# Advancing Validity Evidence and Methodology for Forced-Choice Assessment

Jake Plantz

Department of Psychology, Faculty of Science

McGill University, Montréal

October 2024

A thesis submitted to McGill University in

partial fulfillment of the requirements of the degree

of the Doctor of Philosophy

## Table of Contents

**Chapter 3. From the 1940s to 2020s: A Review of The Current State of Forced-Choice Methodology**

**Chapter 4. Detecting DIF in Forced-Choice Assessments: A Simulation Study Examining the Effect of Model Misspecifications**

# List of Tables

Chapter 1 - Introduction and Background

Chapter 2 - Validity Evidence Reported for Non-Cognitive Educational Assessments: Current Practice and Next Step

Chapter 3 - From the 1940s to 2020s: A Review of The Current State of Forced-Choice Methodology

Chapter 4 - Detecting DIF in Forced-Choice Assessments: A Simulation Study Examining the Effect of Model Misspecifications

Chapter 5 – General Discussion

## List of Figures

Chapter 1 - Introduction and Background

Chapter 2 - Validity Evidence Reported for Non-Cognitive Educational Assessments: Current Practice and Next Step

Chapter 3 - From the 1940s to 2020s: A Review of The Current State of Forced-Choice Methodology

Chapter 4 - Detecting DIF in Forced-Choice Assessments: A Simulation Study Examining the Effect of Model Misspecifications

Chapter 5 – General Discussion

## Abstract

This dissertation's objective is to advance the validity and methodology of forced-choice (FC) non-cognitive assessments. To this end I have written three manuscripts presented in full in Chapters 2-4. In Chapter 1, I begin with an introduction to the FC test format and its methodology. I then provide background on non-cognitive assessments before summarizing the remaining chapters.

In Chapter 2, I review non-cognitive assessments and their validity arguments. Non-cognitive assessments measure skills such as teamwork, intellectual engagement and others that are outside of cognitive abilities like math and reading. These assessments are increasingly used for high-stakes decisions in workforce and educational settings and need strong validity arguments. To understand the state of validity evidence reported for these assessments, we undertook a mapping review of 46 validity studies. Our findings indicated that validity evidence related to consequences of testing, response processes, and measurement bias was not often reported. Further, demographic data for validation samples were also limited. This raised questions about the assessments' use with diverse populations. Through this review, we identified validity issues with the non-cognitive assessments examined, particularly the role of response bias and faking in high-stakes contexts. This led me to consider the FC test format as a means of reducing the potential for faking.

In Chapter 3, we conduct an in-depth literature review of more than eight decades of research related to FC assessments from 1940 to 2024. This exhaustive survey established a

coherent picture of the existing methodology for FC assessments. Our review and synthesis of the various areas of FC assessment is the first step toward creating a more standardized test construction procedure in the future. It also revealed a gap in FC methodology: there are no thoroughly developed methods for testing differential item functioning (DIF) in FC models. Testing for DIF is a necessary part of ensuring the test is fair for all respondents.

In Chapter 4, I examine a method for testing DIF in FC assessments. This work builds on Lee and colleagues' (2021) latent-scoring-based approach. They evaluated this approach in conditions where the model is correctly specified, but this is not typical in practice. To further investigate these methods, we conducted a simulation study examining the effects of model misspecification on DIF detection. Our study indicated that when misspecification is present the methods cannot accurately detect blocks with or without DIF.

In Chapter 5, I summarize and discuss the limitations and implications of Chapters 2-4. I then consider the impact of my entire body of work in this dissertation. I conclude with its limitations and potential future directions for my research in FC assessment.

# Résumé

L'objectif de cette thèse est de faire progresser la validité et la méthodologie des évaluations non cognitives à choix forcé. À cette fin, j'ai rédigé trois manuscrits présentés dans leur intégralité dans les chapitres 2 à 4. Dans le chapitre 1, je commence par présenter le format du test (CF) et sa méthodologie. Je présente ensuite le contexte des évaluations non cognitives avant de résumer les autres chapitres.

Au chapitre 2, je passe en revue les évaluations non cognitives et leurs arguments de validité. Les évaluations non cognitives mesurent des compétences telles que le travail d'équipe, l'engagement intellectuel et d'autres qui ne relèvent pas des capacités cognitives comme les mathématiques et la lecture. Ces évaluations sont de plus en plus utilisées pour des décisions à enjeux forts dans le monde du travail et de l'éducation et nécessitent de solides arguments de validité. Pour comprendre l'état des preuves de validité rapportées pour ces évaluations, nous avons entrepris une analyse de 46 études de validité. Nos conclusions indiquent que les preuves de validité liées aux conséquences des tests, aux processus de réponse et aux biais de mesure ne sont pas souvent rapportées. En outre, les données démographiques relatives aux échantillons de validation étaient également limitées. Cela soulève des questions quant à l'utilisation des évaluations auprès de diverses populations. Cet examen nous a permis d'identifier les problèmes de validité des évaluations non cognitives examinées, en particulier le rôle du biais de réponse et de la falsification dans des contextes à enjeux élevés. Cela m'a amené à considérer le format du test de CF comme un moyen de réduire le potentiel de simulation.

Au chapitre 3, je procède à une analyse documentaire approfondie de plus de huit décennies de recherche liée aux évaluations de la CF, de 1940 à 2024. Cette étude exhaustive a permis de dresser un tableau cohérent de la méthodologie existante pour les évaluations de la CF. Mon examen et ma synthèse des différents domaines de l'évaluation de la CF constituent la première étape vers la création d'une procédure d'élaboration de tests plus standardisée à l'avenir. Elle a également révélé une lacune dans la méthodologie de la CF: il n'existe pas de méthodes bien développées pour tester le fonctionnement différentiel des items (FDI) dans les modèles de la CF. Les tests de DIF sont indispensables pour s'assurer que le test est équitable pour tous les répondants.

Au chapitre 4, j'examine une méthode permettant de tester le FDI dans les évaluations de la CF. Ce travail s'appuie sur l'approche de Lee et de ses collègues (2021) basée sur la notation latente. Ils ont évalué cette approche dans des conditions où le modèle est correctement spécifié, ce qui n'est pas le cas dans la pratique. Pour approfondir l'étude de ces méthodes, nous menons une étude de simulation examinant les effets d'une mauvaise spécification du modèle sur la détection de DIF. Notre étude indique qu'en présence d'une mauvaise spécification, les méthodes ne peuvent pas détecter avec précision les blocs avec ou sans FDI.

Au chapitre 5, je résume et examine les limites et les implications des chapitres 2 à 4. Je considère ensuite l'impact de l'ensemble de mon travail dans cette thèse. Je conclus sur les limites et les orientations futures potentielles de mes recherches dans le domaine de l'évaluation de la CF.

# ACKNOWLEDGEMENTS

This dissertation is the result of three years of work that was only possible with six years of learning, growth, and the support of a handful of people who came during and before. I am grateful to take this opportunity to express my immeasurable appreciation for them.

I am deeply indebted to my supervisor, Dr. Jessica Flake. Jess, I would never have come this far if not for you. You allowed me the freedom to pursue what I found interesting. You provided the support I needed to improve every paper. When the PhD started, you met me where I was and treated me as an equal. You know I can be quite wordy in my writing and could go on forever, so I will end with this: Your humanity, resilience, and ferocity in the face of adversity gave me something to strive to be. Throughout it all, you taught me not only how to be a better scholar, but a better person.

I am incredibly grateful for Dr. Keith Wright. While Jess allowed me to explore academia, you provided me with an outlet to explore opportunities outside of it. Your passion for the work and incredible support at every turn helped propel me forward. You are a phenomenal mentor, co-author, and person. You have energized me for the future, and I appreciate that dearly.

Among my many mentors, there was the first who fostered my curiosity in science. Dr. Tess Neal was the first to show me how a question can be investigated and the joy of finding evidence of an answer. Tess, I appreciate you greatly. You were the catalyst that led to my love of psychometrics. I would never have pursued this PhD if not for your encouragement, support, and mentorship throughout my undergraduate and master’s degrees.

I also thank the members of my dissertation and supervisory committee who gave their time and feedback to support me throughout the degree. Thank you to Drs. Melanie Dirks, Carl Falk, and Bärbel Knäuper. I also would like to thank Dr. Anna Brown. Anna, your support and feedback are invaluable. Your papers are why I found myself interested in forced-choice and I appreciate you as a mentor greatly.

I would also like to thank my therapist, Skylar Ofili. You were there when I got rejected from every PhD program, accepted into this one, and through all the life in between. You gave me the skills I needed to forge a path toward a healthy life. You taught me that life is one big mountain to climb. Joy is found along the way with people who sometimes need to walk their own path. I am fortunate to have had you along my journey thus far.

So many others have been part of that climb up the mountain. Emma, Lindsay, Jacey, Mairead, Elizabeth, Tess, Logan, my parents, and so many more. So many have touched my life and inevitably played a role in me being where I am—completing this work. I cannot express my gratitude enough to every single person who has helped me get here in one way or another. You were all vital.

## Contribution To Original Knowledge

Chapter 2, which has been invited for revision at the *Journal of Applied Educational Measurement*, expands on prior reviews conducted on the types of validity evidence offered by non-cognitive assessments. Cox and colleagues (2019) examined assessments measuring collaboration, perseverance, and self-regulated learning, whereas Cordier and colleagues (2015) focused on search terms relating to social skills. Halle and Darling-Churchill's (2016) review also focused on these types of constructs but sought to identify the best assessments based on their validity, reliability, and the breadth of skills measured. In my review I took a broader perspective by focusing on the umbrella term “non-cognitive.” I also examined and compared the validity arguments presented by assessments from the grey and peer-reviewed literature. Finally, I provided practical recommendations for strengthening the validity arguments of non-cognitive assessments that were not present in the literature broadly.

Chapter 3, under review at *Psychological Methods,* exhaustively reviews the FC test format since its inception in psychological research in 1940. The primary contribution of this chapter is a comprehensive review of FC methods at each phase of their development. The chapter also contains an annotated bibliography of the more than eight decades of literature I reviewed. This represents the first step toward creating a structured construction procedure by laying out all of the various phases of construction and the state of the test format.

Chapter 4, accepted in principle as a Stage 1 Registered Report (indicating the background, methods, hypotheses, and proposed analyses were peer-reviewed and accepted) at *Peer Community In*, focuses on examining an approach for testing DIF in FC assessments.

Through an extensive simulation study where conditions representative of real-world testing scenarios were examined, the chapter found evidence that the only proposed method for DIF testing in dominance models, is not accurate when misspecification is present. This chapter also provides important replication work of some conditions tested by Lee and colleagues (2021). Finally, a different formulation of the model used by Lee et al. was implemented which allows for uniform and nonuniform DIF to be tested separately.

My dissertation comprehensively examined the challenges of constructing FC non-cognitive assessments and non-cognitive assessments more broadly. Chapter 2 uncovered validity issues present in non-cognitive assessments from across the field and made recommendations on how to improve them. Chapter 3 comprehensively reviewed the methods at all phases of constructing an FC assessment. Chapter 4 advanced the state of FC assessment methodology by examining a method for detecting DIF. The primary contribution of this dissertation as a whole is the synthesis and advancement of the methodology available for FC assessments.

## Contribution Of Authors

This dissertation follows a manuscript-based format. These manuscripts are as follows:

- Chapter 2.
    - Publication [1]: **Plantz, J.W.**, Elbaz, S., Brenneman, M., & Flake, J.K. (invited for revision, *Journal of Applied Educational Measurement*). Validity evidence reported for non-cognitive educational assessments: current practice and next steps.
- Chapter 3.
    - Publication [2]: **Plantz, J.W.,** Flake, J.K., Wright, K. (under review, *Psychological Methods*). From the 1940s to 2020s: A review of the current state of forced-choice methodology.
- Chapter 4.
    - Publication [3]: **Plantz, J.W.,** Brown, A., Wright, K., Flake, J.K. (accepted in principle, *Peer Community In Registered Reports*). Detecting DIF in forced-choice assessments: a simulation study examining the effect of model misspecifications.

All three manuscripts are presented in full in Chapters 2-4.

I am the first and corresponding author for each manuscript. Dr. Jessica Flake was involved in the supervision and editing of each manuscript. She was also responsible for the initial idea in [1], contributed to the development of the ideas within [2], and helped in the development of the simulation study in [3]. Dr. Anna Brown also assisted with the simulation study planning in [3]. All other listed authors contributed to the manuscripts through editing and providing feedback. Outside of these contributions, I was responsible for the entire body of work for this dissertation. This includes the analyses, programming, simulation studies, and writing of the manuscripts.

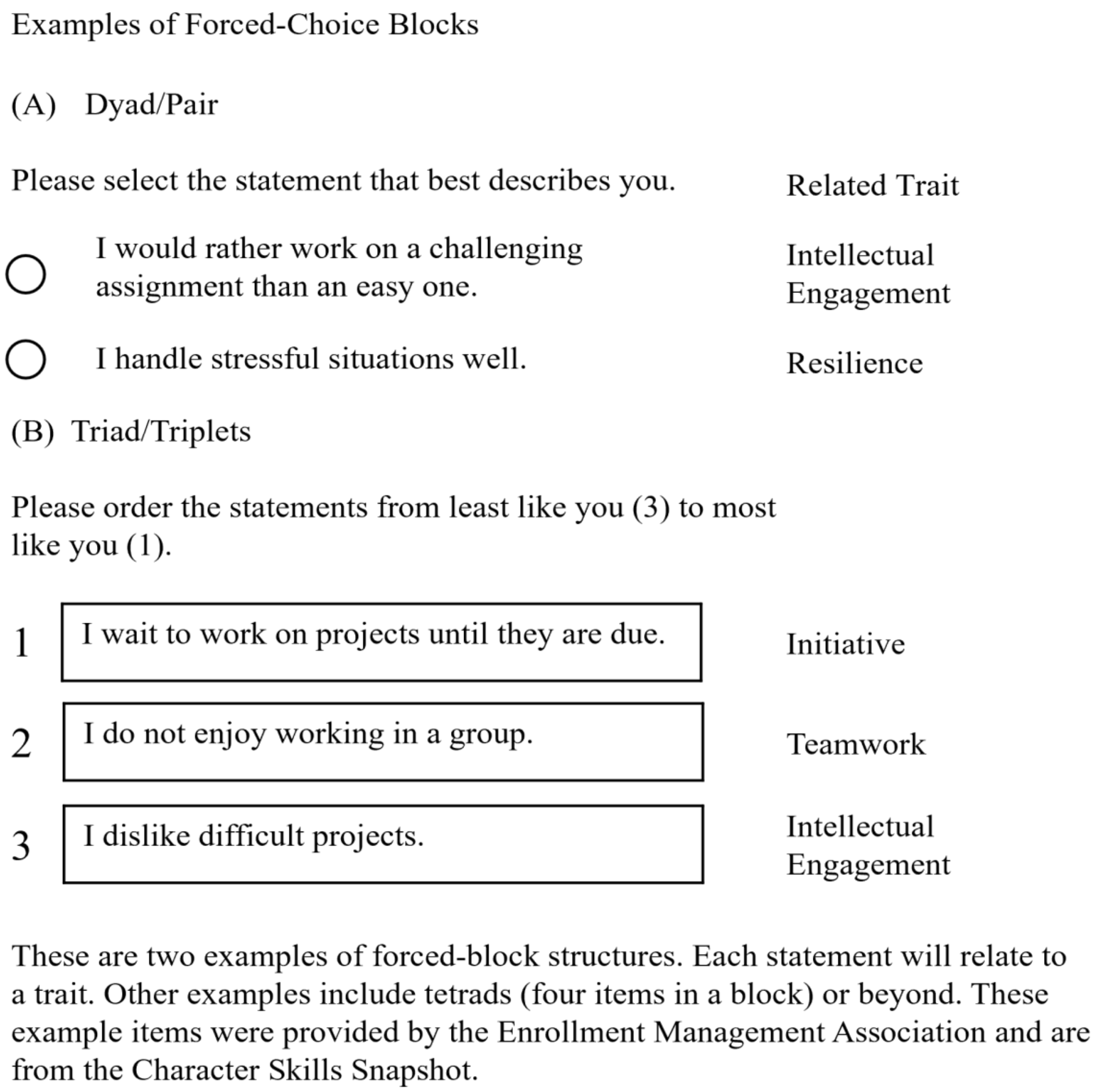
# Chapter 1. Introduction and Background

# Chapter 1. Introduction and Background

## Forced-Choice Assessment

Imagine you apply for a supervisor job that you are very excited about. When you apply, the company requests that you respond to a non-cognitive test. These types of tests are increasingly used in high-stakes decisions like personnel selection. This test may measure abilities like teamwork, resilience, and leadership, or anything else outside of cognitive abilities such as math and reading, where questions have clear right and wrong answers. As this is a job you really want, you will naturally try to perform the best you can. In a rating scale test, you may be asked to answer items like, "I have trouble organizing a team," on a scale of 1 (completely disagree) to 5 (completely agree). As you want to present your best possible self, you may answer "1" even if your true response may be different. This is known as faking, and in scenarios where stakes are high, it may be more likely to occur (Christiansen et al., 2005). As non-cognitive assessments become more widely used in high-stakes scenarios, test developers look to formats outside of rating scales that reduce the opportunity for respondents to fake their responses. The forced-choice (FC) assessment format is perhaps the most promising of these, as it can reduce faking among other social desirability biases (Brown & Maydeu-Olivares, 2011). Rather than being shown a single item, respondents are asked to choose or rank items from a block of items that are most or least like them. Each item in the block relates to a single trait, such as in Fig. 1, where the items relate to teamwork, resilience, and openness. The reasoning behind FC assessments is intuitive: By juxtaposing multiple items together and asking respondents to rank or select the item most or least like them, the chance of respondents identifying and selecting the most socially desirable response is reduced (Bartlett et al., 1960; Brogden, 1954; Brown & Maydeu-Olivares, 2011; Cao & Drasgow, 2019; Christiansen et al., 2005; Edwards, 1957; Goffin et al., 2011; Jackson et al., 2000; Joubert et al., 2015; Lee & Smith, 2020; Wetzel et al., 2020; Wetzel & Frick, 2020). Research shows this may result in stronger validity evidence compared to rating scale tests (Bartram, 2007; Lee et al., 2018; Wetzel & Frick, 2020; Vasilopoulos et al., 2006). These reasons are likely why FC is being increasingly used in high-stakes settings such as personnel selection and education. For example, one of the most widely used workplace assessments in the world, the OPQ32-r, is an FC assessment (Bartram & Brown, 2011). Meanwhile, in the field of education several of the most widely used assessments are FC formatted (see Enrollment Management Association, 2023; Walton et al., 2022). However, while FC has several benefits it comes with additional complexities due to the data it generates which must be considered.

Fig. 1 - Forced-Choice Example



Examples of Forced-Choice Blocks

(A) Dyad/Pair

| Please select the statement that best describes you. | Related Trait |
|---|---|
| ○ I would rather work on a challenging assignment than an easy one. | Intellectual Engagement |
| ○ I handle stressful situations well. | Resilience |

(B) Triad/Triplets

Please order the statements from least like you (3) to most like you (1).

| | | |
|---|---|---|
| 1 | I wait to work on projects until they are due. | Initiative |
| 2 | I do not enjoy working in a group. | Teamwork |
| 3 | I dislike difficult projects. | Intellectual Engagement |

These are two examples of forced-block structures. Each statement will relate to a trait. Other examples include tetrads (four items in a block) or beyond. These example items were provided by the Enrollment Management Association and are from the Character Skills Snapshot.

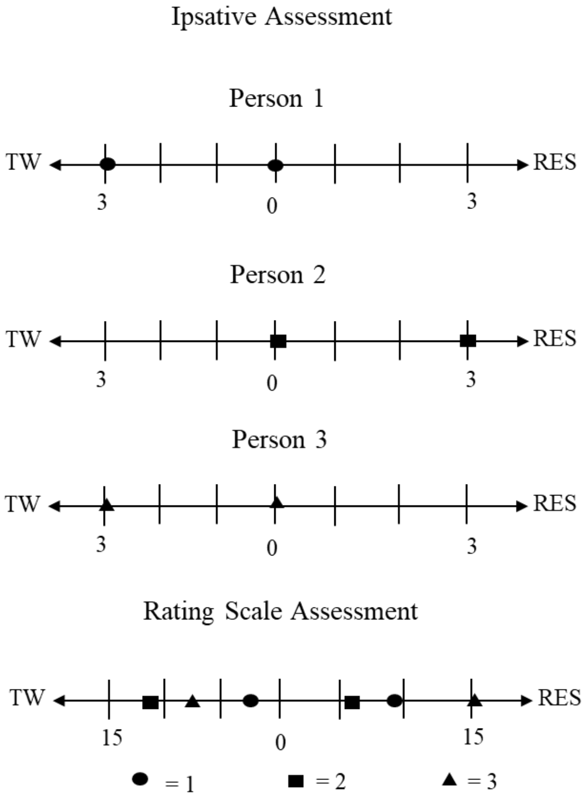
***Ipsative Data***

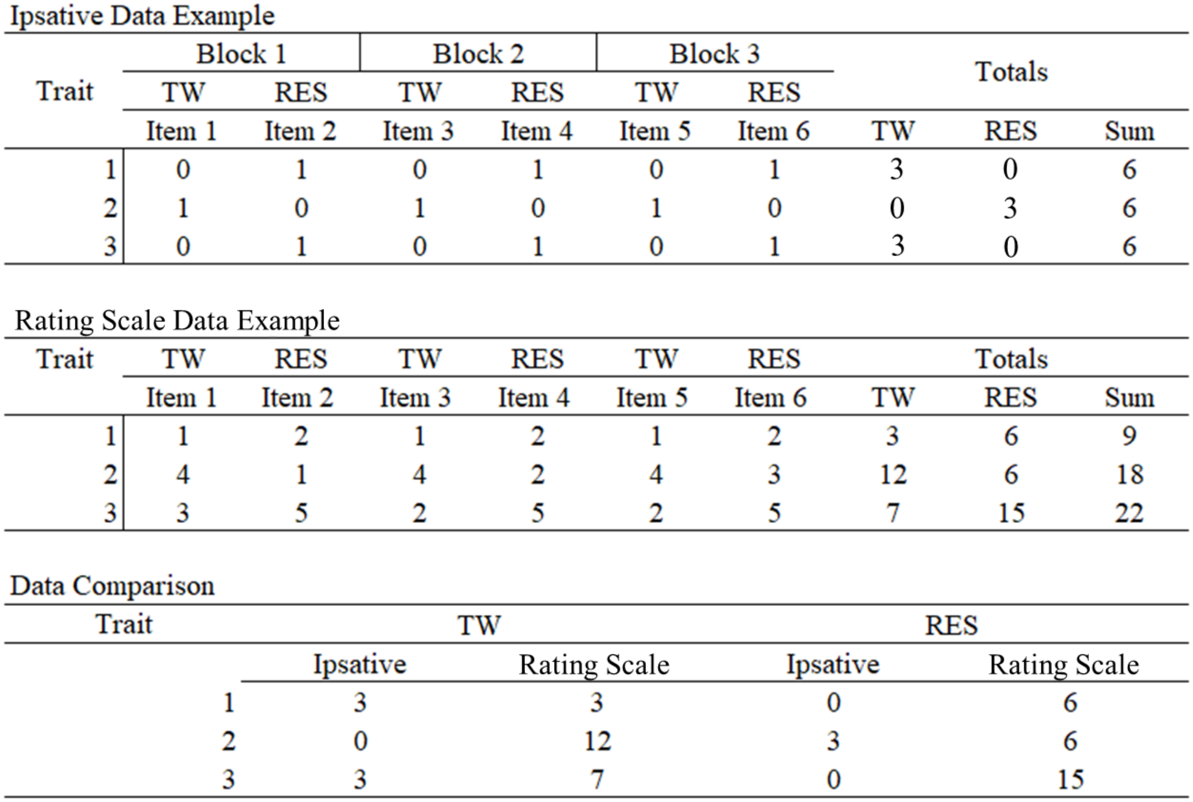
Ipsative data implies that the responses to each item are interdependent, and the sum of the variables equals a constant score for all respondents (Baron, 1996). This is because the scores for each trait are relative to each other within a given block, meaning that as one trait is ranked higher, another must be lower. As a result, ipsative scores can only be compared within an individual. Fig. 2 demonstrates the differences between ipsative and rating scale assessments.

The relative nature of ipsative scores also means that two individuals with identical ipsative scores could have very different response patterns. For example, in Fig. 2, Persons 1 and 3 have identical scores for Resilience and Teamwork, but Person 3 has higher scores for both traits when considering the absolute values of their scores. The interdependence of item scores in ipsative data has implications for statistical analyses that typically depend on correlated items but are not wholly dependent on each other. Ignoring item co-dependencies can result in inaccurate estimates (Baron, 1996). Modern psychometric methods incorporate these dependencies to estimate reliability and assess model fit. In particular, the Thurstonian-IRT (TIRT) model has been proposed to accurately examine the characteristics of FC models.

Fig. 2 - Ipsative Data Example

Example: Three people answered six items on a FC and 5-point rating scale with 5 indicating "Completely Agree". Three items measured Teamwork (TW) and three measured Resilience (RES).





Ipsative Data Example

| Trait | Block 1 | | Block 2 | | Block 3 | | Totals | | |
|---|---|---|---|---|---|---|---|---|---|
| | TW | RES | TW | RES | TW | RES | | | |
| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | TW | RES | Sum |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | 3 | 0 | 6 |
| 2 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 3 | 6 |
| 3 | 0 | 1 | 0 | 1 | 0 | 1 | 3 | 0 | 6 |

Rating Scale Data Example

| Trait | TW | RES | TW | RES | TW | RES | Totals | | |
|---|---|---|---|---|---|---|---|---|---|
| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | TW | RES | Sum |
| 1 | 1 | 2 | 1 | 2 | 1 | 2 | 3 | 6 | 9 |
| 2 | 4 | 1 | 4 | 2 | 4 | 3 | 12 | 6 | 18 |
| 3 | 3 | 5 | 2 | 5 | 2 | 5 | 7 | 15 | 22 |

Data Comparison

| Trait | TW | | RES | |
|---|---|---|---|---|
| | Ipsative | Rating Scale | Ipsative | Rating Scale |
| 1 | 3 | 3 | 0 | 6 |
| 2 | 0 | 12 | 3 | 6 |
| 3 | 3 | 7 | 0 | 15 |

### ***The Thurstonian-IRT Model***

The TIRT model is conceptually grounded in Thurstone's (1927) Law of Comparative Judgment. According to this law, item A is preferred over item B in a pairwise comparison ($y_{AB}$) if the latent utility for item A ($t_A$) is higher than that of the latent utility for item B ($t_B$; Brown & Maydeu-Olivares, 2011). This can be represented as:

$$y_{AB} = 1 \; if \; y^*_{AB} = t_A - t_B \geq 0 \; else \; 0 \quad (1)$$

Here, $y^*_{AB}$ denotes the difference in latent utilities. When there are more than two items in a block, the rankings, are recoded into a set of ñ = *n* (*n*-1)/2 pairwise comparisons, where ñ is the number of pairwise comparisons and n is the number of items in a block. In a triad block with items A, B, and C there are three pairwise comparisons {A,B},{A,C}, and {B,C}. Data are recoded in line with equation 1, such that when the first item is greater than the second it is given a 1, indicating it is preferred, otherwise it is given a 0. Suppose the following rankings were given to three items where 3 is the highest ranking and 1 is the lowest: A = 3, B = 1, C = 2. This would result in {A, B} = 1 as A is higher than B, {A, C} = 1, and {B, C} = 0 as B is lower than C. When an item is ranked higher than the other in the comparison this implies that it also has a higher utility to the respondent. After data is recoded, the TIRT model can be specified as a second-order model, also known as the Thurstonian Factor Model (Maydeu-Olivares & Böckenholt, 2005), to examine model fit and item parameters, whereas a first-order model is used when estimating trait scores is the goal. The specification of both models is further discussed in Chapter 4.

### Forced-Choice for Non-Cognitive Assessments

The FC response style has uses in other research areas, but its main use in psychological research is in non-cognitive assessments. Non-cognitive assessments aim to evaluate a broad spectrum of skills and characteristics that exceed conventional cognitive capacities typically gauged by cognitive tests (Duckworth & Yaeger, 2015). Unlike cognitive skills, which primarily assess intellectual and subject-specific knowledge, non-cognitive skills offer a supplementary perspective on people's abilities, capturing invaluable skills such as teamwork, resilience, and conscientiousness. The importance of these skills has seen increased acknowledgment in their ability to predict academic, economic, occupational, and health outcomes, as well as psychological success (Almlund et al., 2011; Duckworth & Yaeger, 2015; Egalite et al., 2016; Heckman et al., 2006; Heckman & Kautz, 2012; Jones et al., 2015; Ohman, 2015; Smithers et al., 2018; Taylor et al., 2017). Research suggests that these skills are malleable throughout adolescence (Heckman & Kautz, 2013; Kautz et al., 2014). Some studies have shown a positive effect on increasing these skills through interventions (Alagan et al., 2014; Kautz et al., 2014; Kautz & Zanoni, 2014; Martins, 2010). This growing body of research corresponds with a growing desire to measure and create interventions for fostering these skills (Duckworth & Yeager, 2015; Garcia, 2016; Melnick et al., 2017). Measures such as the Character Skills Snapshot (Enrollment Management Association, 2023), the Mosaic (ACT, 2022), the OPQ32r (Brown & Bartram, 2011) and many others cataloged by the Collaborative for Academic, Social, and Emotional Learning (CASEL, 2020) are being increasingly used in high-stakes scenarios. However, using these assessments to make decisions about employment, admissions, or other high-stakes assessment scenarios requires them to have a strong validity argument.

The Standards for Educational and Psychological Testing, henceforth the Standards, define validity as the extent to which evidence and theory support the interpretations of test scores for proposed uses of tests (American Educational Research Association [AERA] et al., 2014). This process is considered unitary, where various forms of evidence collectively contribute to a valid interpretation of scores (Kane, 2013). The sources of validity evidence can come from examining test content, response processes, internal structure, relationships to other variables, and consequences of testing (AERA et al., 2014). All five evidentiary types are vital in supporting the use of an assessment and are summarized in Table 1.

Table 1 - Types of Validity Evidence

| Evidence Type | Definition | Evidence Considerations |
|---|---|---|
| Test Content | The construct has been identified and defined, and content experts were consulted. | Was a construct identified, and if yes, was it defined?<br>Were content experts consulted? |
| Response Process | Whether the theory was examined or individual responses were systematically tested. | Were response processes to items tested with individuals (e.g., think-alouds)? |
| Internal Structure | Any statistical technique to determine if the model reflects the construct it proposes to measure (e.g., factor analyses) and that the scores are reliable.<br>Psychometric properties are equal across groups. | Were any statistics that test for internal structure reported or measured?<br>Was IRT or factor analysis used to test for equality across groups? |
| Relationships with Other Variables | Evidence for how the construct is related to other variables. | This may include correlations, testing for group differences, and predictive testing of the measure with other variables. |
| Consequences of Testing | Includes positive or negative consequences. Evidence that pertained to how the score was interpreted or other evidence of the score's applied purpose and utility was noted (Messick, 1975). | Were consequences considered? |

Research on the strength of the validity arguments made by non-cognitive assessments is limited, but there are some reviews that detail the ones they present (Humphries & Kosse, 2017; Egalite et al., 2016). For instance, Cox and colleagues (2019) focused on assessments measuring collaboration, perseverance, and self-regulated learning, while Cordier and colleagues (2015) centered on search terms related to social-emotional skills (SEL). Halle and Darling-Churchill (2016) reviewed assessments based on their validity, reliability, and the range of skills assessed, primarily focusing on SEL-related constructs. These reviews indicated that internal structure evidence is largely present, but other evidence types are less so, which is inconsistent with the Standards. With increased interest in using non-cognitive assessments for high-stakes decisions, strong validity arguments are needed, and a mechanism for controlling response bias, such as formatting the test as FC, should be present.

