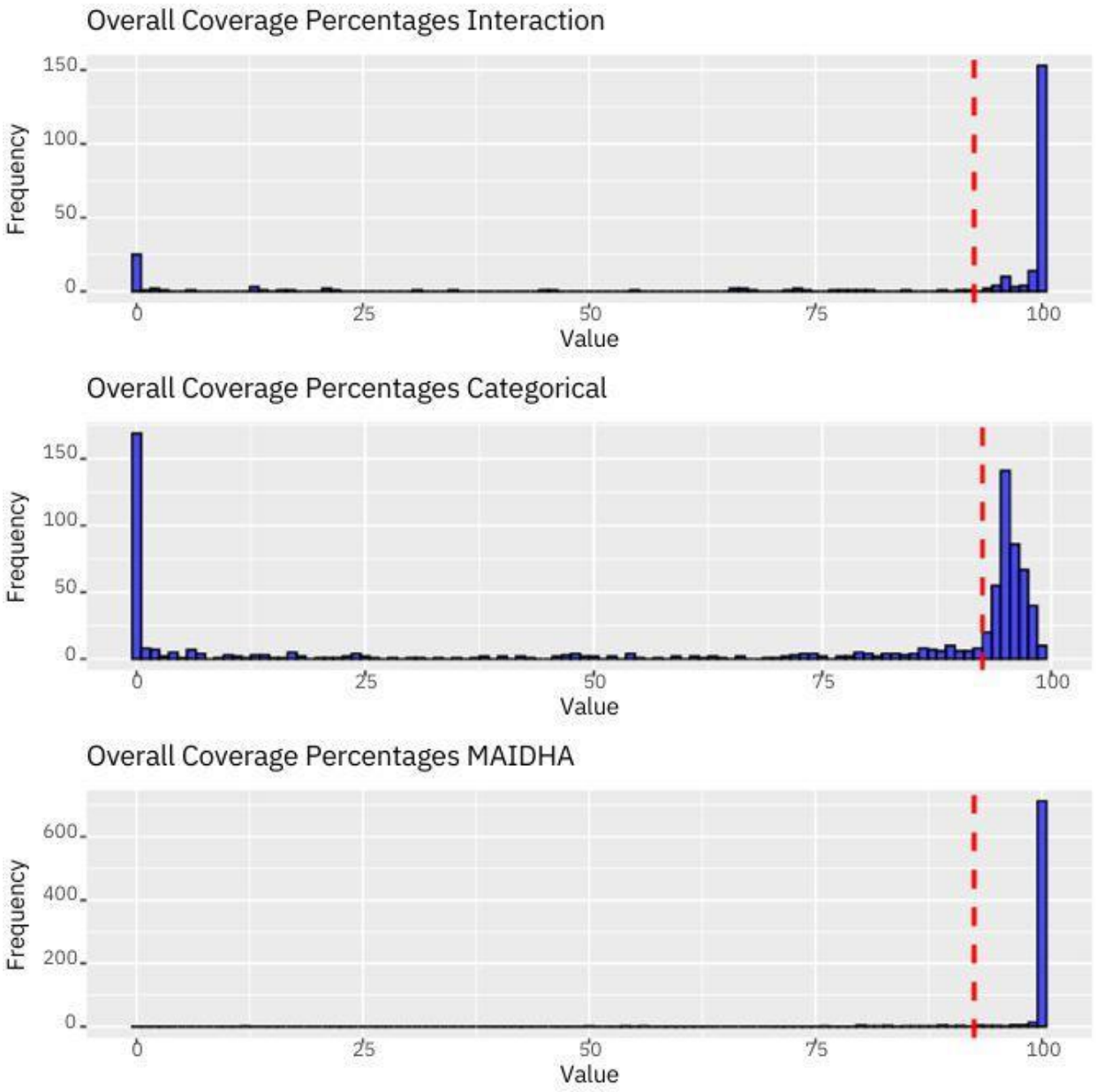


Figure 21

Three Categories: Distribution of Power Flags

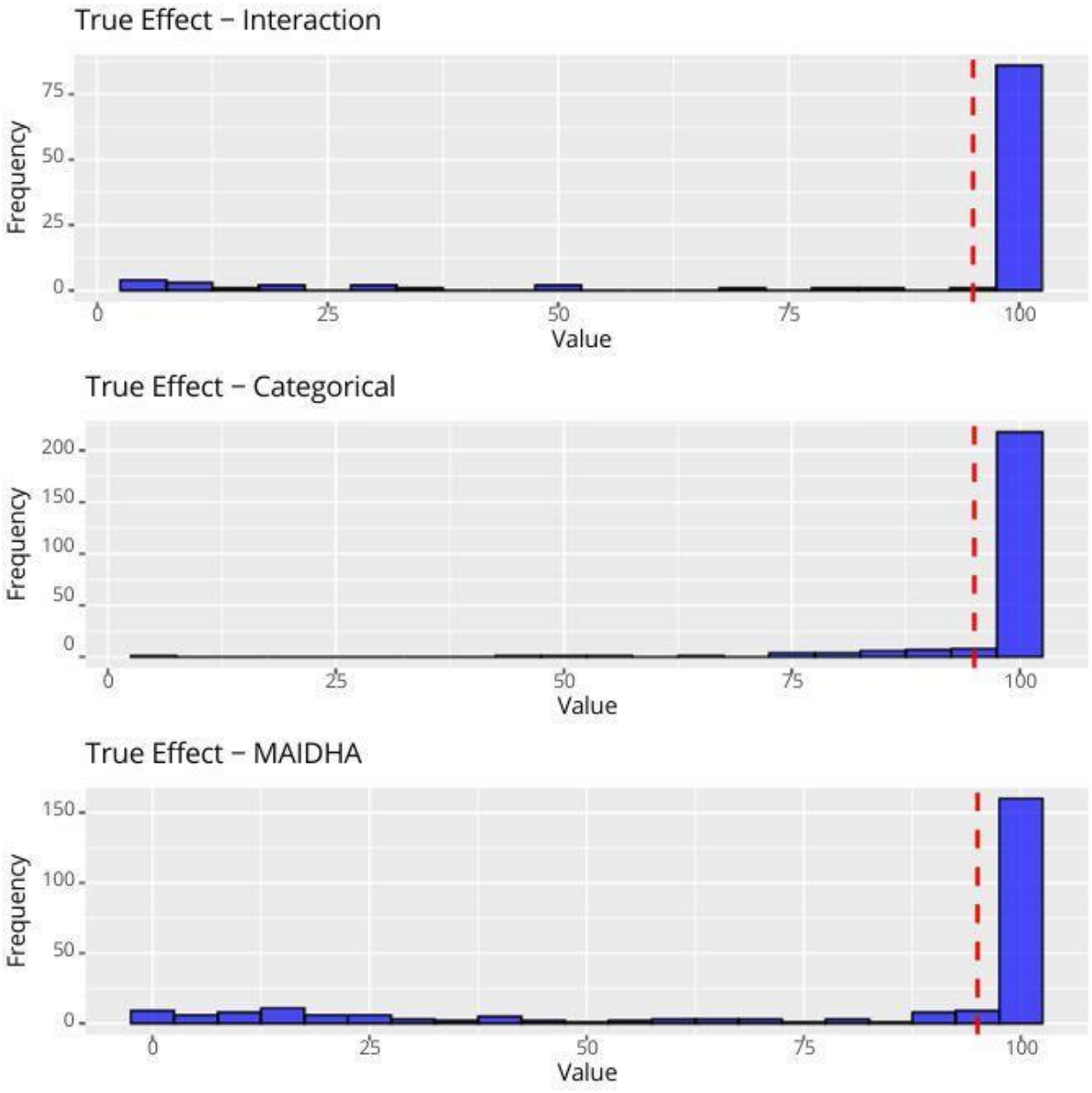


Figure 22