## Objectives and Overview of The Dissertation

In psychological research, the FC format is used in the context of non-cognitive assessments. With increasing evidence showing that non-cognitive skills can predict success, there has been more interest in using them to make high-stakes decisions (Duckworth & Yaeger, 2015). For example, the Enrollment Management Association's FC assessment, the Character Skills Snapshot, is used to inform admissions decisions into private schools (EMA, 2023). Meanwhile, the OPQ32, one of the most popular workplace assessments, is used in personnel selection. When a test is used to make a potentially life-altering decision, it should have a strong validity argument. However, the validity arguments of non-cognitive assessments used today are less well-studied than their cognitive counterparts. Using non-cognitive tests in high-stakes contexts also introduces additional challenges, with the potential for respondents to misrepresent themselves. This may occur when they "fake" their responses or try to make themselves look the best possible instead of answering honestly. FC has been shown to effectively reduce response bias in high-stakes contexts. The control over response bias makes FC a useful tool for ensuring the accuracy of test scores when the stakes are high. However, while various authors have made recommendations on some aspects of their validation, there has yet to be a comprehensive review that synthesizes and lays out the state of the field after 84 years of their usage. There is also room for further advancement of research on the format. My dissertation seeks to understand the issues of current non-cognitive assessment validity. This understanding will then inform the primary objective of this work in synthesizing and advancing the state of methodology for FC assessment.

In Chapter 2, I conduct a mapping review of 46 non-cognitive assessments in the grey and peer-reviewed literature. The chapter begins with an introduction to non-cognitive assessment and describes the framework for forming a validity argument proposed by the Standards (AERA et al., 2014). We then summarize the five types of validity evidence and how they support the validity argument of a test. After this, we discuss the methodology and results of the mapping review I conducted. A mapping review is useful to "identify, select, and critically appraise relevant research" (Moher et al., 2009, p. 1). We then conclude with recommendations on how the field can improve the validity of arguments of non-cognitive assessments based on the findings of the review. This chapter has received a revise and resubmit decision from the *Journal of Applied Measurement in Education.*

In Chapter 3, I review and synthesize 84 years of research on FC assessment. I first describe FC assessment and then trace its history from the 1940s to the present day. Next, I synthesize the methodology for constructing an FC assessment at each phase of development. Finally, we discuss arguments against the use of FC and identify areas for future research, which directly resulted in Chapter 4. As part of this work and to promote engagement with the historical literature, we also created an annotated bibliography for the more than eight decades of articles that we reviewed. This chapter aimed to be the first step in creating a structured construction procedure by laying out all that is currently known in the field. This chapter is under review at *Psychological Methods*.

Chapter 4 presents an approach for detecting DIF in FC assessments. Identifying blocks displaying DIF is vital to constructing a fair assessment. Lee and colleagues (2021) examined a latent-scoring-based approach for testing DIF with promising results. However, their study did not include cases of model misspecification, which is likely in real-world assessment contexts. They also used the first-order TIRT model, which does not allow for a separate examination of nonuniform and uniform DIF. In this chapter, we replicate some of their work using the second-order TIRT model, which can parse out both types of DIF, and examine conditions relating to model misspecification. We found that the latent-scoring approach does not accurately detect DIF or non-DIF items when misspecification is present. My results indicate further work is needed to uncover a reliable method for detecting DIF in FC tests. This chapter was submitted as a registered report and was given an in-principle Stage 1 acceptance at *Peer Community In* (PCI). This indicates the background, methods, code, hypotheses, and proposed analyses were peer-reviewed and approved.

In Chapter 5, the contents of this dissertation are summarized in full. I detail the implications and limitations of each chapter. Then, I consider the implications of this work as a whole. I conclude with an evaluation of the broad limitations of the current work and directions for future research.

# Chapter 2. Validity Evidence Reported for Non-Cognitive Educational Assessments: Current Practice and Next Steps

# Chapter 2. Validity Evidence Reported for Non-Cognitive Educational Assessments: Current Practice and Next Steps

Manuscript: **Plantz, J.W.**, Elbaz, S., Brenneman, M., & Flake, J.K. (invited for revision, *Journal of Applied Educational Measurement*). Validity evidence reported for non-cognitive educational assessments: Current practice and next steps.

## Abstract

There is a growing emphasis on assessing non-cognitive skills in educational settings because they are predictive of educational and workplace success. To uncover the state of validity evidence for these assessments, we conducted a mapping review of 46 studies of assessments published in the peer-reviewed and grey literature. We coded for the five sources of validity evidence consistent with the Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014). We found that most assessments have evidence of content and internal structure, whereas evidence of the consequences of testing and response processes is often lacking. Internal structure evidence is often limited to reliability coefficients with almost no reporting on measurement bias. Validation sample demographics are not routinely reported, making it unclear if they are representative of the diverse groups of students who take the assessments. These results suggest that the issue of fairness and bias in non-cognitive assessments is an important but neglected aspect of validation for assessment developers and practitioners to consider. We conclude with recommendations on how measurement professionals can address these issues.

## Validity Evidence Reported for Non-Cognitive Educational Assessments: Current Practice and Next Steps

Non-cognitive assessments are used in educational settings to measure behaviors, attitudes, personality traits, psychosocial attributes, executive functions, and strategies associated with individual success (Baker, 2013; Egalite et al., 2016; Smithers et al., 2018). The term non-cognitive emphasizes a set of skills other than cognitive ability (e.g., literacy, numeracy) that are important in education and the workforce. Non-cognitive assessments are used for a variety of purposes, including informing high-stakes decisions about students (AERA et al., 2014; Duckworth & Yaeger, 2015; Egalite et al., 2016). For example, the Character Skills Snapshot is used in enrollment decisions and the ACT Mosaic is used to inform program planning (ACT, 2022; Enrollment Management Association, 2023). High-stakes assessments require strong validity evidence, research into equity and bias, and strategies for mitigating potential negative consequences (Kane, 2013). In these contexts, it is essential to consider the test's purpose and how the validity argument supports its use (Mason, 2007; Ryan, 2002). Though these practices have become standard for cognitive assessments (Broer, 2005, for example) it is less clear what standard practice is for non-cognitive assessments as their use increases in educational settings. The purpose of the current work is to conduct a mapping review (Booth, 2016; Grant & Booth, 2009) of current validation practices for non-cognitive assessments used in scholarly research and in educational testing. Our goal was to provide a snapshot of current practice and develop recommendations that assessment developers, psychometricians, educational staff, and researchers can use to strengthen the validity evidence for non-cognitive assessments.

### Literature Review

Before turning to our review of the validity evidence, we overview the frameworks contained under the large umbrella of non-cognitive skills. Then we review the validation process, including an explanation of how different sources of validity evidence can support the valid use of non-cognitive assessments.

### *What Are Non-Cognitive Skills and Why Are They Important?*

The term "non-cognitive," encompasses a wide array of skills and characteristics beyond traditional cognitive capabilities often measured by standardized tests (Duckworth & Yaeger, 2015). While cognitive skills mainly assess intellectual prowess and subject-specific knowledge, non-cognitive skills offer an alternative lens on students' capacities. These skills stem from broader personality traits which are defined as “patterns of thoughts, feelings, and behavior” which result in a great variety of individual skills with their own definitions (Borghans et al., 2008, p. 974). The importance of non-cognitive skills, recognized since the advent of cognitive assessments (Binet & Simon, 1916), has grown over time, particularly in predicting academic, economic, and psychological success (Almlund et al., 2011; Duckworth & Yaeger, 2015; Egalite et al., 2016; Heckman & Kautz, 2012). Non-cognitive skills have significant overlap with a variety of other similar concepts such as Character Skills (Heckman et al, 2014), 21st-century skills (C21 Canada, 2012), social-emotional learning (SEL; Weissberg et al., 2015), life skills, soft skills, and competencies (Cinque et al., 2021). Each framework encompasses different skill sets with varying scopes and the framework used is often discipline-based (Cinque et al., 2021).

The five-factor personality model which encompasses conscientiousness, agreeableness, emotional stability, openness, and extraversion, has been offered as a Rosetta stone to replace the 100+ frameworks for these types of skills in a common language (Berg et al., 2017; Martin et al., 2019). Although, the various skills associated with these frameworks can be categorized into one of the five factors of personality, Soto and colleagues (2021) advocated for them being considered distinct. A key distinction is the difference between skills and traits (Duckworth & Yaeger, 2015; National Research Council, 2012), with aspects of personality being characterized as traits in contrast to non-cognitive skills. Traits are generally considered stable over time and different situations. This is juxtaposed against skills which represent what a person is capable of doing in a particular scenario (Paulhus et al., 1987) and may be able to be further developed . Traits are always present while skills can be called upon or put away to meet the situation at hand (Soto et al., 2021). Personality inventories have a long history of development and validation, and though the debate regarding how separate they are from non-cognitive skills is not settled, measures focused specifically on the skills of students are being newly developed and used. In this work we focus on these more recent measures of non-cognitive skills, rather than the personality traits that may be contained under the larger umbrella.

Regardless of the framework, there is substantial support that non-cognitive skills are important to measure and can be predictive of various educational, occupational, psychosocial, and health outcomes (Durlak et al., 2011; Evans & Rosenbaum, 2008; Heckman et al., 2006; Jones et al., 2015; Ohman, 2015; Smithers et al., 2018; Taylor et al., 2017). There is a growing base of evidence that these skills are malleable and amenable to intervention throughout adolescence (Heckman & Kautz, 2013; Kautz et al., 2014). Research suggests that non-cognitive interventions can have positive effects on outcomes from early adolescence to early adulthood (Alagan et al., 2014; Kautz & Zanoni, 2014; Martins, 2010). Recent pushes in assessing non-cognitive skills in educational measurement and practice highlight the continued need for focused attention and intervention in this area (Duckworth & Yeager, 2015; Garcia, 2016; Melnick et al., 2017). The desire to measure and nurture these skills has resulted in the creation of numerous non-cognitive assessments (Heckman et al., 2014; Kautz et al., 2014). For example, the Collaborative for Academic, Social, and Emotional Learning (CASEL, 2020) has catalogued dozens of instruments on their website. These assessments are used in a myriad of ways. For example, the Character Skills Snapshot (Enrollment Management Association, 2023) is used to support admissions decisions in K-12 private schools, whereas the ACT Mosaic (ACT, 2022) supports program planning, and the OPQ32r is used to inform hiring decisions (Brown & Bartram, 2011). However, research examining the validity evidence of non-cognitive assessments remains limited (Humphries & Kosse, 2017; Egalite et al., 2016), though some studies focusing on frameworks within the larger umbrella provide a piecemeal picture.

Cox and colleagues (2019) examined assessments measuring collaboration, perseverance, and self-regulated learning, whereas Cordier and colleagues (2015) focused on search terms relating to social skills. Halle and Darling-Churchill's (2016) review also focused on constructs that would fall within the SEL framework but sought to identify the best assessments based on their validity, reliability, and the breadth of skills measured. These reviews indicated that there are few assessments that have a strong validity argument and provide a comprehensive evaluation of validity evidence. However, each of these reviews focused on SEL constructs, which is only one of the terms used to describe non-cognitive skills. While they offer some recommendations for validation, they do not provide specific and actionable recommendations for test-makers and users.

To get a more comprehensive picture and develop broader recommendations, we conducted a mapping review. A mapping review “aims at categorizing, classifying, and characterizing patterns, trends, or themes in evidence production or publication” (Moher et al., 2009, p. 14). Our goal was to identify gaps in validation practices from a wide-ranging topic area

(Sutton et al., 2019). Our review expands on the existing literature by including broader search terms and assessments from the grey literature, with the goal of making recommendations that follow the validity evidence described in the Standards for Educational and Psychological Testing, henceforth referred to as the Standards (AERA et al., 2014). Some recent recommendations have been made on creating a strong validity argument for non-cognitive assessments (Mattern & Walton, 2022). We consider these in this review. However, our goal is to provide directed recommendations that are grounded in current practice, based on the snapshot of the validity evidence.

### ***The Validation Process***

The Standards define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p.11). Instead of considering various forms of validity (i.e., content, predictive, etc.) as separate, the Standards advocate for a unitary concept, where multiple forms of evidence collectively contribute to a coherent argument for score interpretation and use (Kane, 2013). Below, we summarize these sources of validity evidence and why they matter in the context of interpreting test scores. They include evidence related to test content, response processes, internal structure, relation to other variables, and consequences of testing (AERA et al., 2014). Though we use the Standards to guide our review, we recognize they reflect decades of scholarship from a large community of researchers. Validation can employ varied theories and methods that we cannot review here but point readers to Slaney (2017) for an in-depth history and review of validity theory.

**Evidence Based on Test Content.** Validity evidence based on test content pertains to the adequacy of test items in measuring the construct of interest. This evidence can come from reviewing the relevant theory, expert review of the items, and identifying any potential underrepresentation of the construct (AERA et al., 2014; Sireci & Faulkner-Bond, 2014). This process should involve a careful review of the literature surrounding the construct and the creation of items or questions that avoid extraneous content that would cause anything but the construct to be measured (Bandalos, 2018). For example, when assessing teamwork skills in children, items would need to be developed based on theory and existing research about collaborative behaviors among young learners. These items can then be assessed by expert judges to confirm the content aligns with the construct (Delgado-Rico et al., 2012).

Evidence of test content supports the accurate interpretation of test scores by confirming that the test measures the intended construct. This is crucial for the credibility and usefulness of the test, particularly in educational settings where decisions about curriculum development, student placement, and instructional effectiveness are made based on scores. For example, imagine a non-cognitive test is used to determine admission into a school that values students with high resilience. If a test such as the Character Skills Snapshot (Enrollment Management Association, 2023), which claims to measure resilience, is used but does not thoroughly consider how the test's content relates to this skill, it may end up measuring something irrelevant or too narrow to be useful. This would have negative consequences for the school and students.

**Evidence Based on the Response Process.** Validity evidence based on response processes supports the validity argument by demonstrating that the assessment targets the intended mechanisms underlying what people do or think when responding to an item or task (Hubley & Zumbo, 2017). Here, potential threats like unmotivated or random responding are considered as well as ensuring item content is not misunderstood. Techniques such as think-aloud interviews can be used to evaluate if respondents are thinking and then responding as intended, and to check the developmental appropriateness of items for children (AERA et al., 2014).

The interpretation of test scores is supported by this type of evidence by providing insight into the validity of the inferences made from these scores across individuals and various subgroups. For example, if test-takers use unexpected strategies such as responding in a way they think is desired or misunderstanding items, this indicates a problem with how the test content was constructed. This may introduce construct-irrelevant variance. Gathering response process evidence may also reveal further context to consider when interpreting scores for one group or another as misunderstanding may differ across groups. Understanding response processes helps in refining test items and instructions to better target the construct of interest and potential context effects (Padilla & Benitez, 2014).

**Evidence Based on Internal Structure.** Validity evidence based on the internal structure of the test examines if the relationship between items aligns with theoretical expectations. This type of evidence comes from reporting on three basic aspects: dimensionality, test and item invariance, and reliability (Rios & Wells, 2014). Each component supports the interpretation of test scores in a different way. Dimensionality indicates that the inter-item relationships support the hypothesized theoretical model. For example, if a test is meant to measure a single dimension or skill with 10 items, conducting a factor analysis or using an item response theory model should support those items being so highly related as to be measuring a single dimension. This supports the interpretation of the test score measuring only that dimension. Test and item invariance can be assessed through conducting measurement invariance testing, differential item functioning (DIF) analyses, or other analyses related to examining differences in the statistical properties of items across time or groups. Reporting this type of evidence can indicate if the test scores can be interpreted in the same way across groups and if the items measure subgroups in the same fashion. When differences exist between subgroups this could indicate item or test bias, which could result in an unfair test. For this reason, examinations of test and item bias have been linked to all other forms of validity as well as internal structure evidence (Gomez-Benito et al., 2018).  Finally, reliability provides evidence that the test scores are consistent across administrations (test-retest) or that people respond consistently to a set of items meant to measure the same construct (internal consistency).

**Evidence Based on Relationships to Other Variables.** Validity evidence based on relationships to other variables can be gathered by examining convergent and discriminant evidence or test-criterion relationships (Clark & Watson, 2019). Convergent evidence demonstrates whether scores relate to theoretically similar constructs or are distinct from other constructs (discriminant evidence). There are several ways to assess this type of evidence. Correlations can be used to establish relationships between scores by examining the convergent or discriminant relationships of the test to another trait or variable. The multitrait-multimethod matrix can examine evidence of discriminant and convergent relationships through a systematic analysis of correlation coefficients from multiple traits and multiple methods (e.g., self-report and peer-report; Campbell & Fiske, 1959). Structural equation modeling can also be used to provide evidence of discriminant and convergent evidence through a more complex system of modeling the relationships between variables and latent traits. Providing discriminant and convergent evidence demonstrates that the test is measuring something worthwhile and distinct from existing instruments.

The second form of evidence in this category comes from examining test-criterion relationships, which indicate if the test can predict outcomes of interest and differentiate between groups that are known to be different. For example, a non-cognitive assessment measuring teamwork could be evaluated based on its ability to predict students' success in team-oriented tasks. Examining this type of evidence may also involve assessing the incremental validity of the new test to determine if it is better than another that measures a similar construct.

**Evidence Based on The Consequences of Testing.** Consequences of testing refer to the impacts of test use (AERA et al., 2014). For example, a non-cognitive assessment may be used for admission into a private K-12 school. While the positive consequence could be selecting students who have a strong likelihood of success, negative consequences could include potentially excluding students who could thrive in the environment given the opportunity to learn those skills. Understanding these consequences is crucial to provide evidence for the intended and actual uses of the assessment. Consequences of testing evidence relate to the interpretation of a test score by considering the evidence for the test in its totality and how the test score may be used positively or negatively (Kane, 2013). It prompts evaluators to consider not just the technical aspects of test validity but also the ethical, educational, and social implications of how the test scores may be used and how much weight should be given to them based on the evidence. It should be considered early on and throughout the time the test is used.

## Purpose of the Current Work

Although there is potentially a great deal of validity evidence available for non-cognitive assessments, an understanding of what the standard validation practices are for them is lacking. This review aims to get a snapshot of current practice, identify gaps in the current knowledge base, and present a roadmap for strengthening the validity of test use. Our review will:

1. Review the validity evidence reported for both grey literature and peer-reviewed assessments.
2. Report areas where evidence is lacking, drawing parallels with Cordier and colleagues' (2015) systematic review of the psychometric quality of SEL assessments.
3. Provide concrete recommendations for assessment developers, researchers, and educators rooted in the Standards and stemming from gaps in real-world practice.

## Methods

We conducted a mapping review to, "identify, select, and critically appraise relevant research" (Moher et al., 2009, p.1). This approach allowed us to synthesize a large body of evidence and draw robust conclusions that have implications for practice, policy, and research directions (Siddaway et al., 2019). Our review included searches of the peer-reviewed and grey literature.

### Database Search Strategy

#### *Peer-Reviewed Articles*

We searched two central databases in education and psychology for studies that measure non-cognitive skills: the Education Resource Information Center (ERIC) and PsycINFO. The keywords from each column in Table 1 were combined to ensure that studies including assessments of non-cognitive skills and all related terms were recovered. The symbol * was used to denote truncation and searched for variations in the words (e.g., non-cognitive could also be located when searching noncognitive).

We retrieved peer-reviewed articles written in English and published from 1970 to 2020 with any human sample. We searched for articles as far back as 50 years to ensure all available literature in education and psychology was included. Only articles in peer-reviewed journals were sought, and as such the search filtered out books, reports, conference abstracts and proceedings, and other types of articles. Additionally, we reviewed the reference lists of articles that were read in full to determine if any additional articles could be obtained.

### ***Grey Literature***

We examined all assessments recommended by CASEL as well as the ACT Tessera and the Character Skills Snapshot. These two additional instruments were included because they are popular high-stakes educational assessments at the time of our review.

## Eligibility and Coding Criteria

Peer-reviewed articles were selected for this review in several stages by two reviewers. Once articles were obtained from both databases, duplicates were removed. Article titles were first screened to confirm if non-cognitive skills or any variant of the term (as described above in the search terms) were included. Next, reviewers read and coded the abstract to identify if the article described measuring non-cognitive skills and used an assessment in educational settings, with either students or teachers. The abstracts were also screened to determine if their contents were related to conducting a validity study. Only articles describing a validity study were included. Any disagreements or uncertainties were discussed to reach a common decision. If articles met these criteria, they were selected for full-text review. Both title and abstract review involved a process that erred on the side of inclusion rather than exclusion, to ensure all relevant articles were gathered. Fig. 1 outlines the selection process with a PRISMA diagram (Moher et al., 2009; Page et al., 2020).

The full-text articles were coded to extract a range of information about the assessment that included descriptive information, as well as information about the context in which the assessment was evaluated. All codes were derived from the Standards' five types of evidence used to support a validity argument. We also coded for demographic information to determine what samples were being used to gather validity evidence. Brief descriptions of each code can be found in Table 2.

### *Inclusion/Exclusion Criteria*

For the peer-reviewed assessments, we limited our search to instruments that were specifically identified as assessment tools for non-cognitive skills and collectively measured non-cognitive constructs in educational settings. We then screened the peer-reviewed articles by making use of the PRISMA guidelines (Moher et al., 2009). While we did not conduct a systematic review, these guidelines offer a systematic process for reporting our search and exclusion process. After manuscripts were identified, the abstracts were read by two coders to determine their relevance. Any articles that could not be agreed upon were removed. After exclusions, there were a total of 30 peer-reviewed articles reporting on 27 distinct instruments. For the instruments from the grey literature, we included every assessment that had a validity manual or source of information to review regarding the validity process (all assessments coded, an overview of our findings for each, and their references are in the supplementary materials of this manuscript). This resulted in 16 technical reports or validation studies from the grey literature for a combined total of 46 studies examined across both literature types.

## Results

Table 3 reports the results of our coding for each category described in Table 2. We first describe the dichotomously coded major categories, then provide more detail for each source of evidence. We found that regardless of the literature source (grey or peer-reviewed), response process evidence was least reported, appearing in only 2% (1/46) of validity studies (see Fig. 2). This was followed by evidence related to the consequences of testing, which appeared in only 28% (13/46) of studies. The remaining three types of evidence were reported far more frequently, with evidence relating to other variables appearing in 72% (33/46) of studies, content evidence in 78% (36/46), and internal structure evidence appearing most often in 93% (43/46) of the studies.

## Frequency Differences Between Grey Literature and Peer-Reviewed

Grey literature and peer-reviewed articles both followed the trend of evidence reported above. There were some differences in the frequency of evidence reported worth noting, though due to our small sample, these differences should be interpreted with caution (see Fig 2). Grey literature (*n* = 16) was the only source of response process evidence across all assessments in this review (6%, 1/16). Additionally, grey literature assessments reported evidence relating to other variables more frequently (100%, 16/16) than peer-reviewed assessments (57%, 17/30). Finally, content evidence was reported with less frequency in the grey literature (75%, 12/16) than in the peer-reviewed studies (80%, 24/30). The remaining categories of evidence were largely the same.

## Content Evidence

Though some form of content evidence was largely reported for peer-reviewed and grey literature instruments, most studies did not report all three types we coded for: theoretical framework, definition of constructs, and expert review. Reported least often was an expert review of the items across all articles (reported in 20% of the literature). Additionally, although a theoretical framework was considered in 67% (31/46) of the studies surveyed, the construct being measured was only defined in 57% (26/46) of studies.

### Internal Structure Evidence

Internal structure evidence was the most reported category of evidence. The most often reported evidence was reliability which appeared in 89% (41/46) of all articles coded. However, other evidence was far less frequent with confirmatory factor analysis being reported in 52% (24/46) of studies followed by correlations among items (30%, 14/46). All other evidence was reported less than 20% of the time with DIF and measurement invariance testing being reported in a combined 6% (3/46) of all articles.

### Evidence Related to Other Variables

Evidence relating to other variables was most often provided through correlations (57%, 26/46). Predictive difference testing was less frequently reported (41%, 19/46). Grey literature assessments reported information in both categories more frequently than the peer-reviewed assessments.

### Demographics

Demographic information for the validation sample was reported in 93% (43/46) of studies. The type of demographic information was varied, and no articles reported all types (see Fig. 3). Gender and country were the most reported categories, 72% (33/46), and 63% (29/46) respectively. Ethnicity (48%, 22/46), language (13%, 6/46), and other types of information (33%, 15/46) were reported less frequently. Peer-reviewed assessments reported on all categories except gender less frequently than grey literature.

# Discussion

The development and use of non-cognitive assessments in education is on the rise, presumably to enhance educational opportunities and improve outcomes for students. However, non-cognitive assessments cannot realize these goals without a strong base of validity evidence. The purpose of this review was to report on the state of the evidence and identify where the validation process for these assessments should go next. We reviewed 46 studies and noted three key findings: evidence of response processes and consequences is not routinely reported, thorough evidence from other categories is not reported including analyses to investigate item bias, and demographic information is missing, or samples are too homogeneous to generate strong evidence. We discuss these threats to validity and steps forward that measurement professionals can take to ensure the valid use of these assessments.

## Content-Related

While content-related evidence was reported overall, it is problematic that construct definitions did not always accompany a theoretical framework and vice versa, particularly for non-cognitive constructs that have proliferated in the literature. It is not enough to simply state what the constructs are without an accompanying definition. The content of items and their relation to the construct was also not frequently verified with expert review.

### *Recommendation*

Clear definitions for an assessment's constructs should be reported in every paper. This will ensure that definitions are consistent within the field and from study to study and help clarify how constructs are overlapping or distinct. When new items are being developed, experts should also be asked to review and verify the suspected domain targeted by the item.

## Response Processes

The least reported evidence is that of response processes. Think-aloud interviews as well as other qualitative approaches provide an opportunity for the target population to communicate their thoughts on items and can indicate a lack of clarity or the presence of construct-irrelevant variance. They may also give an early indication if an item is not understood equally by different groups of test-takers. Only one study reported evidence of a response process evaluation across key groups of test-takers. This may mean that the populations of interest were not consulted in the creation or refinement of assessment items. The "accessibility of test content to all members" (AERA et al., 2014, *p*. 26), is an important consideration in validity.

### *Recommendation*

Think-alouds and other types of cognitive interviews should play a prominent role in the creation of items and be conducted more frequently. Though they may be logistically intensive, they can ensure assessment items are being understood and provide qualitative support for quantitative findings.

## Internal Structure

Internal structure evidence was reported most often but current practice lacks a thorough examination of internal structure. The use of factor analysis, IRT modelling, or analyses of item or test invariance was reported infrequently. This is problematic as without an argument or a formal evaluation of the instrument's dimensionality, it is unknown if the theoretical structure of the assessment is supported. The second problem in current reporting standards is a lack of measurement invariance or DIF testing to determine if the scores are comparable across groups. Differences may also be captured in think-aloud interviews. Statistical analyses of item bias would further support an argument for the validity of test scores across diverse groups when conducted in isolation or in conjunction with think-aloud interviews. Without them there exists the potential for unreported construct-irrelevant variance contributing to the test scores. This results in a threat to the validity of score interpretation if the scores mean something different from group to group.

Finally, reliability coefficients were the primary statistics used to support internal structure validity arguments. However, while reliability is an important component of supporting the use of an assessment, it does not capture validity. Most articles reported using coefficient alpha (Cronbach, 1951). Alpha measures the degree of correlation between the item responses (Vaske et al., 2015). This does not support the validity argument of an assessment as it does not show the items are measuring what is intended, only that they may be measuring the same thing. It is also problematic as coefficient alpha has several noted issues in the literature. For example, if the scale measures multiple dimensions or factors, coefficient alpha may be misleading and result in inflated reliability estimates (Agbo, 2010; Streiner, 2003). Additionally, alpha is heavily influenced by the number of items in an assessment (Streiner, 2003). ***Recommendation***

The initial validation of an instrument should always report more information than a reliability coefficient. The correlations between items should always be reported as descriptive analyses. Item and model fit analyses should be assessed when sample size allows. Testing of measurement equivalence across groups and/or DIF testing should be included when there is a potential for score differences based on group identities or contexts.

## Relationship to Other Variables

Evidence relating to other variables was primarily focused on correlations with other constructs and assessments. This is a standard way of providing this type of validity evidence. The problem is that only about half of the articles coded reported it. Another issue in current reporting is the lack of predictive testing. This type of analysis may indicate when there are group differences present in how test scores predict a criterion. Ideally, there is no difference in how test scores predict criteria unless explicitly planned for or hypothesized. This would need to be further examined, resolved, or discussed.

### ***Recommendation***

There are two recommendations for researchers to consider. Conduct correlational, multitrait-multimethod, or other similar analyses to examine the convergent and discriminant evidence. Secondly, when key variables are available to examine, predictive testing should be conducted. This may come later in the validation process but is important to follow up on.

## **Consequences of Testing**

The reporting of item bias, differential prediction, and responses processes across diverse groups is not routine practice, consistent with only 28% of articles reporting on the consequences of testing. These three types of evidence can indicate when a measure is performing differently across groups, indicative of potential measurement bias (Bandalos, 2018). Group differences could be caused by construct-irrelevant variance, causing bias and requiring a thorough evaluation of the consequences for fairness and social justice. There is a lack of relevant analyses being reported on group differences and limited consideration of what these differences could mean for score interpretation, use, and ultimately the consequences for schools, families, and students. Group differences are indicative of a larger consideration of the consequences of testing. Considering the consequences also indicates that the opportunity cost of testing, consideration of the test's purpose, how stakeholders are affected both positively and negatively by the use of the test, and as many potential outcomes as possible have been considered (AERA et al., 2014). Specific reporting of these features was not frequent.

### *Recommendation*

The purpose of the test should be explicitly stated at the outset of constructing the assessment. The types of evidence that could assist in determining if a measure is biased should be examined during the validation of an assessment. The results of these analyses should be discussed, and the consequences of any group differences considered in the consequences of using the assessment in a specified way. Differences should be discussed with key stakeholders such as assessment users, educators, and families. When this information cannot be obtained it should be discussed as a limitation of the assessment.

## Limitations

The first limitation of this study is that we did not capture all grey literature and possibly missed articles in the peer-reviewed literature. For the grey literature we relied on only one database and there are more non-cognitive assessments not recommended by CASEL. Secondly, we were not always able to find a full validity manual for grey literature assessments and used what was available. Another limitation of this work is that often validity evidence tends to be spread across multiple studies. It is possible that during this review we did not fully capture each assessment's full validity evidence. If a different review approach was taken and more articles were included, this may have changed the recommendations we made. In the revision of this paper for submission we will address these comments by conducting a more thorough review of the validity evidence of each instrument we identified and adjusting our recommendations accordingly. This study also only focused on non-cognitive constructs as they are typically discussed in the field of educational assessment. As such, we did not include personality assessments which have a longer history of development (see Mammadov, 2022; Propat, 2009), as well as work force assessments used on adults (e.g. the OPQ32r, Brown & Bartram, 2011). Finally, while we extrapolated this review out to make recommendations for the field at large, this review is limited by the fact that these recommendations are based only on the assessments we reviewed.

## Conclusion

The interest in measuring non-cognitive skills is increasing as well as the number of assessments being created to measure them. As use increases, so too do the stakes for schools, educators, families, and students. Through this work, we have identified several areas in which researchers and assessment developers need to improve the state of the evidence. In particular, the consideration of student diversity throughout the validation process is needed, from ensuring samples are representative, items are understood by all students, and psychometric properties are equivalent across groups. While making an assessment with a strong validity argument is onerous, we have outlined a path that will enable non-cognitive assessments to realize their potential of supporting students’ success.

# References

References marked with an asterisk indicate studies were included in the mapping review.

Agbo, A. A. (2010). Cronbach's alpha: Review of limitations and associated recommendations. *Journal of Psychology in Africa*, *20*(2), 233-239. https:///10.1080/14330237.2010.10820371

*Act Tessera .(2018). *Act Tessera Technical Bulletin.* Act Tessera. https://www.act.org/content/dam/act/unsecured/documents/pdfs/tessera-tech-bulletin.pdf

ACT. (2022). Mosaic by ACT Social Emotional Learning Assessment Technical Manual: Elementary School Assessment. https://www.act.org/content/dam/act/unsecured/documents/2022/R2138-Mosaic-SEL-ES-Technical-Manual-03-2022.pdf

*Adebayo, B. (2008). Cognitive and non-cognitive factors: Affecting the academic performance and retention of conditionally admitted freshmen. *Journal of college admission*, *200*, 15-21. https://files.eric.ed.gov/fulltext/EJ829456.pdf

Algan, Y., Beasley, E., Vitaro, F., & Tremblay, R. (2014). The impact of non-cognitive skills training on academic and non-academic trajectories: From childhood to early adulthood.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In Handbook of the Economics of Education (Vol. 4, pp. 1-181). Elsevier.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. [AERA, APA, & NCME] (2014). *Standards for educational and psychological testing*. American Educational Research Association.

*Anagün, S. S. (2018). Teachers' perceptions about the relationship between 21st century skills and managing constructivist learning environments. *International Journal of Instruction*, *11*(4), 825-840. https://files.eric.ed.gov/fulltext/EJ1191700.pdf

Anderson, C., Turner, A. C., Heath, R. D., & Payne, C. M. (2016). On the meaning of grit…and hope…and fate control…and alienation…and locus of control…and…self-efficacy…and…effort optimism…and…. *Urban Review*, *48*(2), 198–219. https://doi.org/10.1007/s11256-016-0351-3

*Arslangilay, A. S. (2019). 21st century skills of CEIT teacher candidates and the prominence of these skills in the CEIT undergraduate curriculum. *Educational Policy Analysis and Strategic Research*, *14*(3), 330-346. https://doi.org/10.29329/epasr.2019.208.15

*Aworanti, O. A., Taiwo, M. B., & Iluobe, O. I. (2015). Validation of modified soft skills assessment instrument (mossai) for use in Nigeria. *Universal Journal of Educational Research*, *3*(11), 847-861. https://doi.org/10.13189/ujer.2015.031111

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Baker, E. L. (2013). Testing in a global future. http://gordoncommission.info/rsc/pdfs/baker_testing_global_future.pdf

Booth, A. (2016). EVIDENT Guidance for reviewing the evidence: A compendium of methodological literature and websites. University of Sheffield, Sheffield, 13.

*Brill, R. T., Gilfoil, D. M., & Doll, K. (2014). Exploring predictability of instructor ratings using a quantitative tool for evaluating soft skills among MBA students. *American Journal of Business Education*, *7*(3), 175-182. https://files.eric.ed.gov/fulltext/EJ1053626.pdf

Broer, M. (2005). Ensuring the fairness of GRE writing prompts: Assessing differential difficulty. ETS Research Report Series, 2005(1), i-41.

CASEL. (2020). Collaborative for academic, social, and emotional learning. https://casel.org

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

*Cevik, M. & Senturk, C. (2019). Multidimensional 21st century skills scale: Validity and reliability study. *Cypriot Journal of Educational Sciences*, *14*(1), 11-28. https://files.eric.ed.gov/fulltext/EJ1211726.pdf

*Chamorro-Premuzic, T., Arteche, A., Bremner, A. J., Greven, C., & Furnham, A. (2010). Soft skills in higher education: Importance and improvement ratings as a function of individual differences and academic performance. *Educational Psychology*, *30*(2), 221-241. https://doi-org/10.1080/01443410903560278

Cinque, M., Carretero, S., & Napierala, J. (2021). *Non-cognitive skills and other related concepts: towards a better understanding of similarities and differences* (No. 2021/09). JRC Working Papers Series on Labour, Education and Technology.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412.

Cordier, R., Speyer, R., Chen, Y.-W., Wilkes-Gillan, S., Brown, T., Bourke-Taylor, H., Doma, K., & Leicht, A. (2015). Evaluating the psychometric quality of social skills measures: a systematic review. PLOS ONE, *10*(7), e0132299.

*Coryn, C. L., Spybrook, J. K., Evergreen, S. D., & Blinkiewicz, M. (2009). Development and evaluation of the social-emotional learning scale. *Journal of psychoeducational Assessment*, *27*(4), 283-295. https://doi.org/10.1177/0734282908328619

Cox, J., Foster, B., & Bamat, D. (2019). A review of instruments for measuring social and emotional learning skills among secondary school students. REL 2020-010. Regional Educational Laboratory Northeast & Islands.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, *16*(3), 297-334.

C21 Canada. (2012). A 21st century vision of public education for Canada. http://www.c21canada.org/wp-content/uploads/2012/05/C21-Canada-Shifting-Version-2.0.pdf

Delgado-Rico, E., Carretero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology España*, *12*(3), 449-460.

*DeRosier, M. E., & Thomas, J. M. (2018). Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Developmental Psychology*, *55*, 52-61. https://doi.org/10.1016/j.appdev.2017.03.001

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237-251. https://doi.org/10.3102/0013189X15584327

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students’ social and emotional learning: A meta-analysis of school-based universal interventions. Child Development, 82(1), 405-432.

Egalite, A. J., Mills, J. N., & Greene, J. P. (2016). The softer side of learning: Measuring students’ non-cognitive skills. *Improving Schools*, *19*(1), 27-40. https://doi.org/10.1177/1365480215616313

*Elliott, S. N., Davies, M. D., Frey, J. R., Gresham, F., & Cooper, G. (2018). Development and initial validation of a social emotional learning assessment for universal screening. *Journal of Applied Developmental Psychology*, *55*, 39-51. https://doi.org/10.1016/j.appdev.2017.06.002

Enrollment Management Association. (2023). Summary of the Snapshot research and findings. https://assets-global.website-files.com/62b5ff6837362e413f2bfed4/64e26e4744c35d763506450b_summary-of-the-snapshot-research-and-findings.pdf

*Esen-Aygun, H., & Sahin-Taskin, C. (2017). Identifying psychometric properties of the social-emotional learning skills scale. *Educational Policy Analysis and Strategic Research*, *12*(2), 43-61. https://files.eric.ed.gov/fulltext/EJ1170193.pdf

Evans, G. W., & Rosenbaum, J. (2008). Self-regulation and the income-achievement gap. *Early Childhood Research Quarterly*, *23*(4), 504–514. https://doi.org/10.1016/j.ecresq.2008.07.002

Flake, J. K., & Petway, K. T., II. (2019). Methodologies for investigating and interpreting student-teacher rating incongruence in noncognitive assessment. *Educational Measurement Issues and Practice*, *38*(1), 63–77. https://doi.org/10.1111/emip.12201

Furlong, M. J., You, S., Renshaw, T. L., Smith, D. C., & O'Malley, M. D. (2014). Preliminary development and validation of the social and emotional health survey for secondary school students. *Social Indicators Research*, *117*, 1011-1032. https://doi.org/10.1007/s11205-013-0373-0

Garcia, E. (2016). The need to address non-cognitive skills in the education policy agenda. In *Non-cognitive skills and factors in educational attainment* (pp. 31-64). Brill.

*Gheith, E., & Aljaberi, N. M. (2017). The effectiveness of an interactive training program in developing a set of non-cognitive skills in students at university of Petra. *International Education Studies*, *10*(6), 60-71. https://doi.org/10.5539/ies.v10n6p60

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109..

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health information & libraries journal, *26*(2), 91-108.

*Gresham, F. M. & Elliott, S. N. (n.d.). *SSIS SEL edition rating form – student (elementary & secondary levels): Summary of key psychometric evidence.* *https://measuringsel.casel.org/wp-content/uploads/2018/10/SSIS-SEL-ed-Student-Rating-Form-psychometrics-10_28_18.pdf*

Halle, T. G., & Darling-Churchill, K. E. (2016). Review of measures of social and emotional development. *Journal of Applied Developmental Psychology*, *45*, 8–18.

Heckman, J. J., Humphries, E. J., & Kautz, T. (2014). *The myth of achievement tests: The GED and the role of character in American life*. The University of Chicago Press.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451-464.

Heckman, J. J., & Kautz, T. (2013). Fostering and measuring skills: Interventions that improve character and cognition.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, *24*(3), 411-482.

Hessen, P., Lee, S. D., & Kuncel, N. R. (2022). Assessment of social and emotional learning for admissions and selection. In *Assessing Competencies for Social and Emotional Learning* (pp. 169-188). Routledge.

Hoffman, D. M. (2009). Reflecting on social emotional learning: A critical perspective on trends in the United States. *Review of Educational Research*, *79*(2), 533–556. https://doi.org/10.3102/0034654308325184

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. *Understanding and investigating response processes in validation research*, 1-12.

*Jensen, A., Fiedeldey-Van Dijk, & Freedman, J. (2012). SEI-YV assessor manual. Six-seconds. https://prodimages.6seconds.org/pdf/SEI-YV/SEI-YV_Manual.pdf

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health*, *105*(11), 2283-2290.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success (No. 20749). *National Bureau of Economic Research*. https://doi.org/10.3386/w20749

Kautz, T., & Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. Chicago, IL: University of Chicago.

*LeBuffe, P. A., Naglieri, J. A., & Shapiro, V. B. (2011). *The Devereux student strengths assessment— second step edition (DESSA-SSE).* Kaplan Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality*, *90*(2), 222-255.

*Maras, M. A., Thompson, A. M., Lewis, C., Thornburg, K., & Hawks, J. (2015). Developing a tiered response model for social-emotional learning through interdisciplinary collaboration. *Journal of Educational and Psychological Consultation*, *25*(2-3), 198-223. https://doi-org/10.1080/10474412.2014.929954

Martins, P. S. (2010). *Can targeted, non-cognitive skills programs improve achievement? Evidence from EPIS* (No. 5266). IZA Discussion Papers.

Mason, E. J. (2007). Measurement issues in high stakes testing. *Journal of Applied School Psychology*, *23*(2), 27–46. https://doi.org/10.1300/j370v23n02_03

Mattern, K., & Walton, K. E. (2022). Developing a validity argument for social and emotional learning assessments: Consideration of uses and evidence. In *Assessing Competencies for Social and Emotional Learning* (pp. 43-56). Routledge.

*Mavis, B., & Doig, K. (1998). The value of noncognitive factors in predicting students' first-year academic probation. *Academic medicine: journal of the Association of American Medical Colleges*, *73*(2), 201-203. https://doi.org/10.1097/00001888-199802000-00021

McCoach, D. B., Gable, R. K., Madura, J. P., McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Evidence based on relations to other variables: Bolstering the empirical validity arguments for constructs. *Instrument Development in the Affective Domain: School and Corporate Applications*, 209-248.

*McKown, C., Russo-Ponsaran, N. M., Johnson, J. K., Russo, J., & Allen, A. (2016). Web-based assessment of children's social-emotional comprehension. *Journal of Psychoeducational Assessment*, *34*(4), 322-338. https://doi.org/10.1177/0734282915604564

Melnick, H., Cook-Harvey, C. M., & Darling-Hammond, L. (2017). *Encouraging social and emotional learning in the context of new accountability*. Learning Policy Institute.

Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, *71*(3), 281–292. https://doi.org/10.1037/0022-0663.71.3.281

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mishler, E. (1979). Meaning in context: Is there any other kind? *Harvard Educational Review*, *49*(1), 1–19. https://doi.org/10.17763/haer.49.1.b748n4133677245p

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269.

*Mondell, S., & Tyler, F. B. (1981). Child psychosocial competence and its measurement. *Journal of Pediatric Psychology*, *6*(2), 145-154. https://doi.org/10.1093/jpepsy/6.2.145

*Motallebzadeh, K., Ahmadi, F., & Hosseinnia, M. (2018). Relationship between 21st century skills, speaking and writing skills: A structural equation modelling approach. *International Journal of Instruction*, *11*(3), 265-276. https://doi.org/10.12973/iji.2018.11319a

*Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. "The Devereux Student Strengths Assessment – Mini (DESSA-Mini): Assessment, technical manual, and user's guide," Apperson, 2011.

National Research Council, Division of Behavioral and Social Sciences and Education, Board on Science Education, Board on Testing and Assessment, & Committee on Defining Deeper Learning and 21st Century Skills. (2013). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. National Academies Press. https://doi.org/10.17226/13398

Öhman, M. (2015). *Be smart, live long: the relationship between cognitive and non-cognitive abilities and mortality* (No. 2015: 21). Working Paper.

*Oliveri, M., McCaffrey, D., Ezzo, C., & Holtzman, S. (2017). A multilevel factor analysis of third-party evaluations of noncognitive constructs used in admissions decision making. *Applied Measurement in Education*, *30*(4), 297-313. https://doi.org/10.1080/08957347.2017.1353989

*Ongardwanich, N., Kanjanawasee, S., & Tuipae, C. (2015). Development of 21st century skill scales as perceived by students. *Procedia-Social and Behavioral Sciences*, *191*, 737-741. https://doi-org/10.1016/j.sbspro.2015.04.716

*Ongardwanich, N., Kanjanawasee, S., & Tuipae, C. (2015). Development of 21st century skill scales as perceived by students. *Procedia-Social and Behavioral Sciences*, *191*, 737-741. https://doi.org/10.1016/j.sbspro.2015.04.716

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

*Panorama Education. (2016). *Preliminary report: reliability and validity of panorama's social-emotional learning measures.* Panorama Education. https://panorama-www.s3.amazonaws.com/files/sel/SEL-Validity-Report.pdf

*Park, N., & Peterson, C. (2006). Moral competence and character strengths among adolescents: The development and validation of the Values in Action Inventory of Strengths for Youth. *Journal of adolescence*, *29*(6), 891-909. https://doi.org/10.1016/j.adolescence.2006.04.011

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

*RAND. (2021). *Hollistic student assessment (HSA).* RAND Corporation. https://www.rand.org/education-and-labor/projects/assessments/tool/2007/holistic-student-assessment-hsa.html

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108-116.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *2*(4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Rowe, H. L., & Trickett, E. J. (2018). Student diversity representation and reporting in universal school-based social and emotional learning programs: Implications for generalizability. *Educational Psychology Review*, *30*(2), 559–583. https://doi.org/10.1007/s10648-017-9425-3

*Russo, J. M., McKown, C., Russo-Ponsaran, N. M., & Allen, A. (2018). Reliability and validity of a Spanish language assessment of children’s social-emotional learning skills.

*Psychological Assessment*, *30*(3), 416–421. https://doi-org.lib-ezproxy.concordia.ca/10.1037/pas0000508

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement Issues and Practice*, *21*(1), 7–15. https://doi.org/10.1111/j.1745-3992.2002.tb00080.x

*Seal, C. R., Beauchamp, K. L., Miguel, K., Scott, A. N., Naumann, S. E., Dong, Q., & Galal, S. (2011). Validation of a self-report instrument to assess social and emotional development. *Research in Higher Education Journal*, *14*, 1-20. https://files.eric.ed.gov/fulltext/EJ1068827.pdf

*Seal, C. R., Miguel, K., Alzamil, A., Naumann, S. E., Royce-Davis, J., & Drost, D. (2015). Personal-interpersonal competence assessment: A self-report instrument for student development. *Research in Higher Education Journal*, *27*, 1-10. https://files.eric.ed.gov/fulltext/EJ1056172.pdf

*Search Institute (2016). Technical summary: Search Institute's REACH survey. www.search-institute.org/surveys/REACH.

*Search Institute. (2013). *Developmental assets profile: Technical summary*. Search Institute. https://www.search-institute.org/wp-content/uploads/2017/11/DAP-Psychometric-Information.pdf

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, *70*(1), 747-770.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299-321.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. psychometrika, 74, 107-120.

*Six Seconds. (n.d.). CFA model fit statistics - AWG. https://measuringsel.casel.org/wp-content/uploads/2018/10/SEI_pYV_draft_psychometrics-2.pdf

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions.* Palgrave Macmillan. https://doi.org/10.1057/978-1-137-38523-9

Smithers, L. G., Sawyer, A. C. P., Chittleborough, C. R., Davies, N. M., Davey Smith, G., & Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour*, *2*(11), 867–880. https://doi.org/10.1038/s41562-018-0461-x

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103. https://doi.org/10.1207/S15327752JPA8001_18

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. Child *Development*, *88*(4), 1156-1171.

*Tican, C., & Deniz, S. (2019). Pre-service teachers' opinions about the use of 21st century learner and 21st century teacher skills. *European Journal of Educational Research*, *8*(1), 181-197. https://doi.org/10.12973/eu-jer.8.1.181

Tracey, T. J., & Sedlacek, W. E. (1989). Factor structure of the non-cognitive questionnaire-revised across samples of black and white college students. *Educational and*

*Psychological Measurement*, *49*(3), 637–648. https://doi.org/10.1177/001316448904900316

*Tracey, T. J., & Sedlacek, W. E. (1984). Noncognitive variables in predicting academic success by race. *Measurement & Evaluation in Guidance*, *16*(4), 171–178. https://files.eric.ed.gov/fulltext/ED219012.pdf

*Tripod social and emotional competency survey (Tripod Sel-C). RAND Corporation. (n.d.). https://www.rand.org/education-and-labor/projects/assessments/tool/2018/tripod-social-and-emotional-competency-survey-tripod.html

Wang, M.-T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*(4), 465–480. https://doi.org/10.1016/j.jsp.2011.04.001

Weissberg, R. P., Durlak, J. A., Domitrovich, C. E., & Gullotta, T. P. (2015). Social and emotional learning: Past, present, and future. In J. A. Durlak, R. P. Weissberg, J. A. Durlak, C. E. Domitrovich, & T. P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice*, (Vol. 634, pp. 3–19). The Guilford Press, xxii.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, *38*(1), 148–170. https://doi.org/10.3102/016237371559729

*West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE Districts. *Journal of*

*Applied Developmental Psychology*, *55*, 119-129. https://doi.org/10.1016/j.appdev.2017.06.001

*Woods-Groves, S. (2015). The Human Behavior Rating Scale–brief: A tool to measure 21st century skills of K–12 learners. *Psychological reports*, *116*(3), 769-796. https://doi.org/10.2466/03.11.PR0.116k29w0

*Yang, C., Bear, G. G., & May, H. (2018). Multilevel associations between school-wide social–emotional learning approach and student engagement across elementary, middle, and high schools. *School Psychology Review*, *47*(1), 45-61. https://doi.org/10.17105/SPR-2017-0003.V47-1

*Yüksel, M., Okan, N., Eminoglu, Z., & Akça-Koca, D. (2019). The mediating role of self-efficacy and hope on primary school students' social-emotional learning and primary mental abilities. *Universal Journal of Educational Research*, *7*(3), 729-738. https://doi.org/10.13189/ujer.2019.070312

Table 1 - Search Criteria Keywords

| **Non-Cognitive** | **Competency** | **Measure** |
|---|---|---|
| Noncognitive | Skill | Scale |
| 21st century | Factor | Tool |
| Soft | Competency | Test |
| Psychosocial | Knowledge | Questionnaire |
| social emotional learning | Learning | Assessment |
| SEL | | Inventory |
| Noncognitive | | Instrument |
| 21st century | | |

Table 2 – Code Descriptions

| Code | Description |
|---|---|
| Mentions of Evidence Broadly | |
| Content | Is test content evidence mentioned? |
| Demo | Is demographic information mentioned? |
| RP | Is response process evidence mentioned? |
| Internal | Is internal structure evidence mentioned? |
| Other_var | Is evidence related to other variables mentioned? |
| consq | Is evidence relating to the consequences of testing mentioned? |
| Test Content Evidence | |
| Theory * | Is a theory or framework described for the constructs broadly? |
| const_def * | Is there a definition of the construct being measured provided? |
| Expert_rev | Is there mention of an expert review of items? |
| Demographic Information | |
| Country * | Is a country reported for the measurement's use? |
| Gender * | Is gender reported? |
| Race/ethnicity * | Is race/ethnicity reported? |
| Language * | Is there a language variable reported (English as a second language, English learner, etc.) |
| Other_group * | Are there any other group demographic variables mentioned? |
| population | What population is being examined (1= students, 2= teachers, 3 = other)? |
| ed_lvl | Is education level (i.e., grade) reported? |
| Response Process Evidence | |
| Cog_int | Is there mention of a cognitive interview, think-aloud, or related term? |
| Internal Structure Evidence | |
| Correlations | Is correlation analysis mentioned in support of the internal structure? |
| PCA | Is principal component analysis mentioned in support of the internal structure? |
| EFA | Is exploratory factor analysis mentioned in support of the internal structure? |
| CFA | Is confirmatory factor analysis mentioned in support of the internal structure? |
| MI | Is there a measurement invariance analysis reported that is not specified? |
| SEM_other | Is there an SEM analysis reported that is not specified? |
| IRT | Is item response theory mentioned in support of the internal structure? |
| DIF | Is differential item functioning analysis mentioned in support of the internal structure? |
| Reliability * | Is there a reliability coefficient mentioned? |
| Int_notes | Extra internal structure notes |
| Evidence based on relation to other variables | |
| Corr_ov | Is there mention of the use of correlations or some other type of analysis toward identifying convergent, discriminant, criterion relationships? |
| pred_diff * | Is there mention of some type of analysis (ANOVA, t-tests, SEM, Etc.) being used to show differential prediction between groups? |
| Evidence based on consequences of testing | |
| Group_diff * | Are the consequences of group differences with regard to the scale information considered in the results? Code covers the reporting of any consideration of interpreting the results in light of group differences. |

*Notes.* * indicates a code where a short descriptive category follows to specify information relating to it.

Table 3 – Summary of Coding Results

| Code | Grey Literature (*n = 16)* | Percent | Peer-Reviewed (*n = 30)* | Percent | Total (*n* = 46) | Percent |
|---|---|---|---|---|---|---|
| Content | 12 | 75 | 24 | 80 | 36 | 78 |
| Theory | 11 | 69 | 20 | 67 | 31 | 67 |
| Cons_def | 9 | 56 | 17 | 57 | 26 | 57 |
| Expert_rev | 2 | 13 | 7 | 23 | 9 | 20 |
| Demo | 15 | 94 | 28 | 93 | 43 | 93 |
| Country | 13 | 81 | 16 | 53 | 29 | 63 |
| Gender | 12 | 75 | 21 | 70 | 33 | 72 |
| Race_ethnicity | 9 | 56 | 13 | 43 | 22 | 48 |
| Language | 4 | 25 | 2 | 7 | 6 | 13 |
| Other_group | 6 | 38 | 9 | 30 | 15 | 33 |
| RP | 1 | 6 | 0 | 0 | 1 | 2 |
| Cog_int | 1 | 6 | 0 | 0 | 1 | 2 |
| Internal | 16 | 100 | 27 | 90 | 43 | 93 |
| Corr | 4 | 25 | 10 | 33 | 14 | 30 |
| PCA | 1 | 6 | 6 | 20 | 7 | 15 |
| EFA | 3 | 19 | 6 | 20 | 9 | 20 |
| CFA | 11 | 69 | 13 | 43 | 24 | 52 |
| MI | 1 | 6 | 1 | 3 | 2 | 4 |
| SEM_other | 0 | 0 | 3 | 10 | 3 | 7 |
| IRT | 1 | 6 | 0 | 0 | 1 | 2 |
| DIF | 0 | 0 | 1 | 3 | 1 | 2 |
| Reliability | 15 | 94 | 26 | 87 | 41 | 89 |
| Other_var | 16 | 100 | 17 | 57 | 33 | 72 |
| corr_OV | 13 | 81 | 13 | 43 | 26 | 57 |
| Pred_diff | 7 | 44 | 12 | 40 | 19 | 41 |
| Consq | 4 | 25 | 9 | 30 | 13 | 28 |
| Group_diff | 4 | 25 | 9 | 30 | 13 | 28 |

Fig. 1 – PRISMA diagram for peer-reviewed article selection

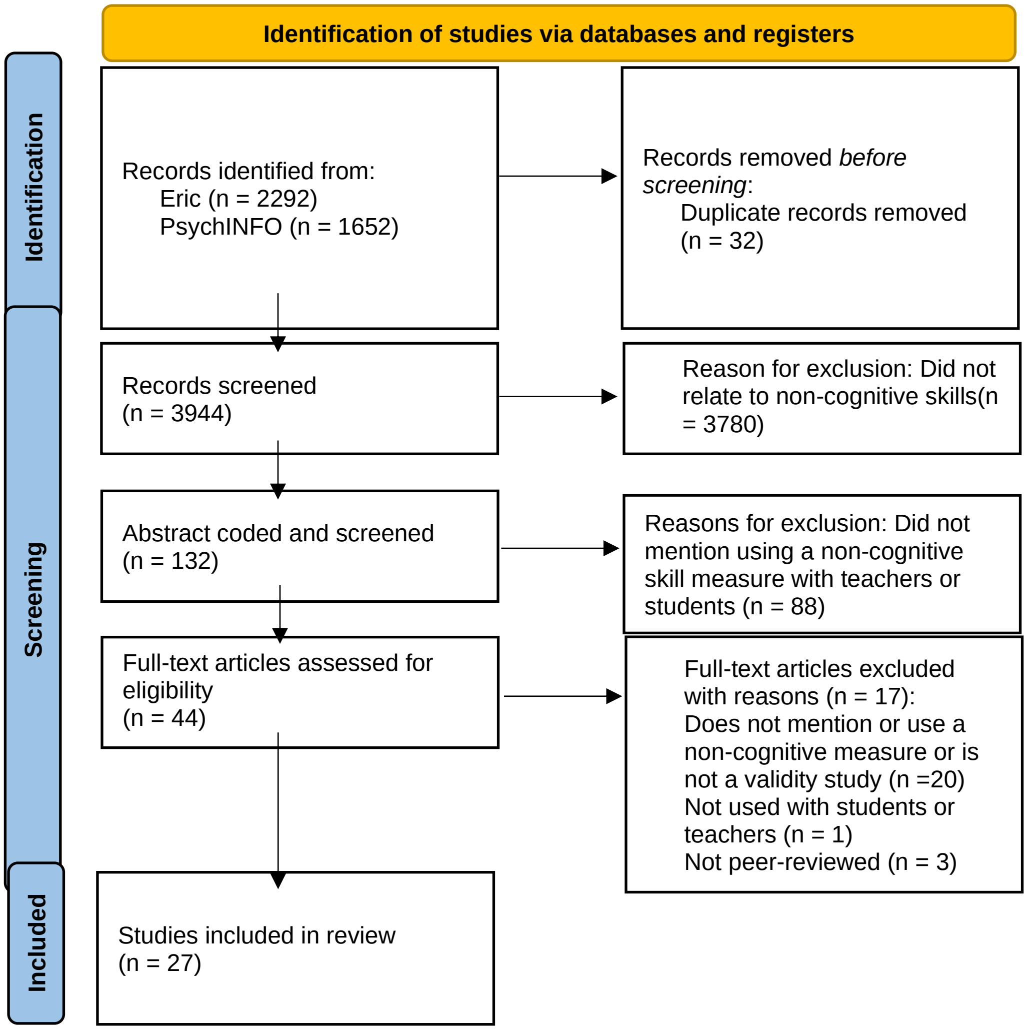

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Eric (n = 2292)
PsychINFO (n = 1652)
Records removed *before screening*:
Duplicate records removed (n = 32)
Records screened
(n = 3944)
Reason for exclusion: Did not relate to non-cognitive skills(n = 3780)
Abstract coded and screened
(n = 132)
Reasons for exclusion: Did not mention using a non-cognitive skill measure with teachers or students (n = 88)
Full-text articles assessed for eligibility
(n = 44)
Full-text articles excluded with reasons (n = 17):
Does not mention or use a non-cognitive measure or is not a validity study (n =20)
Not used with students or teachers (n = 1)
Not peer-reviewed (n = 3)
Studies included in review
(n = 27)

Fig. 2 – Differences in evidence reported between grey and peer-reviewed literature

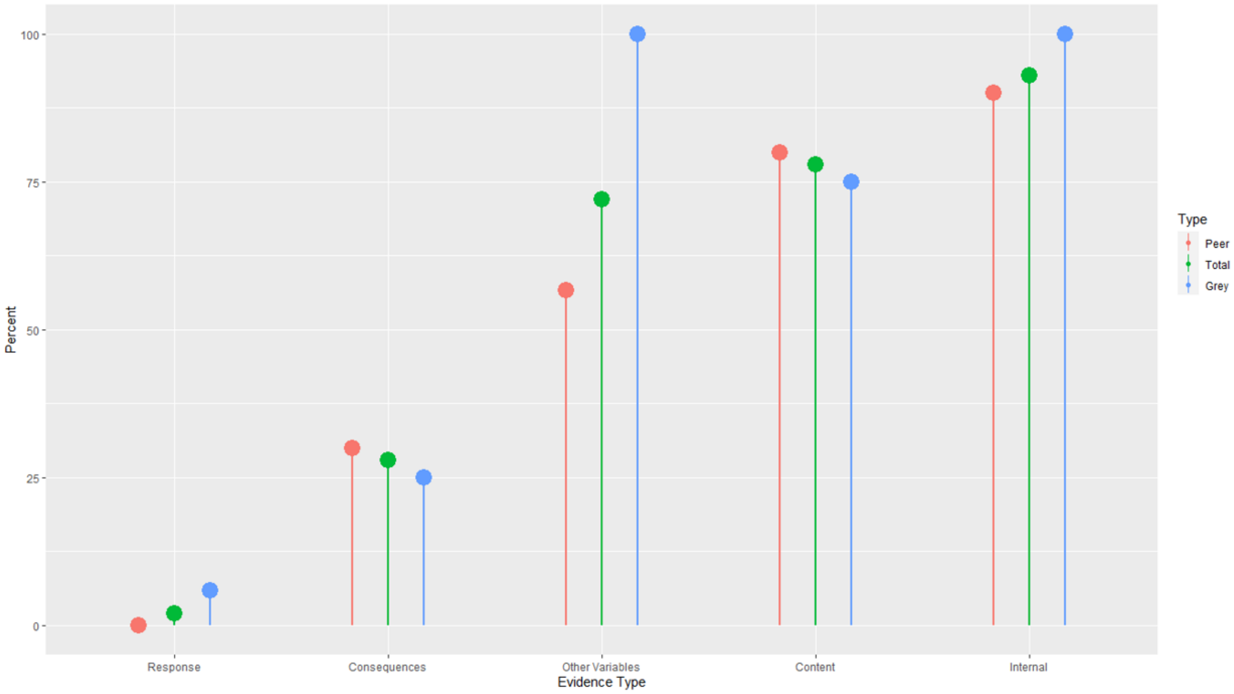
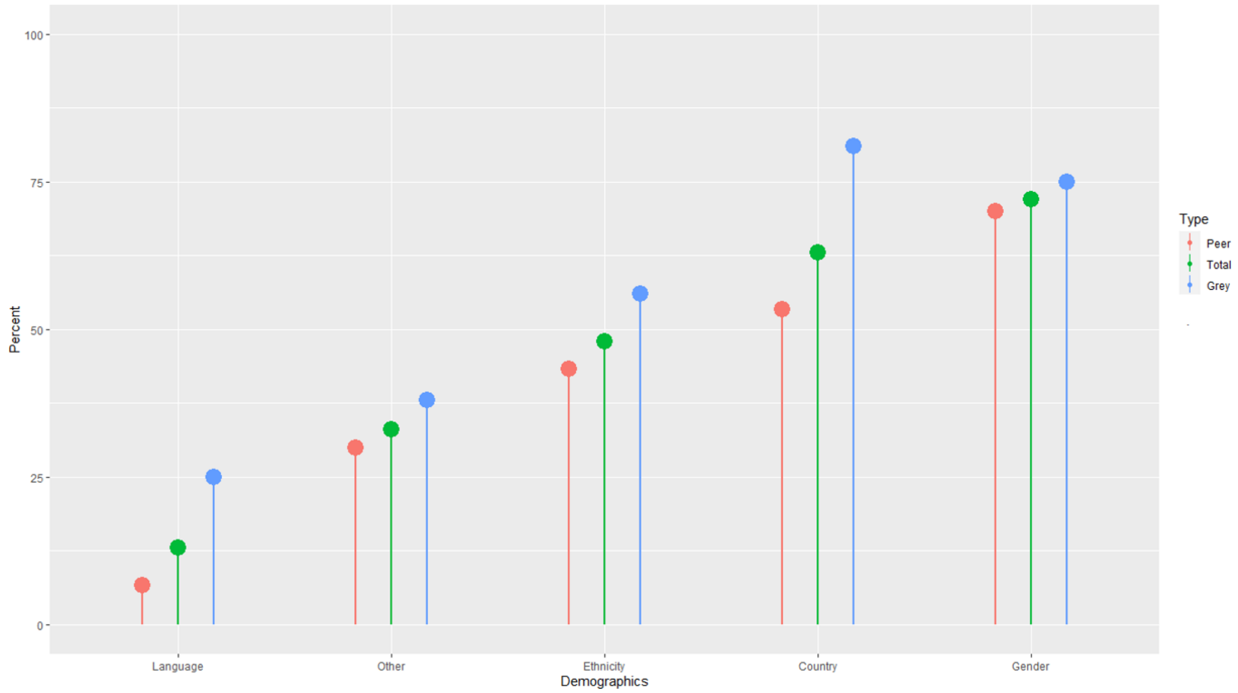