Three Categories: Distribution of Type 1 Error Flags

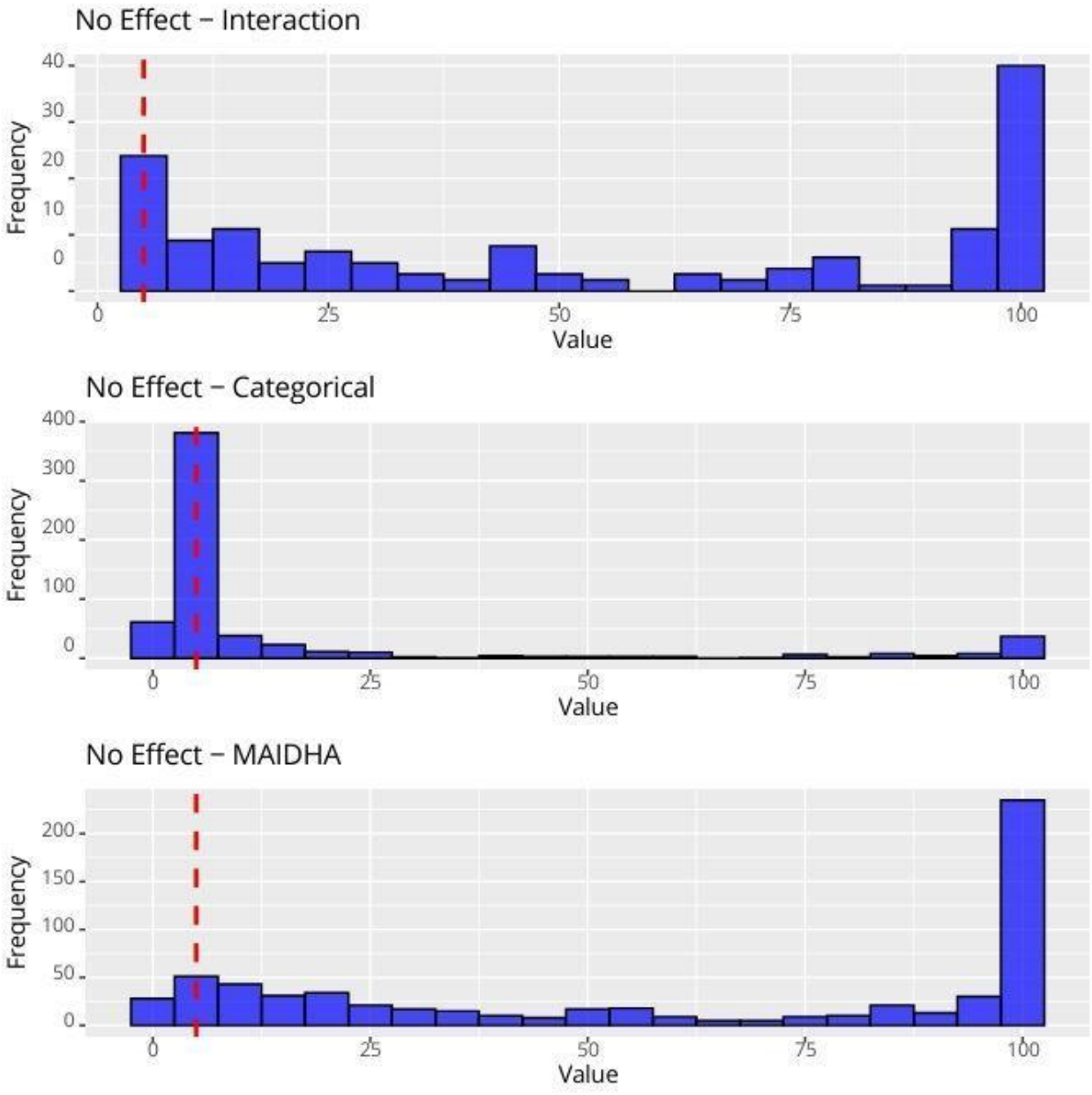


Figure 23

Three Categories: Percentage of Small