Fig. 3 – Types of Demographic Information Reported



# Appendices

## Appendix 1. Table of Assessment Specific Evidence Reported

| Instrument | Content | Demo. | RP | Internal | Other Variables | Consequences | Authors |
|---|---|---|---|---|---|---|---|
| **Peer-Reviewed** | | | | | | | |
| 21st century skills awareness questionnaire | 1 | 1 | 0 | 1 | 1 | 1 | (Motallebzadeh et al., 2018) |
| 21st century skills scale | 1 | 1 | 0 | 1 | 1 | 0 | (Anagün, 2018) |
| The Constructivist Learning Environment Scale- Teacher Form (TCLES) | 1 | 1 | 0 | 1 | 1 | 0 | (Anagün, 2018) |
| 21st century skills scale | 1 | 1 | 0 | 1 | 0 | 0 | (Ongardwanich et al., 2015) |
| Non-Cognitive Questionnaire (NCQ) | 0 | 1 | 0 | 1 | 1 | 0 | (Tracey & Sedlacek, 1984) |
| Non-Cognitive Questionnaire (NCQ) | 1 | 1 | 0 | 0 | 1 | 1 | (Adebayo, 2008) |
| Non-Cognitive Questionnaire (NCQ) | 0 | 1 | 0 | 0 | 1 | 1 | (Mavis & Doig, 1998) |
| Non-Cognitive Skill Scale | 1 | 1 | 0 | 1 | 0 | 0 | (Gheith & Aljaberi, 2017) |
| Personal Potential Index (PPI) | 0 | 1 | 0 | 1 | 0 | 0 | (Oliveri et al., 2017) |
| Personal-Interpersonal Competence Assessment (PICA) [revised from SED-I] | 1 | 1 | 0 | 1 | 0 | 0 | (Seal et al., 2015) |
| Psychosocial competence Incomplete Stories Test (PCIST) | 1 | 1 | 0 | 1 | 1 | 0 | (Mondell & Tyler, 1981) |
| SELweb-S | 1 | 1 | 0 | 1 | 1 | 1 | (Russo et al., 2018) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Social Emotional Development Instrument (SED-I). | | | | | | | (Seal et al., 2011) |
| Social-Emotional Learning Scale<br>(SELS) | 1 | 1 | 0 | 1 | 0 | 0 | (Yüksel et al., 2019) * Replaced by (Arslan & Akin, 2013) for original validation |
| Social-Emotional Learning Scale<br>(SELS) | 1 | 1 | 0 | 1 | 1 | 1 | (Coryn et al., 2009) |
| Social-Emotional Learning Scale<br>(SELS) | 1 | 1 | 0 | 1 | 0 | 0 | (Esen-Aygun & Sahin-Taskin, 2017) |
| Teaching of Social and Emotional Competencies (TSEC) subscale of the Delaware Techniques Scale–Student (DTS-S) | 1 | 1 | 0 | 1 | 1 | 1 | (Yang et al., 2018) |
| The 21st Century Learner Skills Use Scale | 1 | 1 | 0 | 1 | 1 | 0 | (Tican & Deniz, 2019) |
| 21st century skills scale | 1 | 1 | 0 | 1 | 0 | 0 | (Ongardwanich et al., 2015) |
| Devereux Student Strengths Assessment (DESSA) | 1 | 1 | 0 | 1 | 0 | 0 | (Maras et al., 2015) |
| Devereux Student Strengths Assessment-mini (DESSA-mini) | 0 | 1 | 0 | 1 | 1 | 1 | Naglieri et al.2014 |
| Goldsmith soft skills inventory | 1 | 1 | 0 | 1 | 1 | 1 | (Chamorro-Premuzic et al., 2010) |
| Human Behavior Rating Scale: Brief (HBRS: Brief) | 1 | 1 | 0 | 1 | 0 | 0 | (Woods-Groves & Choi, 2017) |
| ICT/ (Information and Communication Technology) Twenty-First Century Skills Student Questionnaire. | 1 | 1 | 0 | 0 | 0 | 1 | (Cohen et al., 2017) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Malaysian 21st century skills instrument (M-21CSI) | 1 | 0 | 0 | 1 | 0 | 0 | (Osman et al., 2010) |
| Malaysian 21st century skills instrument (M-21CSI) | 1 | 0 | 0 | 1 | 0 | 0 | (Soh et al., 2012) |
| McCann Business Soft Skills Assessment Tool | 1 | 1 | 0 | 1 | 1 | 0 | (Brill et al., 2014) |
| Modified Soft Skills Assessment Instrument (MOSSAI) | 0 | 1 | 0 | 1 | 0 | 0 | (Aworanti et al., 2015) |
| Multidimensional 21st century skills scale | 1 | 1 | 0 | 1 | 1 | 0 | (Cevik & Senturk, 2019) |
| Multidimensional 21st century skills scale | 0 | 1 | 0 | 1 | 1 | 0 | (Arslangilay, 2019) |
| The 21st Century Teacher Skills Use Scale | 1 | 1 | 0 | 1 | 1 | 0 | (Tican & Deniz, 2019) |
| | | | **Grey Literature** | | | | |
| ACT Tessera | 1 | 1 | 0 | 1 | 1 | 0 | (ACT Tessera, 2018) |
| Six Seconds Youth Version (SEI-YV) | 1 | 1 | 0 | 1 | 1 | 0 | (Jensen et al, 2015) |
| California Healthy Kids Survey (CHKS) - Social and Emotional Health (SEHS) | 1 | 1 | 0 | 1 | 1 | 1 | (Furlong et al, 2014) |
| CORE Districts Social Emotional Learning Surveys | 1 | 1 | 0 | 1 | 1 | 0 | (West et al. 2018) |
| Social and Emotional Competency Survey for Students (SEL-C) | 0 | 1 | 1 | 1 | 1 | 0 | RAND (a manual could not be located but RAND search tools report information from a preliminary technical manual) |
| Developmental Assets Profile (DAP) | 0 | 1 | 0 | 1 | 1 | 0 | (Search institute, 2013) only technical manual located |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| SSIS Social Emotional Learning (SELA) | 1 | 1 | 0 | 1 | 1 | 1 | (Elliot et al, 2018) |
| Holistic Student Assessment (HSA) | 1 | 1 | 0 | 1 | 1 | 0 | (RAND, 2021) |
| Panorama Social-Emotional Learning – Student Measures | 1 | 1 | 0 | 1 | 1 | 0 | (Panorama Education, 2016) |
| REACH survey | 1 | 1 | 0 | 1 | 1 | 0 | (Search institute, 2016) |
| VIA Youth Survey | 1 | 1 | 0 | 1 | 1 | 0 | (Park & Peterson, 2006) |
| Six Seconds Emotional Intelligence Perspective Youth Version (SEI-pYV) | 0 | 0 | 0 | 1 | 1 | 0 | Six Seconds, n.d. |
| ZooU | 0 | 1 | 0 | 1 | 1 | 1 | (DeRosier & Thomas, 2018) |
| SELweb | 1 | 1 | 0 | 1 | 1 | 0 | (McKown et al., 2016) |
| Social Skills Improvement System Rating Forms - Parent/Teacher/Student (SSIS SEL Rating Forms) | 1 | 1 | 0 | 1 | 1 | 0 | (Gresham et al., n.d.) |
| Devereux Student Strengths Assessment Second Step Edition (DESSA-SSE) | 1 | 1 | 0 | 1 | 1 | 1 | (Lebuffe et al, 2011) |

*Notes. Demo = Demographic Information; RP = Response Process Evidence;  0= No ; 1= Yes*

# Chapter 3. From the 1940s to 2020s: A Review of the Current State of Forced-Choice Methodology

## Preface to Chapter 3

Chapter 2 highlighted the importance of non-cognitive skills in educational settings and a growing desire to assess them in high-stakes contexts. A review of 46 validity studies from both the peer-reviewed and grey literature revealed gaps in the types of validity evidence reported. Our main finding was that there was a need for more evidence on the consequences of testing, response processes, and potential measurement bias. There was also a lack of thorough analysis of the internal structure of assessments. We recommended solving this by reporting more than reliability coefficients, including factor analyses, and conducting DIF testing. Making a strong validity argument becomes even more important when non-cognitive assessments are used in high-stakes contexts where response bias may be more likely. Response bias can introduce construct-irrelevant variance, jeopardizing the assessment's internal structure and invalidate the inferences made from the scores (Joubert et al., 2015). One potential way of addressing the heightened risk of response bias is through the FC test format.

The FC format can mitigate faking or socially desirable responding, a type of response bias that threatens the validity of scores. Several methods for constructing FC assessments are available, with many methodological advancements in the past decade. However, I found that the FC literature lacked a comprehensive review that synthesized all available methods for FC assessment construction and classified them into the different phases of development. To address this gap, I synthesized 80+ years of research and methodology on the FC test format. I also identified potential gaps in FC methodology and summarized the ongoing debates about the format’s ability to reduce response bias.

# From the 1940s to 2020s:

# A Review of the Current State of Forced-Choice Methodology

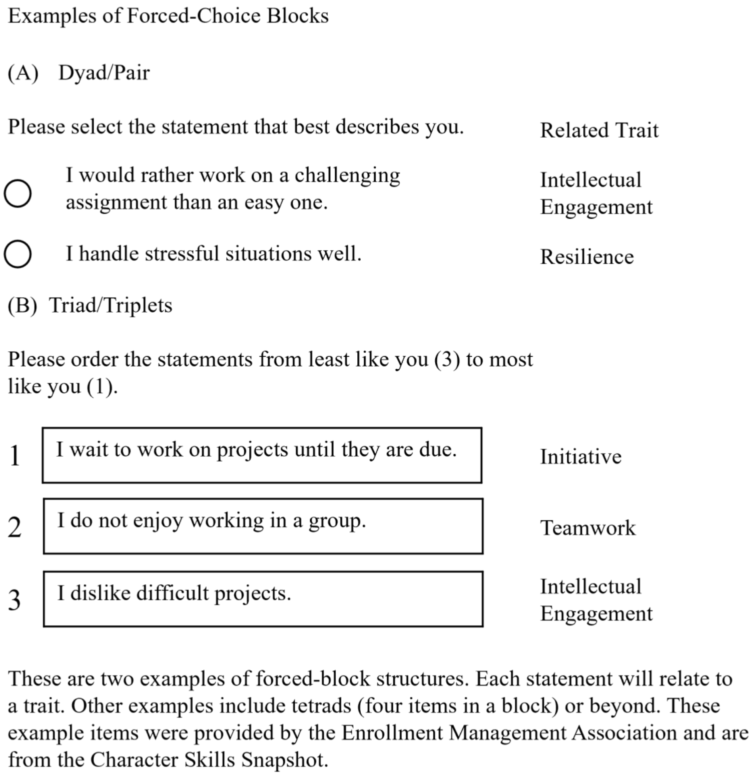
Publication [2]: **Plantz, J.W.,** Flake, J.K., Wright, K. (under review, *Psychological Methods*). From the 1940s to 2020s: A review of the current state of forced-choice methodology

## Abstract

Forced-choice measures are an alternative to rating scale surveys designed to reduce response bias, particularly socially desirable responses. Although forced-choice has been used in psychological testing since at least the 1940s, recent methodological advancements have facilitated an uptick of use in applied areas of psychology. This paper aims to provide a historical chronicling of forced-choice instruments leading into a review of the modern-day methods used for their construction. We review literature relating to the current state of item and block development, the various models used in examining internal structure, measurement bias testing, response processes, and other areas. We then discuss the debates surrounding the use of forced-choice measures and close with future directions for research. To encourage engagement with the historical literature, we provide an annotated bibliography on the more than eight decades of research that we reviewed.

## From the 1940s to 2020s: A Review of the Current State of Forced-Choice Methodology

Forced-choice measurement requires respondents to choose or rank items from a block of several, typically between two and four. Each item in the block relates to a single trait, such as in Fig. 1, where the items relate to initiative, teamwork, or intellectual engagement . The respondent would be asked to rate the items in the block from most like them to least. The primary purpose of forced-choice instruments is to reduce response biases such as socially desirable responding, halo effects, acquiescence, and extreme responses (Brown & Maydeu-Olivares, 2011). The idea is that it is harder to give a 'correct' response or provide a clearly desirable answer when rating items from different traits against one another, thus reducing possible response bias.



Examples of Forced-Choice Blocks

(A) Dyad/Pair

| Please select the statement that best describes you. | Related Trait |
|---|---|
| ○ I would rather work on a challenging assignment than an easy one. | Intellectual Engagement |
| ○ I handle stressful situations well. | Resilience |

(B) Triad/Triplets

Please order the statements from least like you (3) to most like you (1).

| | | |
|---|---|---|
| 1 | I wait to work on projects until they are due. | Initiative |
| 2 | I do not enjoy working in a group. | Teamwork |
| 3 | I dislike difficult projects. | Intellectual Engagement |

These are two examples of forced-block structures. Each statement will relate to a trait. Other examples include tetrads (four items in a block) or beyond. These example items were provided by the Enrollment Management Association and are from the Character Skills Snapshot.

Fig. 1 - Forced-Choice Example

The forced-choice format has a long history in industrial and organizational psychology as well as in educational psychology. Measures used in these settings often have high-stakes decisions associated with them (e.g., getting into a school or getting a job). High-stakes measurement of non-cognitive constructs motivates response bias, which threatens the validity of scores. The forced-choice format provides an alternative to the more easily faked rating scale format (Bartlett et al., 1960). Recent advances in methodology have increased interest in forced-choice measures. However, these advancements are scattered throughout the literature in education, psychometrics, and industrial and organizational psychology. In this paper we bring these methods together into one review. Reviews on forced-choice methodology have been put forth (Brown & Maydeu-Olivares, 2018; Li et al., 2024; Wetzel et al., 2020) but are brief or only focus on a narrow aspect of development. There has yet to be a comprehensive review on the methods for all phases of development and validation.

Our goal is to provide a comprehensive, yet digestible reference for the methodologies needed to construct and score a forced-choice (FC) measure. We also provide an annotated bibliography (see supplementary materials). We begin by giving a historical summary of forced-choice measurement to contextualize its development and provide relevant background on the format and how it evolved over time to address the unique challenges of the data it generates. We then break down the process of constructing an FC measure into three phases: item and block development, internal structure evaluation, and continued validation. For each phase we review the associated perspectives and methodology. We also discuss how each phase relates to the process of forming a strong validity argument as discussed in the Standards of Educational and Psychological Testing (AERA et al., 2014). We conclude with a discussion of where the field stands now, on-going debates, and areas that still need work.

## The History of Forced-Choice Measurement

Many modern-day methods for FC measures were born out of a need to address early challenges with the test format. We start with a circumscribed history of FC to connect the development of modern methods to the long-standing issues they are meant to ameliorate. We then present the current-day methodologies and how they are used for the development and validation of FC measures.

FC instruments were developed in the 1940s out of concerns over the rating scale and yes-no questionnaires of the time. Item creation and methodology on test construction were in their infancy. Jurgensen (1944, p. 445) noted that tests during the 1940s often had items like “Do you daydream frequently?” which has a clear, correct answer in the context of a job application. These issues led to Jurgensen’s (1944) development of the Classification Inventory, one of the first FC instruments. This was followed by Sisson's New Army Rating Scale (1948). Sisson noted an overall increase in the validity of the ratings given by officers compared to the previous ‘how much or how little’ rating system. This ‘increase’ in validity was subject to criticism and debate (e.g., Baier, 1951; Travers, 1951). Travers (1951) argued that FC measures were unnecessary and flawed in comparison to rating scale methods. Others concluded that FC measurement provided an improvement in validity over rating and yes-no rating scales (Gordon, 1951).

In the realm of personnel selection, several examinations of the test format were conducted in the 1950s (Berkshire & Highland, 1953; Ghiselli, 1954; Taylor & Wherry, 1951). Work in this era focused primarily on how to create new measures, as well as early thoughts on how to develop items and put them together in blocks (see Nagel, 1954). Scoring during this time used scoring keys where points were assigned for making the desirable choice in a pair (Sisson, 1948). For example, a candidate for a supervisory position is asked to rank which item is most like them from this dyad:

I enjoy leading a team. – I take initiative.

In the scoring key approach, a respondent would receive a score of 1 on the block if they choose “I enjoy leading a team” as this is a priority for the position. In some cases, a weighted scoring technique was used where the importance of a particular block was also considered but this was a statistically complex procedure for the time (Jurgensen, 1944). Meanwhile, the educational sector was using FC in a variety of contexts. For example, the efficacy of the test format was examined in elementary schools (Tolle, 1955), for college admissions (Kirkpatrick, 1951), faculty evaluations (Goodenough, 1957; Lovell & Haner, 1955; Mazzitelli, 1957), and student evaluations (Runyon & Stromberg, 1953).

By the late 1950s, the FC technique was being discussed as a viable alternative to rating scale measures that could reduce response bias (Cronbach, 1956) and noted for its ability to control social desirability biases (Brogden, 1954; Edwards, 1957). Several studies found FC measures better predicted key outcomes such as work performance over yes-no and rating scale measures (Newman et al., 1957). FC research was also beginning to focus on how the test format could be improved. For example, Kay (1959) examined how 'Critical Incidents', a type of format targeted at understanding how a respondent would act in a real-world situation, could be combined with FC. Meanwhile, Berkshire (1958) examined a variety of indices used for pairing items together in blocks and determined the 'Job Importance Index' to be the best at the time. This index is a rating of how important a particular item is to a given job context and foreshadows the desirability ratings that we discuss later. Some consideration was also given to how FC improved on rating scale measures with Osburne and colleagues (1954) finding evidence that FC measures provided more information on the respondent's trait score than rating scale measures after surpassing a certain threshold of item quantity.

Research in the 1960s continued to improve the format. Some developments included Bartlett and colleagues (1960) formalizing that FC controlled for social desirability bias over and above other formats (Merenda & Clarke, 1963). However, another finding pointed to the importance of properly pairing FC items, or response bias may be reintroduced due to an item being clearly preferable (Brogden, 1954; Feldman & Corah, 1960). Evidence for the efficacy of the test format in reducing response bias accumulated (Saltz et al., 1962), though these claims were not without scrutiny (Hedberg, 1963). Concurrently, there was the continued construction of instruments for use in industry (Cunningham, 1964; Schwartz & Gekoski, 1960) and in other areas such as clinical and personality measurement (e.g., Bergs & Martin, 1961; Berkowitz & Wolkon, 1964). In the 1960s, Zavala (1965) wrote the first review of FC methodology, which detailed issues of efficiency in reducing response bias, validity, and format concerns.

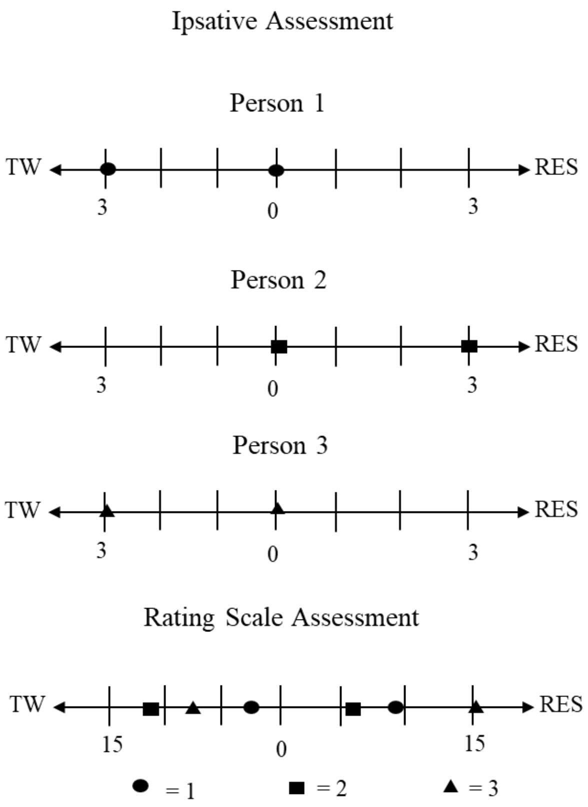
Research in the 1970s continued the debate around FC and its main advantage over rating scale measures: the reduction of social desirability bias (Bernhadwn & Fisher, 1971). Meanwhile, Jackson and colleagues (1973) presented the challenges of estimating the reliability and scores of respondents in the ipsative data that FC instruments produce. Ipsative data denotes item responses that are reliant upon each other which makes the total scores of respondents incomparable (Block, 1957; Hicks, 1970). Ipsative data also creates challenges for examining the internal structure of an FC instrument and poses a significant challenge in the validation process, which we detail in the next section. The 1970s also saw interest in the construction and validation of FC measures (Dubeck et al., 1971; Gertzen, 1976; Pearson & Powell, 1979). This interest would continue throughout the 1980s and 90s as the test format neared what we consider the modern era of FC. Over these years, a variety of FC measures were developed with more rigorous construction approaches (see Bernardin, 1987; Villanova et al., 1994). While a factor analysis approach was introduced to score the ipsative data (see Jackson & Alwin, 1980), scoring key approaches continued to be favored. These approaches require researchers to generate scoring keys rather than estimate scores from a latent variable model. While it was the best method available at the time, skepticism remained about the approach due to it not completely addressing the issue of ipsative data (Hicks, 1970).

### Ipsative Data and Scoring

The FC format creates ipsative data, where item responses are dependent upon each other, and the set of variables sums to a total score that is constant across respondents. This is due to the item and, by extension, trait scores being relative to each other in a block (i.e., as one trait moves up a point, another necessarily moves down a point). As the score of each trait is relative, it can only be compared to another from the same person, and it does not provide much information for comparing people. This differs from rating scale measures where the data can be compared across people because each item response is unique (i.e., the response on item A does not directly affect item B’s). The differences in ipsative and rating scale scores can be observed in Fig. 2. The relative nature of ipsative scores also means that although two respondents may have the same ipsative scores, their response patterns could be quite different. This is the case for persons 1 and 3 in the data comparison in Fig. 2. Both respondents have a score of three on Resilience and zero on Teamwork, but person 3 is much higher on both traits when the absolute value of their scores is considered.

Fig. 2 - Ipsative Data Example

Example: Three people answered six items on a FC and 5-point rating scale with 5 indicating "Completely Agree". Three items measured Teamwork (TW) and three measured Resilience (RES).



Ipsative Data Example

| Trait | Block 1 | | Block 2 | | Block 3 | | Totals | | |
|---|---|---|---|---|---|---|---|---|---|
| | TW | RES | TW | RES | TW | RES | | | |
| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | TW | RES | Sum |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | 3 | 0 | 6 |
| 2 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 3 | 6 |
| 3 | 0 | 1 | 0 | 1 | 0 | 1 | 3 | 0 | 6 |

Rating Scale Data Example

| Trait | TW | RES | TW | RES | TW | RES | Totals | | |
|---|---|---|---|---|---|---|---|---|---|
| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | TW | RES | Sum |
| 1 | 1 | 2 | 1 | 2 | 1 | 2 | 3 | 6 | 9 |
| 2 | 4 | 1 | 4 | 2 | 4 | 3 | 12 | 6 | 18 |
| 3 | 3 | 5 | 2 | 5 | 2 | 5 | 7 | 15 | 22 |

Data Comparison

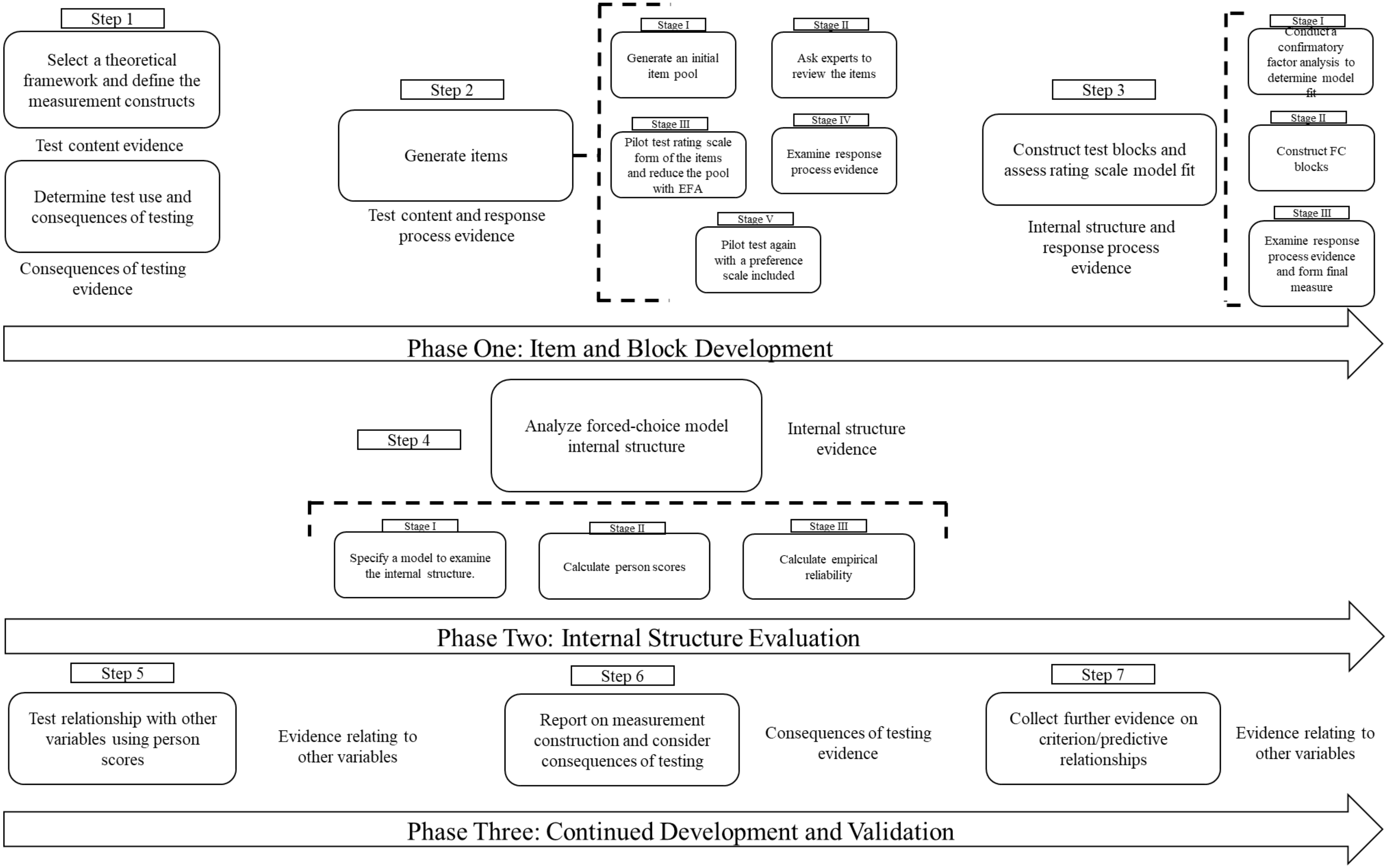
| Trait | TW | | RES | |
|---|---|---|---|---|
| | Ipsative | Rating Scale | Ipsative | Rating Scale |
| 1 | 3 | 3 | 0 | 6 |
| 2 | 0 | 12 | 3 | 6 |
| 3 | 3 | 7 | 0 | 15 |

These features of ipsative data have implications for statistical analyses. In both factor and reliability analyses, item scores are expected to be correlated but not wholly dependent on each other. If the item co-dependencies are not incorporated, factor analysis and standard reliability analyses such as Cronbach's alpha result in inaccurate estimates (Baron, 1996). Traditionally, to score tests a rubric would be used based on the block type. In a simple dyad, an item that was chosen as 'most like me' would be given a 1 and the other a 0. The scores would then be totaled and compared producing fully ipsative scores. Alternatively, a test could introduce negatively keyed-items or use one of several other procedures (see Hicks, 1970) that result in some total-score variability. For example, if a test uses a scoring approach where picking a negative item from a dyad as most preferred results in a -1 for that item (and 0 or 1 otherwise), and positively keyed items receive scores of 1 or 0 depending on the ranking, this will result in participants having different total scores on the traits. This is known as partially-ipsative data which has some between-subject comparability. However, modern psychometric methods incorporate the dependencies of ipsative data to estimate reliability, assess model fit, and generate normative scores which have full between and within subject comparability.

## Modern Methods for Constructing Forced-Choice Measurements

Developing an FC assessment is an intensive process. We reviewed eight decades of literature on FC methodologies to summarize this process in Fig. 3. Many of the steps in FC test construction do not differ from a traditional assessment. Rather than discussing each step in the context of FC assessment, we will review the relevant research at each phase of test development. These phases include item and block development, internal structure evaluation, and continued validation. For those interested in a full discussion on validity evidence and developing other types of measures, we refer them to resources such as the Standards (AERA et al., 2014) or Bandalos (2018).

Fig. 3 – Forced-Choice Measurement Construction Process


Step 1
Select a theoretical framework and define the measurement constructs
Test content evidence
Determine test use and consequences of testing
Consequences of testing evidence
Step 2
Generate items
Test content and response process evidence
Stage I
Generate an initial item pool
Stage II
Ask experts to review the items
Stage III
Pilot test rating scale form of the items and reduce the pool with EFA
Stage IV
Examine response process evidence
Stage V
Pilot test again with a preference scale included
Step 3
Construct test blocks and assess rating scale model fit
Internal structure and response process evidence
Stage I
Conduct a confirmatory factor analysis to determine model fit
Stage II
Construct FC blocks
Stage III
Examine response process evidence and form final measure
Phase One: Item and Block Development
Step 4
Analyze forced-choice model internal structure
Internal structure evidence
Stage I
Specify a model to examine the internal structure.
Stage II
Calculate person scores
Stage III
Calculate empirical reliability
Phase Two: Internal Structure Evaluation
Step 5
Test relationship with other variables using person scores
Evidence relating to other variables
Step 6
Report on measurement construction and consider consequences of testing
Consequences of testing evidence
Step 7
Collect further evidence on criterion/predictive relationships
Evidence relating to other variables
Phase Three: Continued Development and Validation

## Phase One: Item and Block Development

### *Item Development*

FC items begin development like any other assessment, with theory that forms the basis of construct definitions. Once constructs are defined, items must be written to measure them. Many researchers have studied and written guidelines on the best way to write items. Bandalos (2018, p.90-93) reviews literature on item writing and provides a concise list of guidelines. Highlights include keeping statements short (less than 20 words) that clearly tap into one idea. They suggest using language that is at a 5th-grade reading level or lower, depending on the population, and avoiding making statements of fact. Items can also be sent to content area experts to judge how well the items align to or represent the constructs. This allows for experts in the field to identify the quality of items or what traits they relate to through methods such as ranking tasks (Crocker & Algina, 1986) and Q-sorts (Nahm et al., 2002; Sireci & Faulkner-Bond, 2014). The major considerations in developing items for FC assessments are the desirability of the items and the quantity of negatively keyed items, which are central to block construction.

**Item Desirability.** Item desirability refers to how preferable respondents perceive an item given a particular context (Pavlov et al., 2021) and features prominently in FC item development compared to a traditional assessment. Examining item desirability is a necessary aspect of determining how blocks should be formed and is a concept that dates back to the early days of FC testing (see Sisson, 1948). There are various ways to determine an item’s desirability. A common approach is to ask for an explicit rating of desirability. For example, Christiansen and colleagues (2005) asked respondents to rate how desirable an item would be for job applicants on a scale of 1 (not-at-all desirable) to 9 (very desirable). These ratings may come from the population of interest (see Christiansen et al., 2005; Vasilopoulos et al., 2006) or from subject matter experts (Chernyshenko et al., 2009; Jackson et al., 2000). Some recent work has also explored using large-language-models (LLMs) to gather these ratings (Hommel, 2023). Hommel gathered desirability ratings for 521 items from 14 openly-accessible empirical datasets and then generated desirability ratings by using two customized LLMs. They found a correlation of .80 between the LLM and empirical ratings of the items. This suggests that LLMs could reduce the costs of data collection for desirability ratings for FC measures.

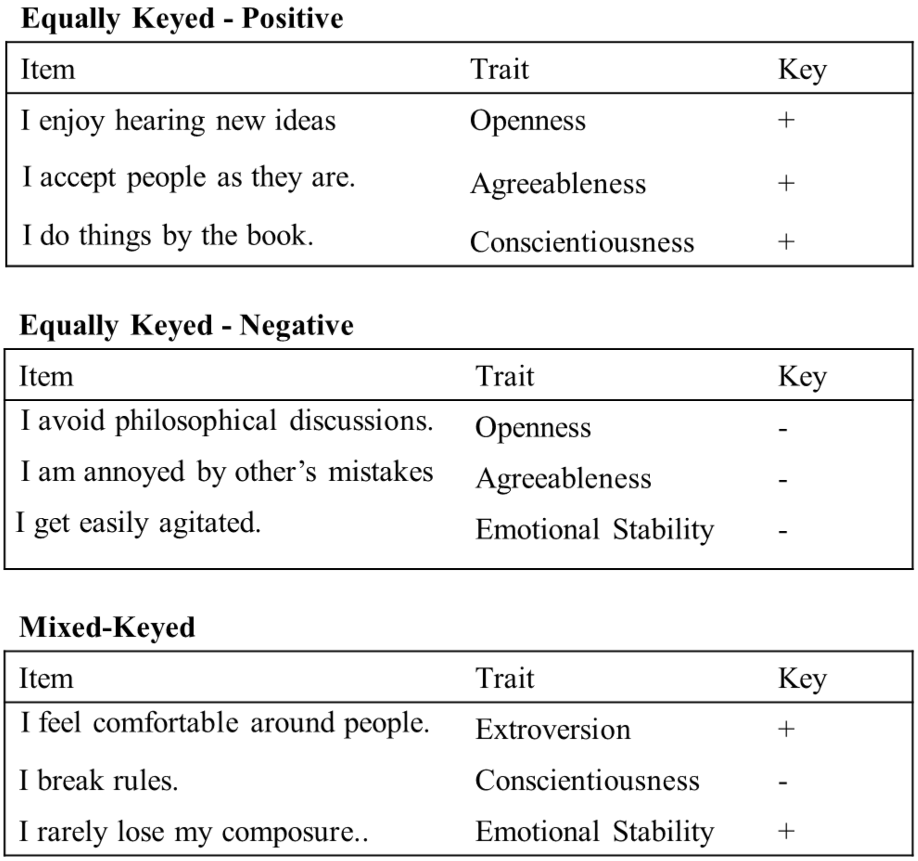
What is desirable depends on the instructions given and context a person is responding in. A common study design in FC research is to conduct a 'fake-good' survey and compare it to an honest response style (Christiansen et al., 2005; Heggestad et al., 2006; Li et al., 2024; Stark et al., 2005, to name a few examples). In the 'fake-good' design, respondents are instructed to rate scale items on the measure as if they were trying to make themselves look like the 'ideal' candidate. In the ‘honest’ condition, they are asked to respond as they normally would. Using this design Pavlov (2024) found substantial differences in item desirability, which led to differences in how blocks would be constructed. Converse et al. (2010) examined how desirability ratings differ across contexts for items from the International Personality Item Pool (Goldberg et al., 2006) by comparing general desirability, a job general applicant, a real estate job, and a police job. They found significant differences in the desirability of items across conditions. This research suggests that using context specific ratings in the development of an FC measure, as opposed to general desirability is best practice.

**Data Collection for Block Construction.** Before moving on to constructing blocks it is necessary to gather rating scale responses and desirability ratings to the test items (Lin et al., 2024). The rating scale responses should be subjected to factor analyses to determine that items are properly loading on to their respective factor. The parameters from this analysis can also be used to check the reliability of the FC form (discussed in Phase Two). Finally, these factor analyses can provide internal structure evidence for the test (AERA et al., 2014). If the model fits well and factor loadings are high, then there is enough evidence to proceed to block construction.

### ***Block Construction***

Block construction begins with two key considerations: block format (which determines how many items are in a block) and how to pair items together in blocks. There are two types of blocks that can be constructed. They may be mixed-keyed or equally keyed. Keying refers to whether an item is positive or negative. Items that are negatively keyed will often be lower in desirability than their positive counterparts (Li et al., 2024). When a block is mixed-keyed there are both positive and negative items in it. Equally keyed blocks will have all negative or all positive items. Fig. 4 shows examples of these blocks formed using International Personality Item Pool items (Goldberg et al., 2006). A good block will have no clear ‘correct’ choice, meaning there is little chance of social desirability bias. Ideally, equally keyed blocks would be the only ones used, but issues in current modeling techniques, discussed in phase two, result in both types being used in practice.

Fig. 4 - Block Examples



**Equally Keyed - Positive**

| Item | Trait | Key |
|---|---|---|
| I enjoy hearing new ideas | Openness | + |
| I accept people as they are. | Agreeableness | + |
| I do things by the book. | Conscientiousness | + |

**Equally Keyed - Negative**

| Item | Trait | Key |
|---|---|---|
| I avoid philosophical discussions. | Openness | - |
| I am annoyed by other's mistakes | Agreeableness | - |
| I get easily agitated. | Emotional Stability | - |

**Mixed-Keyed**

| Item | Trait | Key |
|---|---|---|
| I feel comfortable around people. | Extroversion | + |
| I break rules. | Conscientiousness | - |
| I rarely lose my composure.. | Emotional Stability | + |

**Block Format.** Blocks typically contain between two (dyad), three (triad), or four (tetrad) items. Hontangas and colleagues (2015) describe the three types of FC formats as ones where the respondents 1) PICK the option most like them in a dyad, 2) choose one item as most like them and the other as least in a triad or tetrad (the MOLE format), and 3) RANK the options from most to least like them, typically in a triad. The PICK and RANK formats are full-ranking formats where every item is ordered, and a response is recorded for each item in the block. MOLE is also known as the partial-ranking format when a tetrad is used because only three out of four items in the block have a response recorded (Cao & Drasgow, 2019). When these ordered responses are then transformed into binary responses for analysis, this results in one missing binary outcome (Brown & Maydeu-Olivares, 2013).

Research suggests that each format can reduce social desirability bias; however, the best option is debated. Historically, it was believed MOLE was the superior format (Hicks, 1970). But, in a recent meta-analysis conducted by Cao and Drasgow (2019), they found that scales using the PICK format were more resistant to faking than those that used the MOLE format and recommended that triad blocks should be used. There are other factors to consider in picking a format. A simulation study conducted by Frick and colleagues (2021) indicated that trait estimate recovery was slightly worse as block size increased. This means scores of respondents may be biased from their true score depending on the size of the blocks. Cao and Drasgow (2019) found a similar effect but determined the PICK format resulted in the lowest score inflation. These results taken together suggest the use of triads with the PICK format to be the best option.

**Desirability Matching.** Blocks are formed by ensuring no item is clearly the most desirable. There are different approaches for desirability matching in the literature. The first is the mean difference index (Kilmann & Thomas, 1977). In this, the mean value of all responses to the item is calculated. The difference between the mean of that item and other items is calculated systematically, with scores close to zero indicating that they are close in desirability. This provides an absolute difference but does not account for error in responding. Pavlov et al. (2021) have detailed the history of these methods and proposed the use of what they term an inter-item agreement (IIA) coefficient. In their study, Pavlov and colleagues (2021) examined a variety of IIA coefficients and recommended the use of the linearly weighted Brennan-Prediger index ($\mathrm{BP}_l$). The coefficient is based on the level of agreement in desirability ratings from respondents and the error associated with agreement by chance. This results in a value ranging from -1 (complete disagreement) to 1 (complete agreement). For example, if there are three items that participants consistently rate high in desirability, this will result in a positive $\mathrm{BP}_l$ value close to 1, indicating the items are all similarly desirable, making it difficult to make a ‘correct’ choice. If two items are high in rated desirability and one low, this will result in a value around 0 or that is negative, indicating the block has at least one item unmatched with the others. In this case it would likely be clear to someone looking to represent themselves in the most positive light not to choose the lower desirability item, effectively only choosing between the other two. But these indices were only evaluated in the context of dyads. There has been limited research on the use of the indices to form PICK or MOLE formatted blocks, making this an area in need of further research.

A challenge of desirability matching is that a researcher may often have a hundred or more items, which can potentially form thousands of combinations of blocks. To remedy this, Li et al. (2022) introduced the *autoFC* package in R, which automates the process of block construction by using IIA values and user inputs to find optimal block combinations. This package may also assist in finding the balance between equally keyed and mixed-keyed blocks that are currently needed in constructing FC measures.

**Limiting Mixed-Keyed Blocks.** Items can be positive or negative in blocks, with them ideally being all positive within a block to limit desirable responding. However, to accurately estimate scores, there is a need for items to load differently on the trait and for factor loadings within a block to be different from each other (Brown, 2016). To do this a user must carefully assemble blocks using factor loadings calculated in the factor analysis of the rating scale items or by including some mixed-keyed blocks on the test as negative and positive items tend to have different factor loadings (Brown & Maydeu-Olivares, 2011). While the use of factor loadings to assemble a test without any mixed-keyed blocks is possible in theory (Brown, 2016), a formal approach has not been discussed. As such, most research has focused on the use of mixed-keyed blocks and how to limit their usage. Frick et al. (2021) conducted a simulation examining a variety of factors, including various levels of mixed-keyed comparisons. They found that trait-recovery was best when there were 50% mixed-keyed blocks. This balanced recovering the sums of traits and differences between the traits with substantial improvement over having no mixed-keyed blocks but minimal gains when using more than 50%. Lee and colleagues (2022) conducted a simulation that examined the effect of mixed-keyed blocks and found that accurate trait estimation could be accomplished using only 20% mixed-keyed blocks, likewise a recent empirical study provided further evidence that only 20-30% of the test needs to be mixed-keyed (Li et al., 2024). Li et al. also advised that 30% should be the maximum number of mixed-keyed blocks used to limit the opportunity for social desirability bias, but it is most likely not a one-size-fits-all solution.

### ***Response Process Evidence for Forced-Choice Blocks***

Item development and block construction requires making many assumptions about how respondents interpret items. The *Standards* recommends evidence of responses processes in general and FC measures should be no exception. Though research on the response processes involved in FC blocks is limited, this source of validity evidence can offer invaluable insights. Sass and colleagues (2020) investigated respondent reactions to FC tests more broadly. They found evidence for equal motivation to respond to FC and rating scale tests. They also found that blocks containing three items were easier for respondents to answer than ones with four or five. Fuechtenhans and Brown’s (2022) qualitative investigation found that respondents followed a process "Activate-Rank-Edit-Submit" (ARES). In the ARES process, respondents either activate their experience and self-image to produce an accurate ranking or activate their image of an ideal candidate, which will produce a faked ranking. They then may evaluate their ranking and edit it to produce an ideal ranking before submitting their response. They found that the decision to edit is typically motivated by the participant's desire to avoid looking bad. In unequally keyed blocks, there is a much greater chance that responses will be based on an image of the ideal candidate instead of themselves. The investigation of responses process and respondent reactions to FC is an area in need of further research but offers great potential for providing validity evidence for these measures.

### Phase Two: Internal Structure Evaluation

When the FC measure has items grouped into blocks, item response data can be collected to evaluate the psychometric properties of the instrument. There are a variety of models that have been introduced to examine the internal structure of FC measures (see Brown (2016) for a review). We summarize and extend their discussion about the models for forced-choice to include ones published after their paper, summarized in Table 1. Each model is grounded in a theoretical decision model. Thurstone’s Law of Comparative Judgment states that when choosing between two items, a respondent will select the item that is most like them (i.e., with more utility; Thurstone, 1928). The second decision model, the unfolding preference model, is a special case of Thurstone's law and was first introduced by Coombs (1950). In this model respondents will choose the item with a utility closest to their true trait score. Finally, the Bradley-Terry (1952) model assumes the choice of the item is proportional to the comparisons between two items and their true rating on the utility.

The decision models described above indicate how a respondent selects an item within a block based on its utility. Their associated measurement models provide a link between the item utility and underlying trait. There are two types of measurement models. The first is ideal-point where the underlying trait score is measured through a series of pairwise comparisons. In this model the respondent will pick the item with a utility closest to their true trait score. The second type of measurement model is linear factor analysis, also known as the dominance model. In this model the underlying trait scores are measured by item means among other attributes which suggest that items with the highest utility are picked most frequently (Brown & Maydeu-Olivares, 2011). The models, especially those since Stark and colleagues introduced the Multi-Unidimensional Pairwise Preference Model (MUPP; 2005), are based in Item Response Theory (IRT). IRT is a factor analytic approach used to model ordinal data, where the relationship between latent traits and their observable outcomes with a focus on item-specific parameters such as the item's difficulty or ability to discriminate between people at different levels of the latent trait (Embretson & Reise, 2000). IRT models assume local independence of the items, meaning that items should not be related after taking into account the factor estimated from the model. FC blocks violate this assumption because the item responses from a block are entirely dependent on one another through the respondent having ranked them against one another. This is the core challenge that many of the modern psychometrics models for FC measures address. It is beyond the scope of this review to discuss every model's specification and the particulars of how they do so, but the basic process is to introduce specifications that account for the violation of local independence and then use dichotomized item responses to estimate the trait scores for respondents. For readers interested in further details, we would recommend reading the cited paper from Table 1 that is associated with their model of interest.

While specification and approach differ from model to model, comparing the models in Table 1, especially those developed after the Thurstonian-IRT (TIRT) model (Brown & Maydeu-Olivares, 2011), users would find that they can all accomplish the task of analyzing FC data with similar results. The major practical differences between the models are the suitability of each to a particular measurement context.

Table 1 - Forced-Choice

Models

| | Model Name | Acronym | Citation | Measurement Model | Decision model |
|---|---|---|---|---|---|
| 1. | Zinnes-Griggs' Unfolding Preference Model | | Zinnes & Griggs, 1974 | LFA | Thurstonian |
| 2. | Simple Squared Difference Model for Pairwise Preferences | SSDMPP | Andrich, 1989 | IP | Bradley-Terry |
| 3. | Simple Hyperbolic Cosine Model | SHCMPP | Andrich, 1995 | IP | Bradley-Terry |
| 4. | Multi-Unidimensional Pairwise Preference Model | MUPP | Stark et al., 2005 | IP | Bradley-Terry |
| 5. | Bayesian Random Block – IRT | BRB-IRT | Lee & Smith, 2020 | IP | Bradley-Terry |
| 6. | Multi-Unidimensional Pairwise Preference Model – 2PL | MUPP-2PL | Morillo et al., 2016 | LFA | Bradley-Terry |
| 7. | Generalized-Graded Unfolding Model | GGUM - RANK | Hontengas et al., 2016 | IP | Bradley-Terry |
| 8. | Confirmatory Multidimensional Generalized-Graded Unfolding Model | CCGGUM | Wang & Wu, 2016 | IP | Bradley-Terry |
| 9. | Zinnes-Griggs Pairwise Preference Item Response Theory Model | ZG-MUPP | Joo et al., 2023 | IP | Bradley-Terry |
| 10. | Forced-Choice Ranking Models | FCRM's | Hung & Huang, 2022 | IP or LFA | Bradley-Terry |
| 11. | Generalized Thurstonian Unfolding Model | GTUM | Zhang et al. 2023 | LFA | Bradley-Terry |
| 12. | Joint-Response-Time Thurstonian-IRT | JRT-TIRT | Guo et al., 2023 | LFA | Thurstonian |
| 13. | Thurstonian-IRT | TIRT | Brown & Maydeu- | LFA | Thurstonian |

Olivares, 2011

*Note.* IP = Ideal-Point. LFA = Factor Analysis. This table does not describe more advanced applications such as faking-mixture models (see Frick, 2021; Brown & Böckenholt, 2023).

For example, the ideal-point-based MUPP model has been successfully used in constructing a CAT-FC measure (Stark et al., 2014). Other examples, as well as a developing interest in CAT-FC, have been discussed in greater depth by Abad and colleagues (2022). Since that review, Lin and colleagues (2023) have also introduced a version of the dominance-based Thurstonian-IRT model (Brown and Maydeu-Olivares, 2011) that can be used to construct CAT-FC measures. Other models include the ability to incorporate response time into the model (Guo et al., 2023) and using Bayesian estimation approaches (Morillo et al., 2016; Lee & Smith, 2020). Tutorials and software have been introduced to ease the access of using some of these models. This includes R packages to estimate the Thurstonian-IRT model (Bürkner et al., 2019) and generalized-graded-unfolding models (Tu et al., 2021), as well as a tutorial on using *Mplus* to estimate the Thurstonian-IRT model (Brown & Maydeu-Olivares, 2012). Readers may also find interest in model specific reviews such as the one conducted on the Thursonian-IRT model by Jansen and Schulze (2024).

Regardless of the model choice, a set of statistics can be used to examine model fit. The type of fit indices examined will depend on the model. For example, when using the Thurstonian-IRT model it is recommended to examine a corrected RMSEA (Brown & Maydeu-Olivares, 2012). When the RMSEA is low this indicates the model fits well which provides a piece of internal structure evidence.  test developed can examine the RMSEA resulting person score estimates can be used to calculate reliability.

### *Reliability*

The reliability of FC tests is calculated in one of three ways: IRT information-based reliability estimates, simulated true-estimate reliability, and test-retest reliability (Lin, 2022). Information-based IRT estimates are calculated using the standard error measurement (SEM). The SEM is the amount of variance in a respondent's trait score. It is calculated as:

$$SEM(\hat{\eta}_a) = \frac{1}{\sqrt{I_P^a(\eta)}} \qquad (1)$$

Where $\hat{\eta}_a$ is the trait score estimate for trait $a$ and $I_P^a$ is the posterior information of the test for trait $a$. The posterior information is calculated in different ways depending on the model used. When test information is high, indicating a reliable test, SEM will be low and vice versa. The average SEM of respondents is the empirical reliability estimate (see Brown & Maydeu-Olivares, 2011 for an example). This type of reliability is used to assess the reliability of an empirical sample's responses to the FC test. The second form of reliability is termed the simulated true-estimated reliability by Lin (2022). It is calculated as:

$$\rho_T^a = corr(\eta_a, \hat{\eta}_a)^2 \qquad (2)$$

$\rho_T^a$ is the true-estimated reliability of trait $a$. $\eta_a$ is a set of simulated "true" scores, which are often simulated to mirror a population of interest. This could be a set of scores for trait $a$ on a standardized metric (-3 to 3). Using the "true" scores and a set of item parameters, such as the factor loadings from a confirmatory factor analysis on the rating scale form of the item, a set of simulated responses to the FC test can be generated. This simulated data can then be analyzed using the model of choice. Assuming the model is set up for scoring, the user would get $\hat{\eta}_a$ which is a set of estimated scores on the trait. The squared correlation between the true and estimated scores is then taken resulting in $\rho_T^a$. Squaring the correlation gives the amount of variance shared between the true and estimated scores. A strong correlation indicates high reliability. This form of reliability is helpful in determining what level of reliability can be expected when an empirical sample for the FC test is collected. Ensuring that there is adequate simulated true-estimated reliability can help avoid a situation where a test is too short or does not have enough high-quality items before proceeding to further construction. However, it is not usable to assess the reliability of an actual test. Finally, test-retest reliability examines the correlation between the estimated scores of one test administration to another's with the same respondents (see Wetzel & Frick, 2020 for an example).

Lin (2022) investigated the comparability of these reliability estimates and found that they all have pros and cons and are not directly comparable. This is because they are calculated based on different types of measurement error. Information-based IRT estimates measure random error (random differences in item responses) and item-specific error (consistent individual differences in item interpretation; Gnambs, 2015; Lin, 2022). Simulated true-estimate reliability, incorporates scale-specific, random, and item-specific error. Scale-specific error is an error resulting from differences in the measurement’s operationalization (e.g., the error in responses resulting from moving from a rating scale to FC). Finally, test-retest reliability incorporates random and transient errors (difference in the situational factors affect the testing occasion). Lin (2022) found that the differences in these three estimates could exceed .15 units due to the types of error being measured. They made three recommendations based on their findings. These include reporting the type of reliability that was calculated, interpreting it considering its assumptions and limitations, and only making cross test comparison with the same reliability metric.

### ***Measurement Bias***

FC measures are often used in high-stakes settings, making evaluating measurement bias a critical aspect of the development process. Measurement bias can be investigated with tests of measurement invariance (using multiple group factor analysis (MG-FA); Schmitt & Kuljanin, 2008) or differential item functioning (using IRT; Thissen et al., 2012). These analyses aim to identify if the psychometric properties of a test are the same across time, test forms, or groups. The main difference in executing these analyses with an FC instrument from a rating scale is that all of the items in a block must be constrained to equal and items cannot be tested independently. Thus, one is testing for “differential block functioning” as opposed to testing at the item level (Lee et al., 2021). This is a new area of development for FC data and the methodologies are limited.

Recent research demonstrates how the Thurstonian-IRT model can be conducted in a multiple-group fashion (Brown & Maydeu-Olivares, 2018). Similar to the process used with traditional rating scales, a series of increasingly restrictive multiple group models are fit to determine equality of blocks using the differences in model fit indices (Vandenberg & Lance, 2000). Lee and Smith (2020) investigated if the cut-off values for the differences in fit indices used with traditional models can be applied to FC models. They detailed several cutoffs that provided accurate detection of measurement invariance in the Thurstonian-IRT model which are different than those recommended for traditional test formats. Lee et al. (2021) examined a “free-baseline” approach for identifying blocks with DIF in a Thurstonian-IRT model using a simulation study and results were promising. Methods have also been developed for ideal-point models (see Seybert et al., 2014; Joo et al., 2022). These approaches directly compare the item parameters or the item characteristic curves of a focal and reference group to determine if there are significant differences between the two. For all models, there is only a small body of work testing these methodologies and further research is needed to determine the efficacy of these methods in realistic testing scenarios.

## Phase Three: Continued Development, Validation, and Reporting

The continued development of an FC assessment largely follows that of a normal assessment after scores are calculated from one of the models above. These scores can be used to examine relationships with other variables, criterion relationships, and other evidence that the Standards (AERA et al., 2014) indicate are important to form a strong validity argument. For example, after scoring, correlations between traits and other variables can be estimated. The measure under development should have strong correlations with variables that measure similar constructs. Likewise, the measure under development should not correlate with variables that measures dissimilar constructs (Clark & Watson, 2019). These relationships, in addition to potential method bias (e.g., differences between an FC and rating scale version of the test), can be assessed using a multitrait-multimethod matrix (MTMM; Campbell & Fiske, 1959) as can be done with any instrument.

Though the methodologies are similar to those used for rating scales, we highlight some examples of continued validation of real FC assessments. The Enrollment Management Association's (2023) Character Skills Snapshot manual describes their continued testing program and developing a new test form. A new test form may be needed in practice for test security or improvements. Using the same skill definitions, they combined new and old items into new blocks and continued with development as described in previous sections. Additionally, one of the most popular tests used in an occupational setting, the OPQ32r (Brown & Bartram, 2011), provides a rich example of examining how the test predicts management competence, a key criterion that motivated the assessment's development. Another example is the ACT Mosaic which is suggested for program planning (Walton et al., 2022). In their work, Walton and colleagues (2022) investigated the ACT Mosaic's criterion relationship with several important outcomes including GPA and disciplinary infractions as related to the skills measured by the test.

One of the final steps in test development is reporting on all the analyses and procedures taken to support the validity of the test scores (AERA et al., 2014). The tests mentioned above can be used as examples on how to structure a report on the construction and on-going validation of an FC measure. They detail item pilot testing, block construction, and various analyses undertaken to strengthen their validity arguments. Lin (2022) has also offered specific guidelines on the reporting of reliability where they advocate detailing the specific form of reliability used so that different FC tests can be more easily compared.

## Discussion

Whenever a test is used for a high-stakes purpose, faking or cheating is a concern. Forced-choice measurement, proposed as a solution to this problem, has a long history in educational and psychological methods, dating back to the 1940s. This review brought together decades of research as a comprehensive resource for researchers interested in learning about what FC measurement is and how to develop one. In this discussion we turn to on-going debates and the next steps for research.

Debates about the validity of FC scores and the ability of the format to reduce response bias have been going on for decades and continue to this day. The first and most prominent argument is against the ability of an FC measure to reduce response bias when mixed-keyed blocks are included on the test to accurately estimate scores. In a mixed-keyed block, one item is negatively worded and clearly less desirable, making it fakeable (Pavlov et al., 2019; Schulte et al., 2019). Bürkner et al.'s (2019) findings indicate that unless 30 or more traits are measured mixed-keyed blocks are necessary for accurate estimation of item parameters and scores in the Thurstonian-IRT model. Many agree that use of mixed-keyed blocks for accurate estimation of the test is problematic (Bürkner et al, 2019; Lin & Brown, 2017; Ng et al., 2021). However, research suggests that even with the inclusion of mixed-keyed blocks there is evidence that FC tests can reduce faking over and above a rating scale test (Lee & Joo, 2020; Li et al. 2024).

In addition to concerns that the FC format cannot reduce response bias, the general concern that the format is not useful above and beyond traditional rating scales has been present since its inception (see Travers, 1951). Pavlov et al. (2019) reported a lack of support for the FC format's ability to reduce faking in a study where they examined how scores varied for a FC and rating scale test across honest and faking conditions. However, it is notable they did not use a desirability index to create blocks. Recent meta-analyses indicate that the key moderator in FC tests reducing faking is related to how well items are matched on desirability within blocks (see Cao & Draswgow, 2019; Speer et al., 2023). Another avenue is considering the predictive power of FC measures in comparison to traditional rating scales, which has been shown to be superior across multiple studies (Bartram, 2007; Lee et al., 2018; Wetzel & Frick, 2020; Vasilopoulos et al., 2006). The development and research on forced-choice measurement have come a long way in 84 years, but the area is still ripe for inquiry to address these concerns and more.

## Future Directions for Forced-Choice Research

Decades of research has developed methodologies for tackling the longstanding challenges of developing and scoring an FC assessment. For example, the decision processes behind block construction are well detailed (see Cao & Drasgow, 2019; Li et al., 2022; Pavlov et al., 2021). Further, there has been development of psychometric models to address the challenges of ipsative data and make it possible to compare people (see Brown, 2016). While these areas of FC measurement construction have received significant attention there are other areas of validation that remain underdeveloped.

### Response Processes and Bias Testing

The Standards (AERA et al., 2014) recommend examining responses to items as evidence that people are engaging in cognitive processes relevant to the construct. In the context of forced-choice and high-stakes assessment this could also help the community determine if and how people distort their answers. There is a need for further investigation into response processes and what they can reveal to test developers about the interpretation of items and how they are grouped into blocks. First, the "Activate-Rank-Edit-Submit" response process (Fuechtenhans & Brown, 2022) developed on a sample of 19- to 63-year-olds from LinkedIn could be retested to determine its replicability in other samples. How younger participants in educational settings respond to these assessments has yet to be studied (e.g., K-12 students). Additionally, investigating whether think-aloud protocols can be used to understand why groups respond differently, providing qualitative evidence of measurement bias that could be triangulated with factor analytical results. The area of response processes investigation is perhaps the ripest for further inquiry as little is known how respondents interpret and respond to this format and the pros and cons of different ways of creating and grouping items into blocks.

The methods for analyzing block and item bias are also in need of further research. Lee and colleagues (2021) formally investigated an approach for DIF detection in dominance models. Plantz et al. (2024) have followed up with a study using more realistic conditions, showing the method can be very sensitive model misspecification, which is likely in practice. There has also been a lack of research on DIF testing in ideal-point FC models. Without an established, well-researched method for conducting measurement bias investigations, some items may be unfairly providing an advantage to one group over another.

### Test Design

The current approach to assemble FC tests is to use mixed-keyed blocks. This allows for accurate estimation of parameters in scores. However, Brown (2016) has highlighted that factor loadings only need be different within blocks and across the trait to accurately estimate scores, implying an assessment can contain all equally keyed positive blocks. While designing items to have different loadings/discriminations is possible, it has not been formally investigated in the literature. This approach could reduce potential response bias on FC tests further by eliminating the use of mixed-keyed blocks, which remain a key point in debates about the utility of FC assessments.

## Conclusion

We have reviewed the historical context and the rapidly advancing methodologies for developing FC measures. As the methodologies mature and best practices are established, a standardized procedure for construction and validation should be formalized. We are beginning to see progress to this end with Li and colleagues (2024) describing a process focusing on block development. Such work will need to continue if FC measures are to be used in high stakes settings like their cognitive counterparts.

## References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association. https://play.google.com/store/books/details?id=clI_mAEACAAJ

Andrich, D. (1989).  *Applied Psychological Measurement*, *13*(2), 193–216. https://doi.org/10.1177/014662168901300211

Andrich, D. (1995). Hyperbolic  for  and  *Applied Psychological Measurement*, *19*(3), 269–290. https://doi.org/10.1177/014662169501900306

Baier, D. E. (1951). Reply to Travers' "A critical review of the validity and rationale of the forced-choice technique." *Psychological Bulletin*, *48*(5), 421–434. https://doi.org/10.1037/h0059746

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. Guilford Publications. https://play.google.com/store/books/details?id=caxCDwAAQBAJ

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56. https://doi.org/10.1111/j.2044-8325.1996.tb00599.x

Bartlett, C. J., Quay, L. C., & Wrightsman, L. S. (1960).  of  of : Likert- and  *Educational and Psychological Measurement*, *20*(4), 699–704. https://doi.org/10.1177/001316446002000405

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x

Bergs, L. P., & Martin, B. (1961). The effect of instructional time interval and social desirability on the validity of a forced-choice anxiety scale. *Journal of Consulting Psychology*, *25*, 528–532. https://doi.org/10.1037/h0042781

Berkowitz, N. H., & Wolkon, G. H. (1964). *Sociometry*, *27*(1), 54–65. https://doi.org/10.2307/2785802

Berkshire, J. R. (1958). Comparisons of *Educational and Psychological Measurement*, *18*(3), 553–561. https://doi.org/10.1177/001316445801800309

Berkshire, J. R., & Highland, R. W. (1953). Forced-choice performance rating? *Personnel Psychology*, *6*(3), 355–378. https://doi.org/10.1111/j.1744-6570.1953.tb01503.x

*Multivariate Behavioral Research*, *6*(1), 63–73. https://doi.org/10.1207/s15327906mbr0601_4

Block, J. (1957). A comparison between ipsative and normative ratings of personality. *Journal of Abnormal Psychology*, *54*(1), 50–54. https://doi.org/10.1037/h0044466

Brogden, H. E. (1954a). A rationale for minimizing distortion in personality questionnaire keys. *Psychometrika*, *19*(2), 141–148. https://doi.org/10.1007/bf02289161

Brogden, H. E. (1954b). A simple proof of a personnel classification theorem. *Psychometrika*, *19*(3), 205–208. https://doi.org/10.1007/BF02289185

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*, 135-160.

Brown, A., & Bartram, D. (2011). *OPQ32r Technical Manual*. https://kar.kent.ac.uk/44780/

Brown, A., & Böckenholt, U. (2022). Intermittent faking of personality profiles in high-stakes assessments: A grade of membership analysis. *Psychological Methods*, *27*(5), 895–916. https://doi.org/10.1037/met0000295

Brown, A., & Maydeu-Olivares, A. (2011).  of  *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. https://doi.org/10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2018). Modelling forced-choice response formats. In *The Wiley Handbook of Psychometric Testing* (pp. 523–569). Wiley. https://doi.org/10.1002/9781118489772.ch18

Bürkner, P.-C., Schulte, N., & Holling, H. (2019).  *Educational and Psychological Measurement*, *79*(5), 827–854. https://doi.org/10.1177/0013164419832063

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *The Journal of Applied Psychology*, *104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Chen, C.-W., Wang, W.-C., Mok, M. M. C., & Scherer, R. (2021).  for  *Frontiers in Psychology*, *12*, 573252. https://doi.org/10.3389/fpsyg.2021.573252

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance*, *22*(2), 105–127. https://doi.org/10.1080/08959280902743303

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005).  for  *Human Performance*, *18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412.