Standard Deviation Flagged Instances

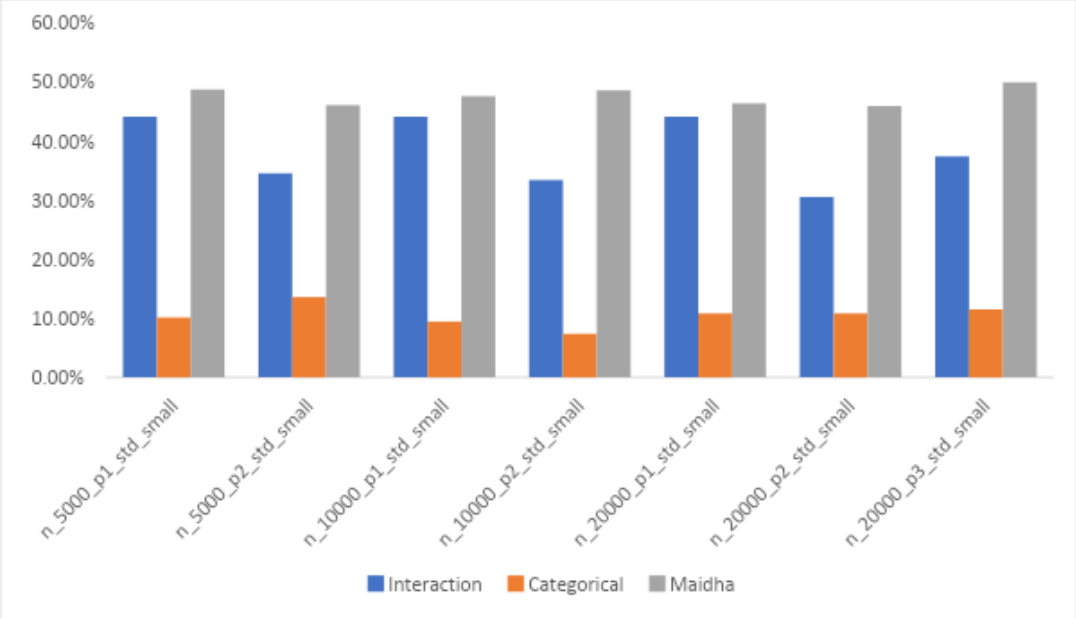


Figure 24

Three Categories: Distribution of Outcomes Across Models