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010).  in  for *Human Performance*, *23*(4), 323–342. https://doi.org/10.1080/08959285.2010.501047

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, *57*(3), 145–158. https://doi.org/10.1037/h0060984

Cronbach, L. J. (1956). Assessment of individual differences. *Annual Review of Psychology*, *7,* 173–196. https://doi.org/10.1146/annurev.ps.07.020156.001133

Cunningham, C. J. (1964).  to https://search.proquest.com/openview/1bce8838dced0584431813a86383d950/1?pq-origsite=gscholar&cbl=18750&diss=y&casa_token=SeviAem4VuEAAAAA:JlHSgTsoP9qZMkYuLpnig3FCpL9_D6sJQ74pXkW0sG_6BZluLtapU2I973m2SxPRz7W-m7iX

Dubeck, J. A., Schuck, S. Z., & Cymbalisty, B. Y. (1971). Falsification of the forced-choice guilt inventory. *Journal of Consulting and Clinical Psychology*, *36*(2), 296. https://doi.org/10.1037/h0030744

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. *108*. https://psycnet.apa.org/fulltext/1958-00464-000.pdf

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Enrollment Management Association. (2023). Summary of the Snapshot research and findings. https://assets-global.website-files.com/62b5ff6837362e413f2bfed4/64e26e4744c35d763506450b_summary-of-the-snapshot-research-and-findings.pdf

Feldman, M. J., & Corah, N. L. (1960). Social desirability and the forced choice method. *Journal of Consulting Psychology*, *24*, 480–482. https://doi.org/10.1037/h0042687

Frick, S., Brown, A., & Wetzel, E. (2021). *Multivariate Behavioral Research*, 1–29. https://doi.org/10.1080/00273171.2021.1938960

in the *Psychometrika*, *87*(2), 773–794. https://doi.org/10.1007/s11336-021-09818-6

Fuechtenhans, M., & Brown, A. (2022). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment*. https://doi.org/10.1111/ijsa.12409

Gertzen, R. W. (1976). *D*iagnostic forced-choice evaluation of police officers. ruor.uottawa.ca. https://ruor.uottawa.ca/bitstream/10393/21570/1/EC55171.PDF

Ghiselli, E. E. (1954). The forced-choice technique in self-description. *Personnel Psychology*, *7*, 201–208. https://doi.org/10.1111/j.1744-6570.1954.tb01593.x

Goodenough, E. (1957). as a *The Journal of Educational Research*, *51*(1), 25–31. https://doi.org/10.1080/00220671.1957.10882433

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *The Journal of Applied Psychology*, *35*(6), 407–412. https://doi.org/10.1037/h0058853

Guo, Z., Wang, D., Cai, Y., & Tu, D. (2023). in *Educational and Psychological Measurement*, 00131644231171193. https://doi.org/10.1177/00131644231171193

Hedberg, R. (1962). More on forced-choice test fakability. *The Journal of Applied Psychology*, *46*(2), 125–127. https://doi.org/10.1037/h0038453

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. *The Journal of Applied Psychology*, *91*(1), 9–24. https://doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*(3), 167–184. https://doi.org/10.1037/h0029780

Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, *213*, 112307.

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015).  and IRT Scoring of  *Applied Psychological Measurement*, *39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Hontangas, P. M., Leenen, I., de la Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, *28*(1), 76–82. https://doi.org/10.7334/psicothema2015.204

Hung, S.-P., & Huang, H.-Y. (2022). Forced-choice ranking models for raters' ranking data. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *47*(5), 603–634. https://doi.org/10.3102/10769986221104207

Jackson, D. J., & Alwin, D. F. (1980).  *Sociological Methods & Research*, *9*(2), 218–238. https://doi.org/10.1177/004912418000900206

Jackson, D. N., Neill, J. A., & Bevan, A. R. (1973). An evaluation of forced-choice and true-false item formats in personality assessment. *Journal of Research in Personality, 7*(1), 21–30. https://www.sciencedirect.com/science/article/pii/0092656673900299

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). *Human Performance, 13*(4), 371–388. https://doi.org/10.1207/S15327043HUP1304_3

Jansen, M. T., & Schulze, R. (2024). Linear factor analytic forced-choice models: Current status and issues. *Educational and Psychological Measurement, 84*(4), 660-690.

Joo, S.-H., Lee, P., & Stark, S. (2022). *Applied Psychological Measurement, 46*(2), 98–115. https://doi.org/10.1177/01466216211066606

Joo, S. H., Lee, P., & Stark, S. (2023). Modeling multidimensional forced choice measures with the Zinnes and Griggs pairwise preference item response theory model. *Multivariate Behavioral Research*, *58*(2), 241-261.

Jurgensen, C. E. (1944). Report on the “Classification Inventory,” a personality test for industrial use. *The Journal of Applied Psychology, 28*(6), 445–460. https://doi.org/10.1037/h0053595

Kay, B. R. (1959). The use of critical incidents in a forced-choice scale. *The Journal of Applied Psychology, 43*(4), 269–270. https://doi.org/10.1037/h0045921

Kilmann, R. H., & Thomas, K. W. (1977). Developing a forced-choice measure of conflict-handling behavior: The “mode” instrument. *Educational and Psychological Measurement, 37*(2), 309–325. https://doi.org/10.1177/001316447703700204

Kirkpatrick, J. J. (1951). Cross-validation of a forced-choice personality inventory. *The Journal of Applied Psychology, 35*(6), 413–417. https://doi.org/10.1037/h0061581

Lee, H., & Smith, W. Z. (2020).  *Educational and Psychological Measurement*, *80*(3), 578–603. https://doi.org/10.1177/0013164419871659

*Organizational Research Methods*, *24*(4), 739–771. https://doi.org/10.1177/1094428120959822

Lee, P., Joo, S.-H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods. *Personality and Individual Differences*, *191*, 111555. https://doi.org/10.1016/j.paid.2022.111555

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*, 229–235. https://doi.org/10.1016/j.paid.2017.11.031

Li, M., Sun, T., & Zhang, B. (2022). autoFC: An R Package for Automatic Item Pairing in Forced-Choice Test Construction. *Applied Psychological Measurement*, *46*(1), 70–72. https://doi.org/10.1177/01466216211051726

Lovell, G. D., & Haner, C. F. (1955). Forced-Choice Applied to College Faculty Rating. *Educational and Psychological Measurement*, *15*(3), 291–304. https://doi.org/10.1177/001316445501500309

Mazzitelli, D., & JR. (1957). A forced-choice approach to the measurement of teacher attitudes [search.proquest.com]. https://search.proquest.com/openview/d1fe3c9b945ec1a060af14a6a54e246b/1?pq-origsite=gscholar&cbl=18750&diss=y&casa_token=zR5vlntTwRkAAAAA:uu737BIKF5WH1abq7yHkNETs8p9YmTshHnGckm1lJVJDBtIDMyXiKxu0WvfTdBNIX1Dhphw7

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005).  *Organizational Research Methods*, *8*(2), 222–248. https://doi.org/10.1177/1094428105275374

Merenda, P. F., & Clarke, W. V. (1963). Forced-Choice vs Free-Response in Personality Assessment. *Psychological Reports*, *13*(1), 159–169. https://doi.org/10.2466/pr0.1963.13.1.159

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). *Applied Psychological Measurement, 40*(7), 500–516. https://doi.org/10.1177/0146621616662226

https://search.proquest.com/openview/41d1bbd657fa95423c49bb79e764c9bb/1?pq-origsite=gscholar&cbl=18750&diss=y&casa_token=PSHx39TzeWEAAAAA:tbNu_gdPh44sspjDWN-8muiuShXQSwA0F2L3EoKpAifoFUNgthgWDmBC9G1ljWbaSFSatn4p

Newman, Howell, & Harris. (1957). Forced choice and other methods for evaluating professional health personnel. *Genetic, Social, and General Psychology Monographs*. https://psycnet.apa.org/journals/mon/71/10/1/?casa_token=7hfS_mzy9kMAAAAA:iyQbN2jfy_X5AAriJYID4vFl839rMQkZ6saIm9n257bzdwEJy9wmtU4Lg8X42t1MttVDh9KfJEM7MRMUP2F8Lg

Ng, V., Lee, P., Ho, M. H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The development and validation of a multidimensional forced-choice format character measure: Testing the Thurstonian IRT approach. *Journal of Personality Assessment*, *103*(2), 224-237.

Osburn, H. G., Lubin, A., Loeffler, J. C., & Tye, V. M. (1954). *Educational and Psychological Measurement*, *14*(2), 407–417. https://doi.org/10.1177/001316445401400222

Pavlov, G. (2024). An investigation of effects of instruction set on item desirability matching. *Personality and Individual Differences, 216*, 112423. https://doi.org/10.1016/j.paid.2023.112423

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and likert scores. *Organizational Research Methods*, *22*(3), 710–739. https://doi.org/10.1177/1094428117753683

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, *183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

Pearson, M., & Powell, J. P. (1979). *Assessment in Higher Education*, *4*(2), 136–139. https://doi.org/10.1080/0260293790040205

Runyon, E. L., & Stromberg, E. L. (1953). *Educational and Psychological Measurement*, *13*(2), 170–178. https://doi.org/10.1177/001316445301300203

*Educational and Psychological Measurement*, *81*(2), 262–289. https://doi.org/10.1177/0013164420934861

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human resource management review*, *18*(4), 210-222.

Schwartz, S. L., & Gekoski, N. (1960). The Supervisory Inventory: A forced choice measure of human relations attitude and technique. *The Journal of Applied Psychology*, *44*(4), 233–236. https://doi.org/10.1037/h0047241

Seybert, J., Stark, S., & Chernyshenko, O. S. (2014). Detecting DIF with ideal point models: A comparison of area and parameter difference methods. *Applied Psychological Measurement*, *38*(2), 151–165. https://doi.org/10.1177/0146621613508306

Sisson, E. D. (1948). The new army rating. *Personnel Psychology*, *1*(3), 365–381. https://doi.org/10.1111/j.1744-6570.1948.tb01316.x

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A *Military Psychology*, *26*(3), 153–164. https://doi.org/10.1037/mil0000044

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4*(1), 39–47. https://doi.org/10.1111/j.1744-6570.1951.tb01459.x

Thissen, D., Steinberg, L., & Wainer, H. (2012). Detection of differential item functioning using the parameters of item response models. In *Differential item functioning* (pp. 67-113). Routledge.

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review*, *35*(3), 175–197. https://doi.org/10.1037/h0072902

Tolle, E. R. (1955). https://search.proquest.com/openview/ab853adad4436095402cb8389ac8db65/1?pq-origsite=gscholar&cbl=18750&diss=y

Travers, R. M. W. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin*, *48*(1), 62–70. https://doi.org/10.1037/h0055263

Tu, N., Zhang, B., Angrave, L., & Sun, T. (2021). bmggum : An R Package for Bayesian Estimation of the Multidimensional Generalized Graded Unfolding Model With Covariates. *Applied Psychological Measurement*, *45*(7–8), 553–555. https://doi.org/10.1177/01466216211040488

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). *Human Performance*, *19*(3), 175–199. https://doi.org/10.1207/s15327043hup1903_1

Villanova, P., Bernardin, H. J., Johnson, D. L., & Dahmus, S. A. (1994). The validity of a measure of job compatibility in the prediction of job performance and turnover of motion picture theater personnel. *Personnel Psychology*, *47*(1), 73–90. https://doi.org/10.1111/j.1744-6570.1994.tb02410.x

Wang, W.-C., & Wu, S.-L. (2016). Confirmatory multidimensional IRT unfolding models for graded-response items. *Applied Psychological Measurement*, *40*(1), 56–72. https://doi.org/10.1177/0146621615602855

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253. https://doi.org/10.1037/pas0000781

Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales: Current state of the research. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *36*(4), 511–515. https://doi.org/10.1027/1015-5759/a000609

Zavala, A. (1965). *Psychological Bulletin*, *63*, 117–124. https://doi.org/10.1037/h0021567

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika*, *39*(3), 327–350. https://doi.org/10.1007/BF02291707

# Supplementary Materials

# From the 1940s to 2020s:
# A review of the current state of forced-choice methodology

## Annotated Bibliography

**Abbreviations:**

FC – Forced-Choice

RS – Rating Scale

TIRT – Thurstonian-IRT

CAT – Computerized Adaptive Test

**Notes:**

A * Indicates articles that we could not locate a manuscript for, only a reference to it in other works.

Generative-AI (OpenAI, 2024) was used to assist in editing some annotations.

**References:**

OpenAI. (2023). *ChatGPT-4o (October 2024 version)* [Large language model]. https://openai.com

## Annotated Bibliography

Arnold, H. J., & Feldman, D. C. (1981). Social desirability response bias in self-report choice situations. *Academy of Management Journal, 24*(2), 377–385. https://doi.org/10.5465/255848\

Arnold and Feldman examine the issue of social desirability response bias in self-report choice situations. They proposed a unique 'inferred objective weight method' for managing this bias, though the method appears complex in its application. The authors then test the method with an internally constructed FC scale. Despite displaying high correlations between the different methods, they do not elaborate on which scale should be used.

Abad, F. J., Kreitchmann, R. S., Sorrel, M. A., Nájera, P., García-Garzón, E., Garrido, L. E., & Jiménez, M. (2022). Building adaptive forced-choice tests "On The Fly" for personality measurement. *Psychologist Papers, 43*(1), 29-35.

The authors provided a summary of the FC response format and note its benefits along with a very clear background section on ipsative data. They then discussed the MUPP model (Stark et al., 2005). The authors highlight its flexibility for CAT testing and how it has been used to create popular tests such as the TAPAS (Stark et al., 2014). Their study then focuses on different algorithms that can be used in FC-CAT test construction. Based on their findings they recommend using a genetic algorithm to build a bank equally-keyed blocks out of pre-calibrated items at the time of testing ('on-the-fly').

Bäckström, M., & Björklund, F. (2023). Why Forced-Choice and Likert Items Provide the Same Information on Personality, Including Social Desirability. *Educational and Psychological Measurement*, 00131644231178721. https://doi.org/10.1177/00131644231178721

The authors argued that FC and RS items provide the same information through an extensive simulation study. Their simulations showed that the multidimensional FC format reproduces single-item personality estimates well when the items are mixed-keyed. Factors other than keying were also manipulated, such as the estimation technique and number of items. These results called into question the advantages of FC when item sets are large if they provided the same information. They also showed that when using ‘evaluatively neutral’ items in an RS format, they should be able to transition smoothly to FC.

Baier, D. E. (1951). Reply to Travers’ “A critical review of the validity and rationale of the forced-choice technique.” *Psychological Bulletin*, *48*(5), 421–434. https://doi.org/10.1037/h0059746

Baier responds to Travers' (1951) critique. Baier acknowledges some points of agreement, but primarily provides a strong rebuttal to Travers' arguments. While this article does not offer significant insights relevant to the current understanding and application of the FC technique, it provides an intriguing perspective on an academic debate.

Bartlett, C. J., Quay, L. C., & Wrightsman, L. S. (1960). A Comparison of Two Methods of Attitude Measurement: Likert-Type and Forced Choice. *Educational and Psychological Measurement*, *20*(4), 699–704. https://doi.org/10.1177/001316446002000405

Bartlett et al. conducted a comparative study between the FC and RS assessment methods, focusing on how they control social desirability bias. After administering two different versions of an attitude measurement and then conducting a teaching session, they noted that the FC scores remained constant, suggesting it was more robust to bias.

However, they also found the FC method exhibited lower split-half reliability, indicating a potential trade-off between bias control and reliability.

Bartlett, C. J. (1960). Factors Affecting Forced-Choice Response. *Personnel Psychology*, *13*(4), 399-406. *Discussed in Zavala, 1965

Bartlett, C. J. (1966). The use of an internal discrimination index in forced-choice scale construction. *Personnel Psychology*, *19*(2), 209–213. https://doi.org/10.1111/j.1744-6570.1966.tb02029.x

Bartlett investigated if item-total score correlations could be used as a discrimination index for item pairings. They found that these correlations could be used in place of a traditional discrimination index when a criterion variable is not available for computing one.

Bartlett, C. J. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *The Journal of Applied Psychology*, *68*(2), 218–226. https://doi.org/10.1037/0021-9010.68.2.218

Bartlett examined halo effects in the context of invalid vs valid halo. Invalid halo effects are those that arise from ratings based on irrelevant behaviors while valid halo effects are those garnered from the general impression of the person. Bartlett attempted to partial out invalid halo effects by adding an error component. They did this for 11 studies over six years and found that the methods was effective at reducing invalid halo effects comparably to FC.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x

Bartram conducted a meta-analysis on FC assessments. They examined the criterion validity coefficients of 29 studies. The author found that there were higher validity coefficients when an FC format is used opposed to a RS. FC assessments were also found to reduce the average correlation between scales. Finally, Bartram found that FC tests appeared to have a greater ability to distinguish between traits. This result is limited due to the ipsative nature of FC data compared to RS tests. The author concluded, based on their results, that while FC assessments have complex measurement properties, they can be advantageous.

Batista, E. E., & Brandenburg, D. C. (1978). The instructor self-evaluation form: Development and validation of an ipsative forced-choice measure of self-perceived faculty performance. *Research in Higher Education*, *9*(4), 319–332. https://doi.org/10.1007/BF00991404

The authors detailed the construction of a teacher self-report form using FC tetrads. They found the scale to have high reliability as shown by intraclass correlation coefficients. They also detailed validity evidence they collected and discussed how the instrument should be used. This is a great paper that incorporates a lot of the best practices an author could use at the time.

Berkowitz, N. H., & Wolkon, G. H. (1964). A forced choice form of the f scale-free of acquiescent response set. *Sociometry*, *27*(1), 54–65. https://doi.org/10.2307/2785802\

The authors identified an issue with a current scale for measuring authoritarianism and attempted to construct an FC version of it to reduce response bias. They provided criterion related validity evidence for this new scale and found that the FC scale is slightly less reliable than the original RS scale. They also concluded that there is evidence of the scale reducing response bias. Finally, they found that there may be some order effects with how the items are placed in the blocks.

Berkshire, J. R., & Highland, R. W. (1953). Forced-choice performance rating? A methodological study. *Personnel Psychology*, *6*(3), 355–378. https://doi.org/10.1111/j.1744-6570.1953.tb01503.x

The authors examined several different options within the FC response style with a focus on its potential for bias reduction, validity, and reliability. They found that, overall, there was a reduction in leniency bias. They also concluded that using tetrad blocks with all positive items is the best option. In constructing a scale, they used RS items first, then used a discrimination index to determine the favorability of the items. There are a lot of interesting things in this article that persist over time.

Berkshire, J. R. (1958). Comparisons of Five Forced-Choice Indices. *Educational and Psychological Measurement*, *18*(3), 553–561. https://doi.org/10.1177/001316445801800309

Berkshire compared five types of desirability indices. They concluded that the Job Importance index is the most representative of the frequency that items are selected by supervisors.

Bergs, L. P., & Martin, B. (1961). The effect of instructional time interval and social desirability on the validity of a forced-choice anxiety scale. *Journal of Consulting Psychology*, *25*, 528–532. https://doi.org/10.1037/h0042781

The authors were interested if time intervals and social desirability variance affected the validity evidence of a test’s scores. They found the the time interval control has

implications for the scale with longer periods of time increasing the predictive validity of the scale.

Bernardin, H. J. (1987). Development and validation of a forced choice scale to measure job-related discomfort among customer service representatives. *Academy of Management Journal*, *30*(1), 162–173. https://doi.org/10.5465/255902

Bernardin described the construction of a job discomfort scale. They used cognitive interviews to identify areas of potential discomfort in the stakeholders and then worked with other experts to develop items in each area. They then formed tetrad blocks, gathered data, and provided evidence for the scale's relationship to other variables.

Bernhadwn, C. S., & Fisher, R. J. (1971). The relationship between personal desirability and endorsement with a forced-choice technique. *Multivariate Behavioral Research*, *6*(1), 63–73. https://doi.org/10.1207/s15327906mbr0601_4

The authors were interested in if personal desirability of FC items were reliable and if the desirability rating was related to their likelihood of endorsing it. They gathered a sample of responses to the EPPS scale at three time points. They found that the desirability ratings were reliable. They also found that respondents were more likely to endorse the item that was more preferable in the FC pairs.

Brogden, H. E. (1954). A rationale for minimizing distortion in personality questionnaire keys. *Psychometrika*, *19*(2), 141–148. https://doi.org/10.1007/bf02289161

Brogden detailed a method for computing the 'distortion' component that is filtered out in FC assessments. The main goal of the study is to show that FC assessments can filter out response bias. Their findings indicated that when this is a goal more distortion is filtered out.

Brooks, K. (1957). The construction and testing of a forced choice scale for measuring speaking achievement. *Speech Monographs*, *24*(1), 65–73. https://doi.org/10.1080/03637755709375198

Brooks constructed an FC scale for measuring speaking achievement based on another speech performance scale which was a numeric scale. They compared the two scales and found them to be roughly equivalent in terms of the strength of the validity evidence. They state a limitation of the work is that the scores on the numerical scale may have been biased based on students being familiar with it. They concluded that the FC scale may have value as its scoring system does not need to be explained unlike the numerical scale.

Brovernam, D. M. (1962). Normative and ipsative measurement in psychology. *Psychological Review*, *69*(4), 295–305.

Normative measurement assumes variance between individuals exists within a single universe or spectrum while ipsative measurement assumes each individual is their own universe. Broverman described different theories in psychology that are related to both types of measurement. They discussed the statistical implications of the different modes of measurement which they noted as not being typically considered up to this point. Broverman recommended a normative-ipsative type model where individuals are considered both uniquely and as a whole population.

Brown, A. (2010). Doing less but getting more: Improving forced-choice measures with Item Response Theory. *Assessment and Development Matters*, *2*(1), 21–25.

Brown analyzed the OPQ32i, a large workplace measurement. The goal of the analysis was to improve the psychometric properties of the ipsative test through IRT. After removing low information items identified through IRT the assessment. They then constructed a new test using triad blocks instead of tetrads to reduce the cognitive load of each block for respondents. This test is known as the OPQ32r. Brown found similar or improved construct and criterion validity estimates from the OPQ32r compared to the OPQ32i and a normative version of the test

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. https://doi.org/10.1007/s11336-014-9434-9

Brown presented nearly all of the currently available models for FC data and placed them in a common framework. They assessed the usefulness of each and provided ample background on ideal point and dominance models. They concluded that of the models available Thurstonian-IRT is the best/easiest to use in most cases. We recommend reading this article in full.

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, *20*(1), 121–148. https://doi.org/10.1177/1094428116668036

The authors examined the effect of response bias on 360-degree feedback. This is a method of assessment where rater feedback is gathered from several different sources. The authors examined the degree to which bias was controlled when the 360-degree feedback came from an FC or RS assessment. They found that the RS test was subject to more response bias which could be reduced by including a method effect in the modeling of the assessment. The FC assessment seemed to perform better on average and increased the inter-rater agreement compared to the RS and bias-controlled RS tests.

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology*, *3*(4), 489–493. https://doi.org/10.1111/j.1754-9434.2010.01277.x

The authors compared ideal point and dominance models. The main conceptual difference here is whether respondents compare themselves to the statement's location (AKA an ideal point) or if the endorsement of an item comes from endorsing an item if its utility (or if the item is more like them) than a certain threshold. The latter calls for a dominance model. The authors argue for more intermediate/middle difficulty items to be used to determine which of these processes is at play. They go on to argue that the estimation of item characteristic curves may be less accurate in ideal point models. They also point out that ideal point models are not invariant to reverse scoring.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

The authors formalize the TIRT model which builds on the normal ogive factor analytic model. It allows for the use of blocks of any size unlike other models before it. We recommend reading this article in full. A short summary will not encapsulate it well enough.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. https://doi.org/10.1037/a0030641

This paper summarized the authors' earlier work (see Brown & Maydeu-Olivares, 2011) and added an additional empirical example. They described the TIRT model again as the

background for this paper. Their description here is more approachable for a general audience.

Bürkner, P.C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827–854. https://doi.org/10.1177/0013164419832063

The authors investigated the effects of equally keyed items on reducing social desirability. Additionally, they implemented a Bayesian FC model in R with their custom thurstonianIRT package. Their simulation study showed evidence that mixed-keyed items are necessary to estimate the TIRT model when there are 30 or fewer traits. They conclude that while accurate estimates can be obtained, doing so with mixed-keyed blocks is in opposition to the point of FC reducing response bias.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *The Journal of Applied Psychology*, *104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

The authors conducted a meta-analysis on several FC instruments to determine if they reduced faking/response bias. They found several interesting results: 1) multidimensional FC was less prone to score inflation than unidimensional models, 2) normative scoring produced lower score inflation than ipsative or partially ipsative scoring, 3) instructions for study design matter can result in lower score inflation depending on them, 4) context of the job matters in how respondents fake their scores, 5) the PICK format is more resistant to faking than the MOLE format.

Campbell, J. T., & Rundquist, E. A. (1950). Scale items for inclusion in forced-choice rating forms. *American Psychologist*, *5*, 280. *Discussed in Nagel, 1954

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

The authors investigated if RS assessments are more prone to response bias than FC assessments and how much weight the claims that FC is inappropriate for personality assessment has. They did this over the course of three studies. In study one they investigated FC in the context of a high-stakes scenario and found that it reduced response bias over and above the RS test. They also found that the responses from both styles were correlated suggesting similar constructs being measured. In the second study they examined the predictive power of FC scores. Christiansen and colleagues found that the response style enhanced predictive and criterion validity compared to the RS format. Finally, study three provided evidence that it is more cognitively demanding to try and fake an FC test.

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance*, *23*(4), 323–342. https://doi.org/10.1080/08959285.2010.501047

The authors investigated how desirability ratings influence the reduction of social desirability bias. They found that the instructions given for providing desirability ratings impacted how desirable the items were. Specifically, desirability of items is context dependent (e.g., answer as a police officer vs as a realtor). Subsequently, an assessment made using item desirability ratings that came from context specific instructions performed better than one without and a RS test.

Cozan, L. W. (1959). Forced choice: Better than other rating methods. *Personnel*, *36*(3), 80-83.

*Discussed in Zavala, 1965.

Cunningham, C. J. (1964). *Measures of leader behavior and their relation to performance levels of county extension agents* [Doctoral dissertation, Ohio State University].

This dissertation centered on the construction of an FC scale for leader-behavior. Cunningham described the current state of the literature in leadership assessment then discusses constructing the scale. Cunningham found that the scale worked better when used for supervisors rating subordinates rather than as a self-assessment.

Cronbach, L. J. (1956). Assessment of individual differences. *Annual Review of Psychology*, *7*(1), 173–196. https://doi.org/10.1146/annurev.ps.07.020156.001133

This article covered assessment of individual differences as a whole but provided a good review of FC testing at the time. Cronbach stated that FC items are more saturated with valid variance. However, FC is more useful when this difference matters as they tend to be lengthy and harder to construct than standard measures of the time.

Deaton, W. L., Glasnapp, D. R., & Poggio, J. P. (1980). Effects of item characteristics on psychometric properties of forced choice scales. *Educational and Psychological Measurement*, *40*(3), 599–610. https://doi.org/10.1177/001316448004000305

Deaton and colleagues tested the effects of item length, direction, and modifiers for items presented in an FC format. They found several noteworthy results. First, when item length or direction was manipulated, the scores varied on the test regardless of the item content. Second, depending on how item content was paired with an item modifier, it changed the mean scores on the test.

Denton, J. C. (1954). Building a forced-choice personality test. *Personnel Psychology*, *7*(4), 449–459. https://doi.org/10.1111/j.1744-6570.1954.tb01045.x

Denton developed an FC personality scale to be used in personnel selection. The work is focused primarily on finding the correct factor structure for the assessment. Notably, the construction of the scale starts with a yes-no inventory which is transitioned into an FC test.

Dubeck, J. A., Schuck, S. Z., & Cymbalisty, B. Y. (1971). Falsification of the forced-choice guilt inventory. *Journal of Consulting and Clinical Psychology*, *36*(2), 296. https://doi.org/10.1037/h0030744

This brief report examined how an FC guilt inventory can be falsified. The authors examined if there was a link between IQ and how likely a respondent was to distort their answers on the FC inventory. They found that higher IQ individuals were more likely to successfully fake socially desirable results. An interesting note with this article is that they cautioned that the scores may be further distorted in high stakes situations.

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian Item Response Theory. *Educational and Psychological Measurement*, *79*(1), 108–128. https://doi.org/10.1177/0013164417752782

The authors compared an RS and FC version of a test. They found that there were weak correlations between the traits in both models suggesting they are capturing unique constructs by virtue of the format. The results also indicated that the FC model provided more information at higher levels of the trait while the RS format provided higher info at lower ends of it.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* Dryden Press.

This is a book on social desirability in personality assessment. For this review we read Chapter 7 – The Forced-Choice Inventory. Edwards provided a brief description on the technique and its pros/cons. The general points they made are echoed throughout the papers discussed so far but they do state that there is the possibility of faking still occurring even with the FC method.

Feldman, M. J., & Corah, N. L. (1960). Social desirability and the forced choice method. *Journal of Consulting Psychology*, *24*, 480–482. https://doi.org/10.1037/h0042687

The authors scrutinized the FC response style's ability to reduce social desirability bias. They did this by constructing a scale and then closely matching items based on how closely the items were in scale using RS pilot data. They stated that if FC assessment was accomplishing its goal, items should be picked equally as often. They found that this was not the case. The authors suggest that a method beyond simple matching based on scale units is necessary. They concluded that even when FC items are closely matched the response format cannot reduce response bias and may in fact increase it.

Felipe, A. (1969). Social desirability tendency and endorsement of items in a forced-choice inventory. *Philippine Journal of Psychology, 2*(2).

Felipe attempted to validate a measure of social desirability in respondents by examining it in relation to an FC scale. They found that their scale, the UP-SD, can help predict how students will respond to a dyad FC measure.

Ferrando, P. J., Anguiano-Carrasco, C., & Chico, E. (2011). The impact of acquiescence on forced-choice responses: A model-based analysis. *Psicológica, 32*(1), 87-105.

The authors introduced a model for determining the impact of response bias on an assessment based on Coombs (1948) unfolding preference model. They described the process of estimating the model by using a two-step normal-ogive approach. They estimated a centroid which is considered the response bias factor and then partial it out of the correlation matrix before estimating a second centroid, considered the content factor, from the residual correlation matrix from the first step.

Frick, S. (2022). Modeling faking in the multidimensional forced-choice format: The faking mixture model. *Psychometrika*, *87*(2), 773–794. https://doi.org/10.1007/s11336-021-09818-6

Frick described several other models that have been used for determining the amount of bias within blocks at an individual level. They then go on to describe their proposed model for determining the fakability of each block and provide evidence for it through a simulation and empirical study. Frick also found that good item matching reduces the fakability of FC blocks. They suggested their model should be used to further investigate potential bias within blocks.

Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*, 1–29. https://doi.org/10.1080/00273171.2021.1938960

The authors conducted a simulation study exploring many features of the T-IRT model. They found that the T-IRT scoring approach was better than partially ipsative scores. Other results indicated that: 1) Trait recovery worsens slightly as block size increases, but increased block size provides more information. 2) Trait recovery accuracy decreases when all items are positively keyed and positively correlated but it is a marginal issue

when using mixed-keyed blocks. 3) There are differences in the strength of the validity evidence between TIRT, partially ipsative scores and the RS format.

Fuechtenhans, M., & Brown, A. (2022). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment*. https://doi.org/10.1111/ijsa.12409

This qualitative study investigated the underlying process mechanisms of faking when responding to MFC personality assessment in a high-stakes context. The authors conducted 32 cognitive interviews and coded participant responses. Based on these findings, the authors proposed a new response process model called the Activate-Rank-Edit-Submit (A-R-E-S) model. In this model, respondents either activate their past experience and self-image to produce a real ranking or their image of an ideal candidate which will produce an 'ideal' but faked ranking. They then determine if their real ranking needs to be edited (producing an ideal ranking) and submit their response.

Gertzen, R. (1976). *Diagnostic forced-choice evaluation of police officers*. [Doctoral dissertation, University of Ottawa (Canada))].

This dissertation detailed the construction of an FC evaluation form for police officers. Gertzen proposed that the FC method be augmented by also including graphic and narrative scales in the assessment as well. They also reported on the validity evidence gathered from examining criterion relationships with the scale and graphic scale assessments.

Ghiselli, E. E. (1954). The forced-choice technique in self-description. *Personnel Psychology*, *7*, 201–208. https://doi.org/10.1111/j.1744-6570.1954.tb01593.x

Ghiselli reviewed the FC literature with a specific focus on personnel selection and developed a measure. Ghiselli found that the FC scale had evidence for its validity and hypothesized it would be comparable to other scales in personnel selection. They also concluded that there may not be a need for elaborate procedures to match items based on their findings.

Goodenough, E. (1957). The forced choice technique as a method for discovering effective teacher personality. *The Journal of Educational Research*, *51*(1), 25–31. https://doi.org/10.1080/00220671.1957.10882433

Goodenough constructed an FC scale that targeted effective teacher personalities. The scale is used for rating of other teachers. They found that the scale had high validity coefficients and concluded that the study provided evidence for the efficacy of the FC method in this type of usage.

Goffin, R. D., Jang, I., & Skinner, E. (2011). Forced-choice and conventional personality assessment: Each may have unique value in pre-employment testing. *Personality and Individual Differences*, *51*(7), 840–844. https://doi.org/10.1016/j.paid.2011.07.012

The authors investigated the claims made in Christiansen et al. (2005) using a sample of employees. They found that the FC scale was strongly correlated with the RS one indicating that similar constructs are being measured. They also found that FC was correlated more strongly with general mental ability, but RS was better at incremental prediction of counterproductive work behavior. They concluded both response styles are potentially useful.

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *The Journal of Applied Psychology*, *35*(6), 407–412. https://doi.org/10.1037/h0058853

Gordon examined the differences between an FC and RS version of an assessment. The author first constructed these new scales then examined their validity coefficients. They found that overall, the FC style assessment was ‘more’ valid that the RS in its assessment of student personality. Gordon also found that there was no added advantage to including both types of scales in terms or predictive accuracy.

Guo, Z., Wang, D., Cai, Y., & Tu, D. (2023). An Item Response Theory Model for Incorporating Response Times in Forced-Choice Measures. *Educational and Psychological Measurement*, 00131644231171193. https://doi.org/10.1177/00131644231171193

The authors conducted a simulation study examining a new FC model that incorporates response time to items in a TIRT model. The simulation study showed that the model obtained better recovery of the latent traits and structural parameters than the TIRT model. The authors highlighted the need for optimized test construction to leverage response time data effectively and suggest that this approach could also be beneficial in detecting aberrant behavior in high-stakes testing scenarios. They concluded that there is a necessity for further research with real data to validate these findings.

Hedberg, R. (1962). More on forced-choice test fakability. *The Journal of Applied Psychology*, *46*(2), 125–127. https://doi.org/10.1037/h0038453.

Hedberg examined the survey of interpersonal values using a FC test format. They used two different instruction sets for the same set of participants after two weeks. They found that the responses of some participants changed significantly for between administrations when the instructional set included instructions that encouraged optimal responding. Hedberg concluded that this indicated the FC format is not without issue.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. *The Journal of Applied Psychology*, *91*(1), 9–24. https://doi.org/10.1037/0021-9010.91.1.9

The authors constructed an FC and RS test and then compared the results to the NEO-FFI, a five-factor personality test. They provided details on their scale construction procedure in Study 1. In Study 2 they instructed participants to respond to the two constructed assessments as job applicants and the NEO-FFI honestly. They found the FC test was more resistant to faking based on lower mean scores than the RS test in a fake-good condition. They also found that that the FC measure did not outperform the RS measure in maintaining consistency with the NEO-FFI. These findings indicate the FC test may not be more fake resistant than a RS test.

Heilbrun, A. B., Jr. (1963). Evidence regarding the equivalence of ipsative and normative personality scales. *Journal of Consulting Psychology*, *27*(2), 152–156. https://doi.org/10.1037/h0047490

Heilbrun expounded upon the current state of different types of measurement in the field of psychological measurement. These included normative, interactive or field studies, and solipsistic measurement (a type based on introspection where participants respond to personality surveys). Heilbrun then classified them in terms of the units they measure:

1. Ipsative measures a population of measurements relative to the person.
2. Normative measures a population of measurements relative to a population of people.

3. Heilbrun also focused a great deal of the article on trying to argue for interactive measurement (measurement based on the individual interacting with their environment). They recommend interactive measurement as the standard for interdisciplinary work.

Heineman, C. E. (1953). A forced-choice form of the Taylor Anxiety Scale. *Journal of Consulting Psychology*, *17*(6), 447–454. https://doi.org/10.1037/h0062337

Heineman constructed an FC version of the Taylor Anxiety scale. As a paper on the construction of an instrument, this stands out for its detailed procedure section and discussing how blocks were formed. Compared to the original scale, Heinman found a reduction in social desirability bias.

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612. https://doi.org/10.1177/0146621615585851

The authors proposed an extension of the MUPP model to handle any block format and examine how they perform using different scoring techniques. They found that EAP scoring works better than traditional scoring (totaling the raw data) in accurately estimating the trait scores. They also found the RANK format performed better than PICK and MOLE in both scoring types. The results of their simulation study also suggested having a larger number of blocks with high discrimination items provided better trait recovery

Hontangas, P. M., Leenen, I., de la Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, *28*(1), 76–82. https://doi.org/10.7334/psicothema2015.204

The authors conducted a simulation study that extended the work of Hontengas et al. (2015) by examining different scoring techniques for MOLE, PICK, and PAIR block formats in a dominance model. They found that that the RANK format performed better across traditional and model-based EAP scoring. They also found that increasing block size improved the accuracy of both estimates. Additionally, when comparing the dominance model and unfolding preference model, the dominance model produced scores closer to the true score and true theta values. It also measured those with extreme values of theta more accurately. Finally, trait recovery was better when mixed-keyed blocks were included on the test.

Huber, C. R., Kuncel, N. R., Huber, K. B., & Boyce, A. S. (2021). Faking and the validity of personality tests: An experimental investigation using modern forced choice measures. *Personnel Assessment and Decisions, 7*(1), 3.

The authors examined different instructional sets for faking and compared the RS format to FC. They found that subtly implying respondents should fake elicited the same pattern of results as explicit faking instructions. They also found results indicating the criterion validity of quasi-ipsative scales was not better than RS. Results also pointed to response distortion having an equivalent effect on both response styles.

Hughes, A. W., Dunlop, P. D., Holtrop, D., & Wee, S. (2021). Spotting the "ideal" personality response. *Journal of Personnel Psychology, 20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

The authors investigated the effect of context on what items are rated as most desirable. They first developed an FC questionnaire then instructed respondents to act as an ideal job candidate and then an ideal candidate for a specific job. This elicited significantly different results in which items were ranked as most desirable. The authors concluded that these results show that item matching is context dependent and should be considered in the matching process. They also found evidence that item matching based on desirability results in respondents having a harder time agreeing upon the most desirable item within a block.

Hung, S. P., & Huang, H.-Y. (2022). Forced-choice ranking models for raters' ranking data. *Journal of Educational and Behavioral Statistics, 47*(5), 603–634. https://doi.org/10.3102/10769986221104207

The authors described the construction of FC ranking models (FCRMs) which are flexible unfolding-preference models that allow for blocks of any size. A core feature of the models is the incorporation of rater leniency and rater errors which may occur when the test is being used to rank others. The authors conducted a set of simulation studies with the models and found that it can accurately recover parameters. We recommend readers check out the full article.

Jackson, D. N., & Minton, H. L. (1963). A forced-choice adjective preference scale for personality assessment. *Psychological Reports*, *12*(2), 515–520. https://doi.org/10.2466/pr0.1963.12.2.515

The authors' goal was to create an adjective preference scale using the FC method and determine if social desirability variance is reduced. They were also interested in if the format is more desirable than others for constructing adjective checklists for personality. They found that desirability bias was reduced which increased content reliability. They concluded that the FC style scales are optimal for these types of checklists.

Jackson, D. N., Neill, J. A., & Bevan, A. R. (1973). An evaluation of forced-choice and true-false item formats in personality assessment. *Journal of Research in Personality*, *7*(1), 21–30. https://www.sciencedirect.com/science/article/pii/0092656673900299

Jackson and colleagues investigated the differences between a personality test that used yes-no vs FC responses using the same items in both assessments. They found the FC form had lower reliability than the yes-no test. They went on to discuss issues in the validity of FC and yes-no assessments, concluding that the non-ipsative nature of yes-no assessments will make them the choice for some time.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: does forced choice offer a solution? *Human Performance*, *13*(4), 371–388. https://doi.org/10.1207/S15327043HUP1304_3

The authors investigated if FC could reduce faking on employment tests compared to an RS test. They did this by using a faking condition where respondents were instructed to take the test as if they were a job applicant and compared it to an unprompted condition. The authors found that there was a greater mean difference between the two conditions in the RS format suggesting it is easier to fake. They also found that in the RS condition criterion correlations were lower in the job applicant condition. The job applicant condition can be considered a high-stakes condition, suggesting the RS is less predictive of important outcomes when socially desirable responding is more probable. This is compared to FC where the criterion relationships were roughly equal across both conditions.

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A Comparison of the Psychometric Properties of the Forced Choice and Likert Scale Versions of a Personality Instrument. *International Journal of Selection and Assessment*, *23*(1), 92-97.

The authors examined how closely related an RS and FC version of the OPQ is psychometrically. They found in a low-stakes setting the assessments are equivalent in terms of their correlational pattern, reliabilities, and individual profiles. They also found evidence for measurement and structural invariance in the low-stakes settings. The authors also found support for the FC format controlling for uniform response biases and suggest that it is a better option for high-stakes conditions than the RS version.

Joo, S.H., Lee, P., & Stark, S. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model. *Journal of Educational Measurement, 55*(3), 357–372. https://doi.org/10.1111/jedm.12183

The authors examined the method proposed by Hontengas et al. (2015). Their simulation study provided evidence about numerous features of the model. This evidence includes that tetrads and triads perform better than dyads. They also found that tetrads provided best results in terms of information, reliability, and scoring. They noted that more work is needed for this model such as an empirical example.

Joo, S.-H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods, 52*(2), 761–772. https://doi.org/10.3758/s13428-019-01274-6

The authors conducted a simulation study examining the use of triad and tetrad blocks within the framework of CAT testing. This study is the first to investigate the use of FC-CAT for blocks containing more than two items. In their simulation study the authors use the GGUM-RANK model. Their results indicated that the CAT version of the simulated test outperformed a nonadaptive test for all block formats tested in parameter recovery and score accuracy.

Joo, S.H., Lee, P., & Stark, S. (2021). Modeling multidimensional forced choice measures with the Zinnes and Griggs pairwise preference item response theory model. *Multivariate Behavioral Research*, 1–21. https://doi.org/10.1080/00273171.2021.1960142

The authors examined the Zinnes-Griggs unfolding preference model, for ideal point measurement and introduced software to model it. They also conducted a simulation study on their updated version of the model. They call this the ZG-MUPP model. In their simulation study they found a sample size of 500 was adequate in modeling a five-trait assessment. The model can also simultaneously estimate item and trait estimates. The authors report a lot of other really exciting and interesting information about their model and if the reader is interested in FC modeling, we recommend reading it in full.

Jurgensen, C. E. (1944). Report on the "Classification Inventory," a personality test for industrial use. *The Journal of Applied Psychology*, *28*(6), 445–460. https://doi.org/10.1037/h0053595

One of the first and only FC assessments in the 1940's. Jungsen provided useful descriptions and insights on the attitudes toward personality testing at the time. Then they described the process of constructing the Classifcation Inventory.

Kay, B. R. (1959). The use of critical incidents in a forced-choice scale. *The Journal of Applied Psychology*, *43*(4), 269–270. https://doi.org/10.1037/h0045921

Kay tried to combine the critical incidents technique with FC. They tested this with the evaluation of foreman but found the critical incidents technique, which requires specific and 'objective' phrasing, is too specific for valid FC blocks.

Kilmann, R. H., & Thomas, K. W. (1977). Developing a forced-choice measure of conflict-handling behavior: The “mode” instrument. *Educational and Psychological Measurement*, *37*(2), 309–325. https://doi.org/10.1177/001316447703700204

The authors constructed a five factor scale for conflict handling behavior and made specific reference to the most used FC scale of the time the EPPS and criticisms levied against it that they consider in the construction of their scale. They attempted to solve for this by ensuring the FC dyads are of equal preference and response frequency. After gathering data they reported poor reliability and validity evidence with the conclusion that the measure had enough evidence to be used but further examinations should be made into its external validity.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *The Journal of Applied Psychology*, *65*(5), 507–516. https://doi.org/10.1037/0021-9010.65.5.507

The authors constructed a multidimensional FC scale to reduce halo effects in performance evaluation. The authors are detailed in how they selected items. This is a good example of how to generate an initial item pool. They then constructed tetrad blocks using discrimination indices to decide on how to pair items. They provided internal structure evidence for their scales by analyzing correlations. They also found evidence for the criterion validity and reliability of their measure. They did not find evidence for the halo effect being reduced.

Kirkpatrick, J. J. (1951). Cross-validation of a forced-choice personality inventory. *The Journal of Applied Psychology*, *35*(6), 413–417. https://doi.org/10.1037/h0061581

Kirkpatrick investigated Jugensen’s (1944) Classification Inventory in a cross-validation study targeted at its efficacy in predicting academic achievement for students. They found that the inventory does not adequately predict academic achievement. The article is historically significant as one of the first examinations of FC in educational settings.

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology, 10*, 2309. https://doi.org/10.3389/fpsyg.2019.02309

The authors examined the effects of using an FC and graded response model on the validity evidence and reliability of scores. They found that the strength of the validity evidence was the worst for the graded response model in the simulation studies that they conducted. Results also indicated that the reliability of the FC model was lower than the graded model.

Kreitchmann, R. S., Sorrel, M. A., & Abad, F. J. (2023). On Bank Assembly and Block Selection in Multidimensional Forced-Choice Adaptive Assessments. *Educational and Psychological Measurement, 83*(2), 294–321. https://doi.org/10.1177/00131644221087986

The authors discuss FC methodology and its intersection with CAT, noting its ability to increase trait estimate precision. They then provided a helpful summary of different CAT bank assembly procedures. Traditional block banks contain a large pool of pre-constructed FC blocks that are chosen as the test goes on. “On the fly” banks assemble blocks during the test from a pool of items and provide a larger search space. This allows for many more suitable blocks to be constructed for each respondent thus improving trait measurement. The authors discuss findings about the on-the-fly method which has been

shown to improve reliability and block overlap rates, reduce ipsativity, and achieve lower average standard errors than the traditional approach.

Krug, R. E. (1958). The effect of specific selection sets on a forced-choice self-description inventory. *The Journal of Applied Psychology*, *42*(2), 89.

Krug investigated the effect of different instructions preceding an FC assessment. They found that response bias is not reduced when a prompt such as 'the company is looking for intelligent people' preceded the test. They also concluded that using a preference index to construct FC blocks is not sufficient.

Krug, R. E., & Northrup, D. (1959). Judgment time for forced-choice adjective pairs. *The Journal of Applied Psychology*, *43*(6), 407–410. https://doi.org/10.1037/h0039975

Krug and Northrup investigated the different effects that FC pairings had on subject response times. They found that unfavorable pairings did not require more time to respond to than favorable ones. They concluded that the resistance to unfavorable alternatives did not appear. They also found that when items are equally preferred in a block, they required more time to respond to.

Lanman, R. W., & Remmers, H. H. (1954). The "Preference" and "Discrimination" Indices in Forced-Choice Scales. *Educational and Psychological Measurement*, *14*(3), 541–551. https://doi.org/10.1177/001316445401400309

The authors described several indices being used at the time to determine how to pair items together into blocks. These included a variety of preference and discrimination indices. They concluded that the favorability index detailed by Berkshire (1953) was the best.

Lee, P., & Joo, S.-H. (2021). A new investigation of fake resistance of a multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions*, *7*(1), 4. https://doi.org/10.25035/pad.2021.01.004

Lee and Joo investigated response bias differences between FC and RS through differential item and test functioning (DIF and DTF) analyses. They found that DIF occurred less frequently in FC blocks. The author also found that when RS items were used to make FC items, they did not always show the same DIF results in both formats. This suggests that relationship the relationships between items and DIF is not invariant across changing response formats. The authors also found lower DTF for the FC response style compared to the RS one. The authors concluded the results could indicate that FC is more resistant to faking than RS.

Lee, P., Joo, S.H., & Stark, S. (2021). Detecting DIF in multidimensional forced choice measures using the Thurstonian item response theory model. *Organizational Research Methods, 24*(4), 739–771. https://doi.org/10.1177/1094428120959822

Lee and colleagues tested the free-baseline and constrained-baseline approach for latent-scoring DIF analysis. Their simulation study found that there were lower Type I error rates and higher power for the free-baseline approach. They then tested a sequential-free-baseline approach for a real assessment. They concluded by detailing how to conduct DIF testing on FC assessments using the sequential-free-baseline approach.

Lee, P., Joo, S.H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods. *Personality and Individual Differences, 191*, 111555. https://doi.org/10.1016/j.paid.2022.111555

The authors investigated how item keying affects FC tests. This is a common limitation cited when considering the TIRT model. They conducted a simulation study and found that 20-40% of blocks need to be negatively keyed to achieve good reliability and validity estimates. They also investigated an empirical example and found that there was better criterion validity for their FC test when using mixed-keyed blocks in the test.

Lee, P., Joo, S.H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*, *43*(3), 226–240. https://doi.org/10.1177/0146621618768294

The authors continued the work of Hontengas et al. (2015) and Joo et al. (2018) with a simulation study examining the GGUM model. Their major findings are that 1) larger sample sizes result in more accurate parameter estimation, 2) 30 blocks across 10 traits is sufficient for model estimation, 3) items need to be highly discriminating for accurate parameter estimation.

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*, 229–235. https://doi.org/10.1016/j.paid.2017.11.031

The authors examined several scoring methods for an FC and an RS approach. They found that all FC scoring methods used (partially ipsative classical test theory, partially ipsative graded response model (IRT), and the model-based scores of TIRT) displayed good convergent and discriminate validity. The results also showed the model-based approach was the best of the three with better criterion validity estimates and slightly better convergent and discriminant validity. The authors also found that the reliability of TIRT scored assessments was lower than the RS tests.

Lee, H., & Smith, W. Z. (2020). Fit indices for measurement invariance tests in the Thurstonian IRT model. *Applied psychological measurement*, *44*(4), 282-295.

The authors examined to what extent current rules of thumb for testing measurement invariance can be used with RANK blocks modeled with TIRT. They found that difference values in CFI, RMSEA, and NCI are effective for assessing measurement invariance at the metric invariance level but needed to be more stringent to test scalar invariance. They proposed more stringent cut-off values for differences in CFI at the metric level (.007) and scalar level (.001).

Lee, H., & Smith, W. Z. (2020). A Bayesian random block item response theory model for forced-choice formats. *Educational and Psychological Measurement, 80*(3), 578–603. https://doi.org/10.1177/0013164419871659

Lee and Smith described a Bayesian approach for modeling FC data. It has the benefit of incorporating priors as is the case of all Bayesian models. The trade-offs are computational time and complexity. The authors describe the specification of the parameters and provide evidence of the model's efficacy through a simulation study.

Lepkowski, J. R. (1963). Development of a forced-choice rating scale for engineer evaluation. *The Journal of Applied Psychology*, *47*(2), 87. https://doi.org/10.1037/h0044908

A two-page article on the construction of a scale for engineer evaluation. In this article they cited using FC as a relatively bias free technique. They based the items of their scale and scoring on those from another test.

Li, M., Sun, T., & Zhang, B. (2022). autoFC: An R package for automatic item pairing in forced-choice test construction. *Applied Psychological Measurement*, *46*(1), 70–72. https://doi.org/10.1177/01466216211051726

The authors described their R package autoFC. The package offers several useful functions for automating item desirability matching. It incorporates options to match items based on features of common FC models such as TIRT requiring a certain number of mixed-keyed blocks.

Lin, Y. (2022). Reliability estimates for IRT-based forced-choice assessment scores. *Organizational Research Methods*, *25*(3), 575–590. https://doi.org/10.1177/1094428121999086

Lin provided background and a thorough analysis of the various ways of assessing reliability in FC models with a focus on the TIRT model. They emphasized the need for improved reliability reporting standards in FC assessment studies. They begin by providing background on test-retest reliability, test information-based reliability, and simulated true-estimated reliability. They point out how each estimate covers different types of error making them not directly comparable to each other. They conclude with the various scenarios each estimate can be used and recommendations to clearly report the type of reliability calculated in FC papers.

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389–414. https://doi.org/10.1177/0013164416646162

The authors focused on how FC and CAT can be merged to get the benefits of both. They do this by working on a central problem that would need to be overcome for FC-CAT to work. This is the parameter invariance assumption. The authors described how this becomes more complex as item parameters may shift based on which items are placed in a block together. They used historical data to compare item parameters in different

contexts. They found that they remained largely stable across context providing support for the technique.

Lin, Y., Brown, A., & Williams, P. (2023). Multidimensional Forced-Choice CAT With Dominance Items: An Empirical Comparison With Optimal Static Testing Under Different Desirability Matching. *Educational and Psychological Measurement*, *83*(2), 322–350. https://doi.org/10.1177/00131644221077637

This article examined a multidimensional FC-CAT using dominance items modeled with the TIRT model. Results from a simulation study are included as theoretical benchmarks, and the effects of adaptive testing and social desirability balancing on measurement precision, score distributions, and candidate perceptions are reported. The authors discussed different optimization methods, recommending the Fisher-information (FI) approach. Overall, their study indicated that adopting FC-CAT with dominance items is possible. This allows the use of the many items-pools developed as dominance items, such as the International Personality Item Pool.

Lingel, H., Bürkner, P.-. C., Melchers, K. G., & Schulte, N. (2022). Measuring personality when stakes are high: Are graded paired comparisons a more reliable alternative to traditional forced-choice methods? In *PsyArXiv*. https://doi.org/10.31234/osf.io/8rt3j

The authors summarized the difference between binary FC questionnaires and graded-paired comparisons (GPC). The main goal of GPCs is to maintain the benefit of reducing response bias but increasing the scores' reliability. A GPC takes the form of placing two items on opposite ends of a rating scale and indicating how much more one item is preferred. The authors then tested 288 conditions to determine the efficacy of the method. The simulation found strong support for the graded response FC format. The authors

concluded that GPCs, in combination with TIRT modeling, can allow for reliable and normative trait estimation under the right conditions.

Lovell, G. D., & Haner, C. F. (1955). Forced-choice applied to college faculty rating. *Educational and Psychological Measurement, 15*(3), 291–304. https://doi.org/10.1177/001316445501500309

The authors described their process for creating a FC scale which largely followed the thinking at the time of using preference and discrimination indices. They also gathered their initial item pool by asking the stakeholders (students) for statements about the worst and best teacher they have had.

Maher, H. (1959). Studies of transparency in forced-choice scales: I. Evidence of transparency. *The Journal of Applied Psychology, 43*(4), 275–278. https://doi.org/10.1037/h0046315

Maher investigated if providing prompting to participants would decrease the effect of response bias reduction in FC scales. When told to get a high score, participants purposefully distorted their answers and scored higher than their previous unprompted responses. Maher recommended further examination on how this can be overcome.

Markey, S. C. (1947). Consistency of descriptive personality phrases in the forced-choice technique. *American Psychologist*, *2*, 310. *Discussed in Nagel, 1954

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*(6), 935–974. https://doi.org/10.1080/00273171.2010.531231

The authors constructed an item-response theory model for FC data based on Thurstone's Law of Comparative Judgement (1927). This law states that a pairwise comparison is made between two items such that the item with more 'utility' or that is more like them

will be chosen. In the case of a dyad this results in the item with more utility being given a 1, and the other a 0. This dichotomous data is ipsative in nature and the authors established a model that incorporates correlated error terms into a normal ogive model for estimation of the data. The article details the math underlying the model and provides several simulation studies to show its efficacy. We recommend reading this article in full as it is technical.

Mazzitelli Jr, D. (1957). *A forced-choice approach to the measurement of teacher attitudes*. [Doctoral dissertation, University of Illinois at Urbana-Champaign].

Mazzitelli described the construction of an FC version of the Minnesota Teacher Attitude Scale for measuring teacher attitudes and provides validity evidence for the measure.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow’s ear: retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8*(2), 222–248. https://doi.org/10.1177/1094428105275374

This paper established the multidimensional unfolding pairwise preference model that extends Coombs (1951) unfolding preference model. The authors conducted a series of simulation studies to show the model’s efficacy and ability to retrieve normative scores from an FC test. This is one of the first steps to modern day models for scoring FC tests by retrieving normative information from them. We recommend interested readers check out the full article.

Merenda, P. F., & Clarke, W. V. (1963). Forced-choice vs free-response in personality assessment. *Psychological Reports*, *13*(1), 159–169. https://doi.org/10.2466/pr0.1963.13.1.159

The authors tested a free-choice test against an FC one. The free-choice test allows respondents to provide words that described them. The authors then constructed an FC scale using 50% of the words from these responses and 50% from another test. The students thought the FC test did not have options that actually described them and was more frustrating than the free-choice test.

Milgram, R. M. (1979). Perception of teacher behavior in gifted and nongifted children. *Journal of Educational Psychology*, *71*(1), 125–128. https://doi.org/10.1037/0022-0663.71.1.125

Milgram constructed an FC scale for their study on what students thought were important qualities in teachers. They did not provide much detail on the scale other than how they constructed the blocks.

Miller, N., & Gekoski, N. (1959). Employee Preference Inventory: A forced-choice measure of employee attitude'. *Engineering & Industrial Psychology, 1*, 83-90. *Discussed in Zavala, 1965

Miller, J. D., Gentile, B., Carter, N. T., Crowe, M., Hoffman, B. J., & Campbell, W. K. (2018). A comparison of the nomological networks associated with forced-choice and Likert formats of the Narcissistic Personality Inventory. *Journal of Personality Assessment*, *100*(3), 259–267. https://doi.org/10.1080/00223891.2017.1310731

The authors examined the Narcissistic Personality Inventory (NPI) in the context of FC and RS tests. They found that the nomological networks are the same in both response styles. They also found slight differences in how discriminant and convergent validity evidence presents in the two response styles.

Mitzel, H. E. (1958). Developing a forced-choice form for the appraisal of student-teacher rapport with pupils. *The Yearbook of the National Council on Measurements Used in Education*, *15*, 66–69.

Mitzel described a five-step process undertaken in creating the scale that involved item generation, pilot testing, and cross validation. They were not able to cross-validate the assessment, however.

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov Chain Monte Carlo estimation. Applied Psychological Measurement, 40(7), 500–516. https://doi.org/10.1177/0146621616662226

The authors expand the MUPP model (Stark et al., 2005) to be a two-parameter logistic model (MUPP-2PL) by using a dominance-based approach. Their model also makes use of a Bayesian estimation approach. They provided simulation results showing the model's ability to accurately recover item and person parameters. They found the results to be similar to those from a TIRT model. However, the authors argue that the MUPP-2PL has additional utility as it requires less mixed-keyed blocks for accurate parameter estimation. They concluded with an empirical study and applied the MUPP-2PL to a personality test.

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Revista de Psicología Del Trabajo Y de Las Organizaciones, 35*(2), 75–83. https://doi.org/10.5093/jwop2019a11

The authors investigated to what extent item parameters remain invariant from an RS format analyzed with a graded response model to the FC format. They found evidence that the discrimination and intercept parameters from both response formats were highly correlated, suggesting the invariance of parameters. The authors also found that that, on average, the parameter estimates from FC were lower than the RS format. These analyses were largely undertaken to show a method of conducting assumption testing for the MUPP-2PL model which assumes items are invariant from RS to FC.

Nagel, J. H. (1954). *The construction and validation of the forced-choice performance rating for pharmaceutical salesmen and analysis of the characteristics contributing to overall competence and efficiency of these men* [Doctoral dissertation, New York University].

Nagel provided an in-depth review of many of the articles discussed up to this point. Their work is focused on constructing an FC performance scale for pharmaceutical salesmen. Nagel uses the techniques of the time to accomplish this goal.

Newman, S. H., & Howell, M. A. (1961). Validity of forced-choice items for obtaining references on physicians. *Psychological Reports, 8*, 367. https://doi.org/10.2466/pr0.1961.8.2.367

In this one-page article about the development of an FC scale, the authors described the pilot testing and validity coefficients for their scale.

Newman, S. H., Howell, M. A., & Harris, F. J. (1957). Forced choice and other methods for evaluating professional health personnel. *Psychological Monographs: General and Applied*, *71*(10), 1.

The authors conducted an extensive examination of how different response styles performed in the assessment of health professional performance. Their major conclusions

were that FC was the best of the four styles used which included a checklist and two RS scales. The FC items were also notably adapted from a different organization and remained valid in the sample selected for this article.

Osburn, H. G., Lubin, A., Loeffler, J. C., & Tye, V. M. (1954). The relative validity of forced choice and single stimulus self description items. *Educational and Psychological Measurement*, *14*(2), 407–417. https://doi.org/10.1177/001316445401400222

The authors used the distortion method discussed by Brogden (1954) to examine a RS test against one with the same item content but structured as an FC assessment with dyads. They found that both types are equivalent to a certain point, but the RS test stopped providing any more information about the test-taker after a certain number of items while FC benefitted from more items. They concluded that there is not a universally superior method based on the validity coefficients and the choice of format should depend on the number of items available.

Pavlov, G. (2024). An investigation of effects of instruction set on item desirability matching. *Personality and Individual Differences, 216*, 112423.

Pavlov investigated the effect of different instruction sets in a fake-good condition where respondents pretend to be an ideal job candidate and an explicit condition where respondents rank the items explicit desirability. They found that the condition significantly impacted the item pairings that were generated to form blocks. Pavlov concludes that explicit instruction should be given in the instructions when gathering desirability ratings.

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and Likert scores. *Organizational Research Methods*, *22*(3), 710–739. https://doi.org/10.1177/1094428117753683

The authors examined a moderation framework for estimating the amount of response bias present in an RS assessment and if FC could mitigate this effect. They found that the FC test was no better at reducing faking or moderating its effect except in high values of explicit faking (e.g., "To what extent did a desire to get the position of Human Resource Assistant lead you to distort your answers?"). Their other results point to issues surrounding how faking may influence scores and skew model estimation when not considered.

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, *183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

The authors described various techniques for doing item pairings for FC blocks after discussing their history. They first demonstrated the mean difference option, but note it has several limitations. They went on to discuss inter-item agreement (IIA; a term coined by them) coefficients that have been used throughout the literature. They conclude that the Brenner-Prediger index is the best option currently.

Pearson, M., & Powell, J. P. (1979). Short reports on assessment. *Assessment in Higher Education*, *4*(2), 136–139. https://doi.org/10.1080/0260293790040205

The authors constructed a FC test of interpersonal skills. They stated that one of the reasons FC was used is to determine if the old free-choice format can be adapted to it for a more efficient, large-scale, administration. They found that the free-choice and FC scales provided the same information about the targeted skills.

Richardson, M. W., & Kuder, G. F. (1933). Making a rating scale that measures. *Personnel Journal, 12,* 36–40. *****Discussed in Tolle, 1955*

Richardson, M. W. (1949). An empirical study of the forced-choice performance report. *American Psychologist*, *4*, 278-279. *****Discussed in Nagel, 1954

Ray, J. J. (1980). The comparative validity of Likert, projective, and forced-choice indices of achievement motivation. *The Journal of Social Psychology*, *111*(1), 63–72. https://doi.org/10.1080/00224545.1980.9924273

Ray compared several RS scales, an FC test, and a projective measure. They did not talk about them again past their inclusion and reporting in a table of correlations. It appears that the FC scale included in their analysis performs at least as well in terms of correlations to the criterion of interest as the RS scales.

Ross, P. F. (1955). *A comparison of two methods of matching in forced-choice rating*. [Doctoral dissertation, Ohio State University].

Ross' dissertation is focused on the methodology surrounding FC at the time and extending it. This included a great review on concepts such as preference indices, FC block/item creation, and validity coefficients. Their work examined how best to create an FC block to limit response bias. They specifically looked at the differences between items within a block based on a face validity index and a preference index. They found no differences between the two.

Rundquist, E. A., Winer, B. J., & Falk, G. H. (1950). Follow-up validation of forced-choice items of the Army Officer Efficiency Report. *Amer. Psychologist*, *5*, 359. *Discussed in Nagel, 1954

Runyon, E. L., & Stromberg, E. L. (1953). A forced choice evaluation form for clinical psychology practicum students. *Educational and Psychological Measurement*, *13*(2), 170–178. https://doi.org/10.1177/001316445301300203

The authors described the construction of a scale for assessing the performance of clinical practicum students. They found that over the process of validation, and cross-validation with a second sample, that the FC scale had evidence for its validity and strong reliability. They contextualized this in getting substantially different ratings from supervisors for subpar and excellent students when using the FC scale compared to other response formats.

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology*, *88*(4), 797–834. https://doi.org/10.1111/joop.12098

The authors conducted a meta-analysis on different FC measurements in occupational settings. Their results indicated that the quasi-ipsative approach produced results with the strongest validity evidence of the four types examined (which also included RS normative, RS ipsative, and normative FC). Other results indicated specific skills as being important predictors of occupational performance.

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*(1), 3–30. https://doi.org/10.1080/1359432X.2012.716198

The authors conducted a meta-analysis of Big-Five personality tests with a focus on examining their performance in different response formats. They looked at a variety of non-cognitive skills as well to determine their efficacy. They found support for FC assessments resulting in better criterion validity than RS formats in their meta-analysis. They also found that some non-cognitive skills (e.g., conscientiousness) perform better in the FC format. Finally, there was evidence that the reliability coefficients of FC tests are similar to those reported in RS.

Saltz, E., Reece, M., & Ager, J. (1962). Studies of Forced-Choice Methodology: Individual Differences in Social Desirability. *Educational and Psychological Measurement*, *22*(2), 365–370. https://doi.org/10.1177/001316446202200209

The authors investigated if the Edwards PPS FC scale reduced response bias. They found that there was evidence of reduced or eliminated social desirability bias at the group level. The authors stated it remained unknown how the items have individual social desirability (or rather how each item is uniquely believed to be socially desirable to an individual). They discussed Gordon's (1951) assumption of a projective principle where the most socially desirable responses to the individual tend to be those most like themselves.

Sass, R., Frick, S., Reips, U.D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, *27*(3), 572–584. https://doi.org/10.1177/1073191118762049

The authors examined the response processes involved with FC assessments. They found that it was largely the same as the those found in the rating scale format, however, there were some additional steps involved in how a respondent answered a question. This was examined through think-aloud interviews. They found that the judgement stage involved further processing of the item compared to the RS format where they weigh items in each block. Their findings also indicated that respondents take two different approaches to responding. They either make preliminary judgements about the items and what is being assessed or weigh the items all together at the end. They also found test motivation across FC and RS formats to be equal.

Schulte, N., Holling, H., & Bürkner, P.C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement*, *81*(2), 262–289. https://doi.org/10.1177/0013164420934861

The authors investigated the effect of an increasing number of traits on the reliability and accuracy of ipsative and model-based scores for FC tests. Their results indicated that the reliability of tests is generally higher when factor loadings are as well. It appeared that model-based scores were equal or better than ipsative ones. The overall trend of the results indicated that model fit improves when the number of traits increased and mixed-keyed blocks were unnecessary at a high number of traits. However, scores could not be accurately estimated with lower dimensionality unless mixed-keyed blocks were incorporated.

Schwartz, S. L., & Gekoski, N. (1960). The Supervisory Inventory: A forced choice measure of human relations attitude and technique. *The Journal of Applied Psychology*, *44*(4), 233–236. https://doi.org/10.1037/h0047241

The authors jumped straight into the construction of a new scale for assessing supervisors after stating that current measures were lacking. They described the process of developing items and blocks then moved on to validity information. They concluded that there is support for the scale overall based on its relationship to other variables.

Schünemann, A. L., & Ziegler, M. (2023). “Use the Force!” Adaptation of Response Formats. *Psychological Test Adaptation and Development*, *4*(1), 218–234. https://doi.org/10.1027/2698-1866/a000044

The authors reformatted the B5PS (Big Five Inventory of Personality in Occupational Situations), into an FC questionnaire. The authors then provided an array of evidence to support the validity of the transformed assessment. Overall, the paper acted as a validation study where the authors concluded there is sufficient evidence to use the assessment for research purposes.

Scott, W. A. (1968). Comparative validities of forced-choice and single-stimulus tests. *Psychological Bulletin*, *70*(4), 231–244. https://doi.org/10.1037/h0026262

Scott believed prior reviews had not been critical enough of the claim that FC had stronger validity evidence than RS tests and aimed to rectify that with their article. To test this they examined three different scales had been reformatted into FC tests from their RS counterparts. They found that the scales showed relatively equivalent reliability and validity arguments. They also found that there was a high correlation between the scale scores.

Seeley, L. C. (1948). Construction of three measures of instructor evaluation. *Technical Report-SD 383–1–5*. *Discussed in Tolle, 1955

Sisson, E. D. (1948). Forced choice? The new army rating. *Personnel Psychology*, *1*(3), 365–381. https://doi.org/10.1111/j.1744-6570.1948.tb01316.x

Sisson discussed the construction of an FC test to be used with army officers and rating their subordinates. Interestingly they point to an article by Rundquist that described steps for creating an FC assessment (we could not find this article). The article marks a major usage of FC as it became the standard in the Army at the time for officer ratings of subordinates.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, *70*(1), 66–77. https://doi.org/10.1093/poq/nfj007

The authors compared the check-all that apply format and FC dyads in online survey research (note, this is just a yes-no style option). They found that the FC method takes much longer. They concluded that respondents are processing the FC questions more deeply as indicated by the length of time and number of questions answered.

Speer, A. B., Wegmeyer, L. J., Tenbrink, A. P., Delacruz, A. Y., Christiansen, N. D., & Salim, R. M. (2023). Comparing forced-choice and single-stimulus personality scores on a level playing field: A meta-analysis of psychometric properties and susceptibility to faking. *The Journal of Applied Psychology*. https://doi.org/10.1037/apl0001099

The authors conducted an extensive meta-analysis comparing RS tests to FC after ‘placing them on an even playing field.’ This means that they only compared matched assessments where the context of the same assessment was a factor. After minimizing the extraneous variance caused by meta-analytic factors such as different assessment contexts, they found that the RS and FC formats overlap quite a bit, especially when developed from the same item pool. In the context of faking, they found that FC assessments exhibited less susceptibility to it than RS assessments in terms of mean score inflation and correlations. The authors concluded that while there are similarities between the two formats, FC assessment appears to provide meaningful benefits of RS.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement*, *29*(3), 184–203. https://doi.org/10.1177/0146621604273988

The authors introduced the Multi-Unidimensional Pairwise-Preference Model (MUPP) mode. Stark et al. first describe the math underlying the model. They then conduct several simulation studies to show parameters estimates can be accurately recovered. We recommend interested readers check out the full article.

Taylor, E. K., Carroll, J. B., & Winer, B. J. (1949). Validity of the Army's officer efficiency report. *Amer. Psychologist*, *4*, 284. *Discussed in Nagel, 1954

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology*, *4*(1), 39–47. https://doi.org/10.1111/j.1744-6570.1951.tb01459.x

The authors examined the differences between graphic ratings and FC. The variance in response patterns is reduced with the use of FC in both research and workplace settings.

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *The Journal of Applied Psychology*, *73*(4), 749–751. https://doi.org/10.1037/0021-9010.73.4.74

Tenopyr examined the effect of scale interdependency on internal consistency. They used the KR-20 coefficient to assess reliability. Tenopyr concluded that construct interpretations based on an FC scale should be made with extreme caution due to lower reliability.

Tolle, E. R. (1955). *A Critical Analysis of the Forced-choice Rating Technique when Used in Rating Elementary Classroom Teachers of a Metropolitan School System*. [Doctoral dissertation, Wayne State University].

Tolle's dissertation presented another great review of the literature at the time and pointed to articles that are not accessible (mostly in early issues of the American Psychologist). Ross developed an FC test to be used in rating elementary school teachers. They used the current techniques at the time to accomplish this goal. It also represents the largest effort of using FC assessment in education to date.

Tolle, E. R., & Murray, W. I. (1958). Forced choice. *The Journal of Educational Research*, *51*(9), 679–685. https://doi.org/10.1080/00220671.1958.10882519

This article described the basic process of constructing an FC scale using the current methodology of the time. This included six steps, which have been largely followed in the prior scale construction papers in the review thus far. They are:

1. Obtaining narrative data on job performance.
2. Analysis of the initial data and development of a list of job performance characteristics.
3. Obtaining administrative and peer rankings of those for whom the scale is to be developed.
4. Use of the characteristics listed to determine basic indices (e.g., a discrimination or preference index to form blocks.
5. Development and administration of an experimental scale.
6. Development of the final scale and scoring method.

Travers, R. M. W. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin*, *48*(1), 62–70. https://doi.org/10.1037/h0055263

Travers' critiqued the FC format and the lack of clarity on how it reduces variance. They were also unhappy with how little was reported in other articles led by authors Richardson and Baier. He closed with a broad generalization that the FC format is not useful.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, *19*(3), 175–199. https://doi.org/10.1207/s15327043hup1903_1

The authors investigated the validity of FC personality tests in relation to the cognitive ability of respondents. They also had the goal of determining how the response format affected criterion validity. Vasilopoulos and colleagues collected data on an FC and RS formatted assessment in a fake-good condition or honest condition. The authors found that when low-cognitive ability respondents try respond in a way that will make them appear more qualified (the fake-good), they have the reverse effect. They also found the FC scores were more predictive of GPA than RS scores, implying better criterion validity of the FC assessment in the study.

Villanova, P., Bernardin, H. J., Johnson, D. L., & Dahmus, S. A. (1994). The validity of a measure of job compatibility in the prediction of job performance and turnover of motion picture theater personnel. *Personnel Psychology*, *47*(1), 73–90. https://doi.org/10.1111/j.1744-6570.1994.tb02410.x

The authors described the construction of an FC measure for movie theatre employees. They developed their scale using the Job Compatability Questionnaire as a foundation.

The authors are detailed in their process and described reducing the number of items, the selection procedure, rationale for using tetrads, and how they formed their blocks. They then presented validity evidence to support the instruments usage.

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment*, *27*(4), 706–718. https://doi.org/10.1177/1073191119843585

The authors compared several scoring methods for FC assessment. These included a model-based approach using TIRT and ipsative scores. They found that the TIRT scores had worse discriminant and convergent validity than ipsative scores. The authors stated that while this may be the case, the use of TIRT scores is advantageous as it allows for factor analyses and between subject comparisons. They cautioned that it should be ensured the discriminant validity of the test is good before using the TIRT model scores.

Wang, Q., Zheng, Y., Liu, K., Cai, Y., Peng, S., & Tu, D. (2023). Item selection methods in multidimensional computerized adaptive testing for forced-choice items using Thurstonian IRT model. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-02037-6

The authors proposed three new item selection methods for multidimensional computerized adaptive testing for FC items (MFC-CAT) based on the TIRT model. The authors suggested an extension of the Kullback–Leibler (KL) based item selection methods from the single-statement MCAT context to the MFC-CAT context. After conducting two simulation studies that compared the extended KL methods against the more common Fisher-index (FI) methods, they found that any of the KL methods

examined in the study outperformed FI. The authors recommended using a KI optimization method when assembling an MFC-CAT.

Waters, L. K., & Wherry, R. J. (1961). Evaluation of two forced-choice response formats. *Personnel Psychology, 14*(3), 285–289. https://doi.org/10.1111/j.1744-6570.1961.tb01234.x

The authors investigated a combination of response formats to determine if it reduced respondent resistance to the survey. This involved combining FC and RS style items into one format. It appeared that this was more acceptable to participants than FC alone.

Waters, L. K., & Wherry, R. J. (1962). The effect of intent to bias on forced-choice indices. *Personnel Psychology*, *15*(2), 207–214. https://doi.org/10.1111/j.1744-6570.1962.tb01862.x

The authors investigated to what extent response bias was correlated with an attractiveness/preference index at varying levels of response bias. They found evidence that this is the case.

Waters, L. K., & Wherry, R. J. (1962). A note on alternative methods of scoring a forced-choice form. *Personnel Psychology*, *15*(3), 315–317. https://doi.org/10.1111/j.1744-6570.1962.tb01626.x

The authors examined three types of FC format questions, a traditional one, one with added RS items, and one with an added graphical scale. They found that the FC scale alone had better validity coefficients.

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253. https://doi.org/10.1037/pas0000781

The authors examined the validity of trait estimates across FC and RS formats. They found evidence for several things in regard to the Big Five measures they used. 1) The constructs remained constant across response style with the exception of one. 2) The FC test had lower reliability. 3) Criterion validities between the two scores were comparable but the FC test had better convergent validity (and worse discriminant).

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156–170. https://doi.org/10.1037/pas0000971

The authors investigated differences between FC with mixed-keyed blocks and only equally-keyed blocks and the RS format. They found that the RS format was more susceptible to faking than the mixed-keyed test. Their results also indicated that the mixed-keyed test was more fakable than the equally-keyed one. Other results from their investigation indicated instructed faking reduced criterion validity in all formats.

Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales: Current state of the research. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *36*(4), 511–515. https://doi.org/10.1027/1015-5759/a000609

The authors provided a brief overview of FC assessment with a focus on the modeling aspects of it.

Wetzel, E., Roberts, B. W., Fraley, R. C., & Brown, A. (2016). Equivalence of Narcissistic Personality Inventory constructs and correlates across scoring approaches and response formats. *Journal of Research in Personality*, *61*, 87-98.

The authors examined to what extent the Narcissistic Personality Inventory (NPI) results from prior studies can be considered accurate when the proper scoring technique for its FC items is used. They compared the mean and model-based scores. They found that across two test formats and scoring approaches the same constructs were being measured. However, the correlations between constructs were better captured when using the model-based scoring approach.

Wherry, R. J. (1959). An evaluative and diagnostic forced-choice rating scale for servicemen. *Personnel Psychology*, *12*, 227–236. https://doi.org/10.1111/j.1744-6570.1959.tb00807.x

Wherry developed a scale to be used in employment situations. They first gathered relevant items from the stakeholders, used preference indices to make the FC test blocks, and then provided validity and reliability evidence for the test.

Xiao, Y., Liu, H., & Li, H. (2017). Integration of the forced-choice questionnaire and the Likert scale: A simulation study. *Frontiers in Psychology*, *8,* 806. https://doi.org/10.3389/fpsyg.2017.00806

The authors investigated a model that combined the TIRT framework with RS response options. Results indicated that doing so allowed for two and three factor models to be sufficiently estimated. They also found that introduction of RS options allowed for less mixed-keyed blocks to be incorporated into the model in their simulation study.

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin*, *63*(2), 117–124. https://doi.org/10.1037/h0021567

Zavala reviewed the state of FC literature at the time. They opened with a discussion about the many areas which the response style is used. They went on to talk about the evidence for validity of FC scales and highlighted that some of the time there is evidence

for better validity coefficiencts with FC scales but it can be conflicting. They then discussed the FC style's ability to reduce response bias. They went on to consider the acceptability of FC to respondents and pointed out that the style is especially unacceptable when negative alternatives are included in the questions. They stated that of different block sizes terads with all positive items performed the best.

Zhang, B., Luo, J., & Li, J. (2023). Moving beyond Likert and Traditional Forced-Choice Scales: A Comprehensive Investigation of the Graded Forced-Choice Format. *Multivariate Behavioral Research*, 1–27. https://doi.org/10.1080/00273171.2023.2235682

The authors compared a graded-response FC model to a rating scale assessment with the same amount of response options. They were asked to respond when there were two, four, or five response options to indicate which was more like them. To investigate the effect of increased response options, they gathered a sizeable empirical sample who responded to the assessment and their preference for the different versions of it. The authors found that there were not any differencea in the desirability of having more options in the FC tests. All FC assessment versions were less susceptible to response styles than their rating scale counterparts. Finally, while preference did not differ, the graded-response format with five options did result in the highest reliability.

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569–590. https://doi.org/10.1177/1094428119836486

The authors investigated the psychometric equivelance of the RS and FC formats. They collected data from two samples of respondents who responded to a non-cognitive measurement called the TAPAS. They found that the reliability of the two response styles displayed similar patterns, however, the RS test was on average higher. The authors also found that the two formats appear to be measuring the same underlying constructs. Additional findings include respondents slightly preferring the RS format over FC. The authors concluded that the measures were mostly equivelant, however, depending on the trait being measured this may change. Both response styles also had high convergent and discriminant validity with FC displaying slightly better discriminant validity.

# Chapter 4. Detecting DIF in Forced-Choice Assessments: A Simulation Study Examining the Effect of Model Misspecifications

## Preface to Chapter 4

Chapter 2 investigated the validity evidence of non-cognitive assessments in the peer-reviewed and grey literature and made several recommendations on how test makers could improve the validity of their assessments. Chapter 2 also described the importance of making strong validity arguments for high-stakes assessments. This led to a thorough review and synthesis of FC methodology in Chapter 3 as it may decrease response bias present in these contexts and increase the validity argument of the assessment.

Chapter 3 examined the FC format with a comprehensive synthesis of 84 years of literature on FC assessments from 1940 to 2024. Our review is the first step toward creating a more structured construction procedure by laying out what is known. We also discussed gaps in current FC methodology in this chapter, the largest being the lack of a reliable method for testing DIF. Meanwhile, in Chapter 2, we found that item bias and DIF were neglected, as well as the consequences of testing. These two results are linked as item bias can cause negative consequences for test takers and contribute to unfair testing. The lack of thoroughly tested methods to identify DIF is a serious limitation of the FC test format that will preclude its use in many high-stakes contexts. Chapter 4 addresses this gap by thoroughly examining a method for testing DIF in FC assessments. Drawing on the latent-scoring approach proposed by Lee and colleagues (2021), we modify and extend their work to explore conditions typical of real-world assessment contexts. Our proposed approach also allows for the separate examination of non-uniform and uniform DIF.

# Chapter 4. Detecting DIF in Forced-Choice Assessments:

# A Simulation Study Examining the Effect of Model Misspecifications

Manuscript: **Plantz, J.W.,** Brown, A., Wright, K.D., Flake, J.K. (accepted in principle). Detecting DIF in forced-choice assessments: A simulation study examining the effect of model misspecifications. *Peer Community In Registered Reports*.

## Abstract

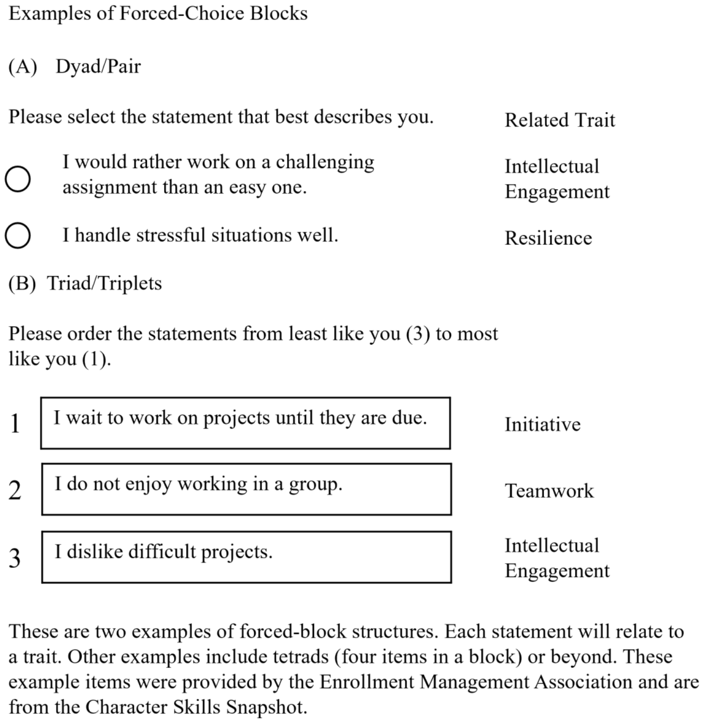
On a forced-choice (FC) questionnaire, the respondent must rank two or more items instead of indicating how much they agree with each of them. Research demonstrates that this format can reduce response bias. However, the data are ipsative, resulting in item scores that are not comparable across individuals. Advances in Item Response Theory have made scoring FC assessments possible, as well as evaluating their psychometric properties. These methodological developments have spurred increased use of FC assessments in applied educational, industrial, and psychological settings. Yet, a reliable method for testing differential item functioning (DIF), necessary for evaluating test bias, has not been established. In 2021, Lee and colleagues examined a latent-variable modeling approach for detecting DIF in forced-choice data and reported promising results. However, their research was focused on conditions where DIF items were known, which is not likely in practice. To build upon their work, we carried out a simulation study to evaluate the impact of model misspecification, using the Thurstonian-IRT model, on DIF detection, i.e., treating DIF items as non-DIF anchors. We manipulated the following factors: sample size, whether the groups being tested for DIF had equal or unequal sample size, the number of traits, DIF effect size, the percentage of items with DIF, the analysis approach, the anchor set size, and the percent of DIF blocks in the anchor. Across 336 simulated conditions, we found that the free-baseline approach performed well and better than the

constrained-baseline approach when there was no misspecification. However, when misspecification was present, Type I error rates greatly increased and power was inconsistent. These results indicate more research is needed to identify a DIF method that can be used to reliably and accurately assess DIF for FC assessments.

## Detecting DIF in Forced-Choice Assessments:

## A Simulation Study Examining the Effects of Model Misspecifications

Forced-choice (FC) is a response format that requires respondents to arrange a set of items - known as a block - in the order that best represents them (see Fig. 1; Cao & Drasgow, 2019; Dueber et al., 2019). This technique was initially conceived to counteract construct-irrelevant variance by limiting respondents' capacity to endorse all desirable items (Bartram, 2007; Lee et al., 2018; Vasilopoulos et al., 2006). It is beneficial in high-stakes or sensitive assessment situations where desirable responding is more likely (Jackson et al., 2000). This includes educational and industrial settings where a test may help inform an admission or hiring decision, as well as the measurement of sensitive topics. For example, the Character Skills Snapshot is used in school admissions (Enrollment Management Association, 2023), the Mosaic (ACT, 2022) is used to inform program planning, and the Occupational Personality Questionnaire which is used in hiring decisions (Brown & Bartram, 2011). While a well-designed FC assessment can effectively reduce response bias, it produces ipsative data. Ipsative data occurs when the responses are directly dependent on each other: e.g., if you rank “I do not enjoy working in a group” first, then you necessarily must rank the other options in Fig. 1. In contrast, a Likert style item allows for



Examples of Forced-Choice Blocks

(A) Dyad/Pair

| Please select the statement that best describes you. | Related Trait |
|---|---|
| ○ I would rather work on a challenging assignment than an easy one. | Intellectual Engagement |
| ○ I handle stressful situations well. | Resilience |

(B) Triad/Triplets

Please order the statements from least like you (3) to most like you (1).

| | | |
|---|---|---|
| 1 | I wait to work on projects until they are due. | Initiative |
| 2 | I do not enjoy working in a group. | Teamwork |
| 3 | I dislike difficult projects. | Intellectual Engagement |

These are two examples of forced-block structures. Each statement will relate to a trait. Other examples include tetrads (four items in a block) or beyond. These example items were provided by the Enrollment Management Association and are from the Character Skills Snapshot.

Fig. 1 - Forced-Choice Block Examples

the selection of any response option, regardless of the response to the last item. Forced-choice assessments produce the same total score for each participant, making interindividual comparisons difficult. This type of data cannot be analyzed with standard methods. Advances in Item Response Theory (IRT) have enabled the evaluation and scoring of FC data, but methodologies for testing differential item functioning (DIF) still need development. Lee and colleagues (2021) introduced a latent variable modeling approach for DIF testing with FC data and evaluated the method using a simulation study. Their study showed promising results when anchor items were correctly specified. We expanded on this work by evaluating the method under conditions of realistic model misspecification when DIF items are included in the anchor set and different testing situations.

This study aimed to further develop DIF testing methods for use in real-world FC testing scenarios. We begin with a review of IRT models and how they have been expanded to account for the complexities of FC data. We then explain the Thurstonian-IRT model and review the literature on methods for testing DIF with FC data to contextualize the current simulation study. Next, we state the research questions we sought to answer with this study and explain the various simulation conditions. Finally, we discuss the results and implications of this work.

## Modeling Forced-Choice Data

Modern forced-choice models are based on IRT. IRT is a class of models used to understand the relationship between item responses and latent traits (Embertson & Reise, 2000). Most of these models assume local independence of items, meaning that after accounting for the factor, the items should be uncorrelated. This is a feature of traditional assessments where item A does not directly influence the score on item B after controlling for the latent trait. In FC assessments, however, the responses to each item within the same block are interdependent (the item-level data are ipsative; Baron, 1996), which also leads to the ipsative trait scores when traditional sum-score scoring approaches are used. With ipsative data, it is not possible to compare the scores of different individuals as each trait score is only comparable within an individual's responses. Brown and Maydeu-Olivares (2011) developed the Thurstonian-IRT model to address these challenges.

## The Thurstionian-IRT Model

The Thurstonian Item Response Theory (TIRT) model is a normal ogive model with some special features. It is conceptually grounded in Thurstone's (1927) law of comparative judgment. According to this law, item *j* is preferred over item *k* in a pairwise comparison ($y_{jk}$) if the latent utility of item *j* ($t_j$) is higher than that of *k* ($t_k$) (Brown & Maydeu-Olivares, 2011). This can be represented as:

$$y_{jk} = 1 \; if \; y_{jk}^{*} = t_j - t_k \geq 0 \quad (1)$$
$$y_{jk} = 0 \; if \; y_{jk}^{*} = t_j - t_k < 0$$

Here, $y_{jk}^{*}$ denotes the difference in latent utilities. When the differences in utilities yields a value below 0, item *k* is preferred over item *j*, and $y_{jk}$ will equal 0.

### *Model Specification*

The TIRT model can be specified as a second-order (Fig. 2) or first-order model (Fig. 3). The first-order model is primarily used for generating factor scores for participants. It involves estimating the thresholds and loadings from the pairwise comparison directly. In the second-order model, the observed responses are functions of the item utilities as described in Equation 1, and the utilities in turn are a linear function of the latent traits as described in Equation 2, where the utility equals the sum of the item mean (μ), the product of the loading (λ) and the measured traits ($\eta$; expressed as a factor score), and an error term (ε) that is independent of the other items.

$$t_j = \mu_j + \lambda_j \eta_j + \varepsilon_j \quad (2)$$

$$t_k = \mu_k + \lambda_k \eta_k + \varepsilon_k \quad (3)$$

Fig. 2 is a diagram for a simple second-order FC model consisting of three blocks with three items in each, measuring three traits. In this diagram, it can be seen that the pairwise item responses are a function of utilities which are then a function of the latent trait. This makes it a second-order model.

In the first-order model, the specification is reformulated in terms of which item will be preferred in a pairwise comparison between the items in equations 2 and 3. The differences between the traits are:

$$y_{jk} = 1 \; if \; y_{jk}^{*} = -\left(\mu_j - \mu_k\right) + \left(\lambda_j \eta_j - \lambda_k \eta_k\right) + \left(\varepsilon_j - \varepsilon_k\right) \geq 0 \quad (4)$$

Furthermore, the difference between item means is often reduced to the threshold parameter:

$$y_{jk} = 1 \; if \; y_{jk}^{*} = -\gamma + \left(\lambda_j \eta_j - \lambda_k \eta_k\right) + \left(\varepsilon_j - \varepsilon_k\right) \geq 0 \quad (5)$$

The first-order model specification can be seen in Fig. 3.

The main difference between the second-order equations in 2/3 and the first-order equation in 5 is that in the first-order model, differences in parameters values are examined for DIF across groups ($\mu_j - \mu_k$) verse just a utility mean ($\mu_{t_j}$) in the second-order model. We focus on the second-order model here because it estimates a single parameter to test for DIF across groups, instead of a difference score.

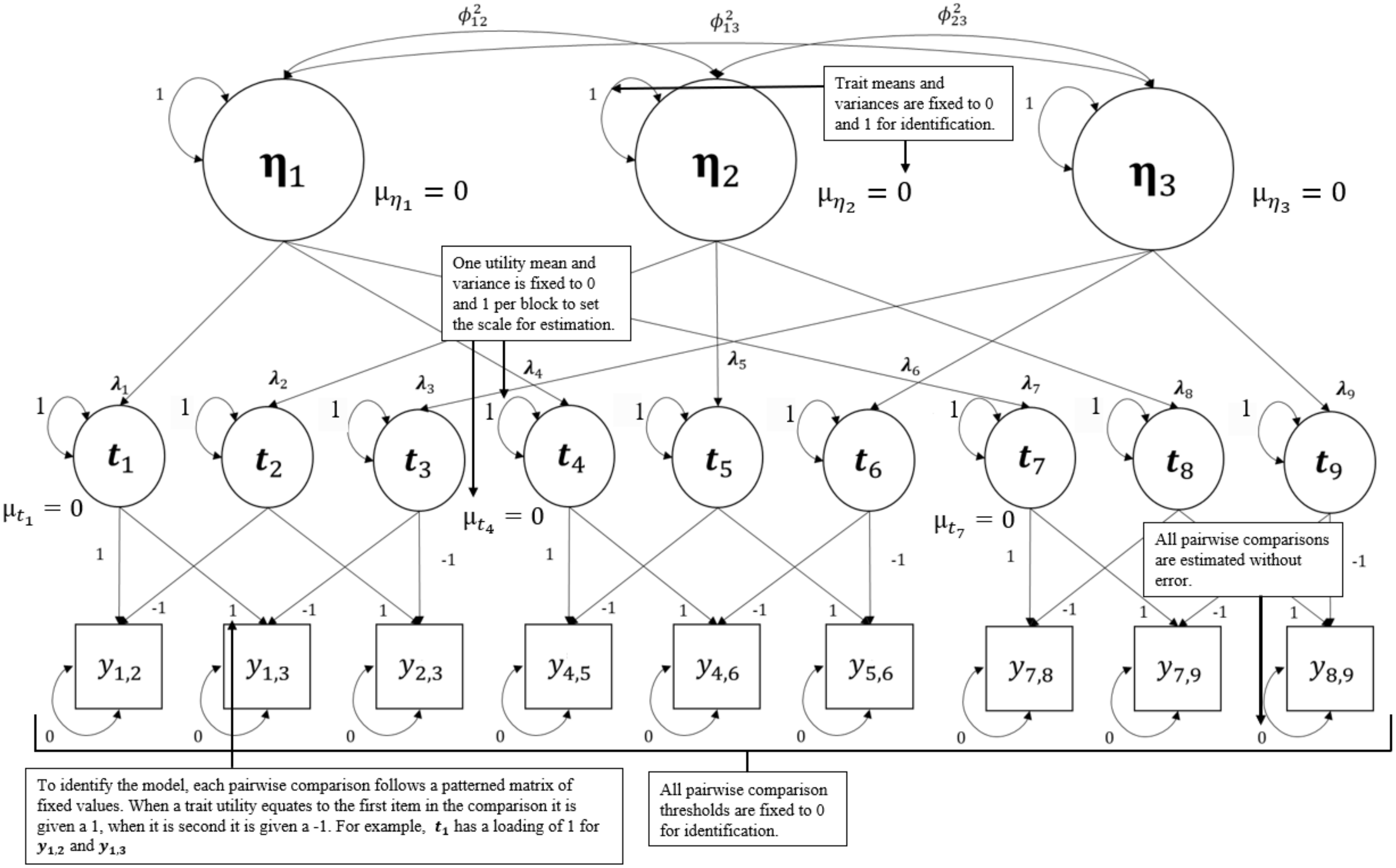
***Model Identification***

For model identification purposes, the loadings of each outcome of pairwise comparison ($y_{jk}$) on the utilities $t_j$ and $t_k$ follow a patterned matrix of fixed values (either 1 or -1) as Fig. 2 shows. Each $y_{jk}$ is estimated without error as Equation 1 shows. Additionally, each trait's variance is set to 1 and their means are set to 0. Also, to set the metric of the utilities, one utility uniqueness per block is fixed to 1. Finally, to identify the scale origin of the latent utilities, one item utility mean per block is set to 0 (for example, the first item in each block) and all thresholds of the pairwise comparisons are explicitly fixed to 0 (Maydeu-Olivares & Brown, 2010). When this model is extended to multiple groups in the context of DIF testing, all constraints mentioned extend to both groups. Additionally, all utility error variances are set to be one in both groups along with a subset of the utility means (*t*). The amount of utility means set to be equal will vary depending on the size of the anchor set.

### *Model Estimation*

The TIRT model is estimated in three stages. In stage one, the thresholds and tetrachoric correlations of pairwise comparisons are calculated. In stage two, the model parameters are estimated using a limited information estimator, such as unweighted least squares (ULS; Brown & Maydeu-Olivares, 2011). The utility means are then estimated such that the mean squared error of the residuals is minimized. The loadings are estimated from the tetrachoric correlations, again minimizing the residuals. Optionally, in stage three, the latent trait scores for respondents are estimated using maximum a posteriori estimation (MAP), which can be implemented in Mplus.

Fig. 2 - Second-Order Thurstonian-IRT Model Diagram



Trait means and variances are fixed to 0 and 1 for identification.

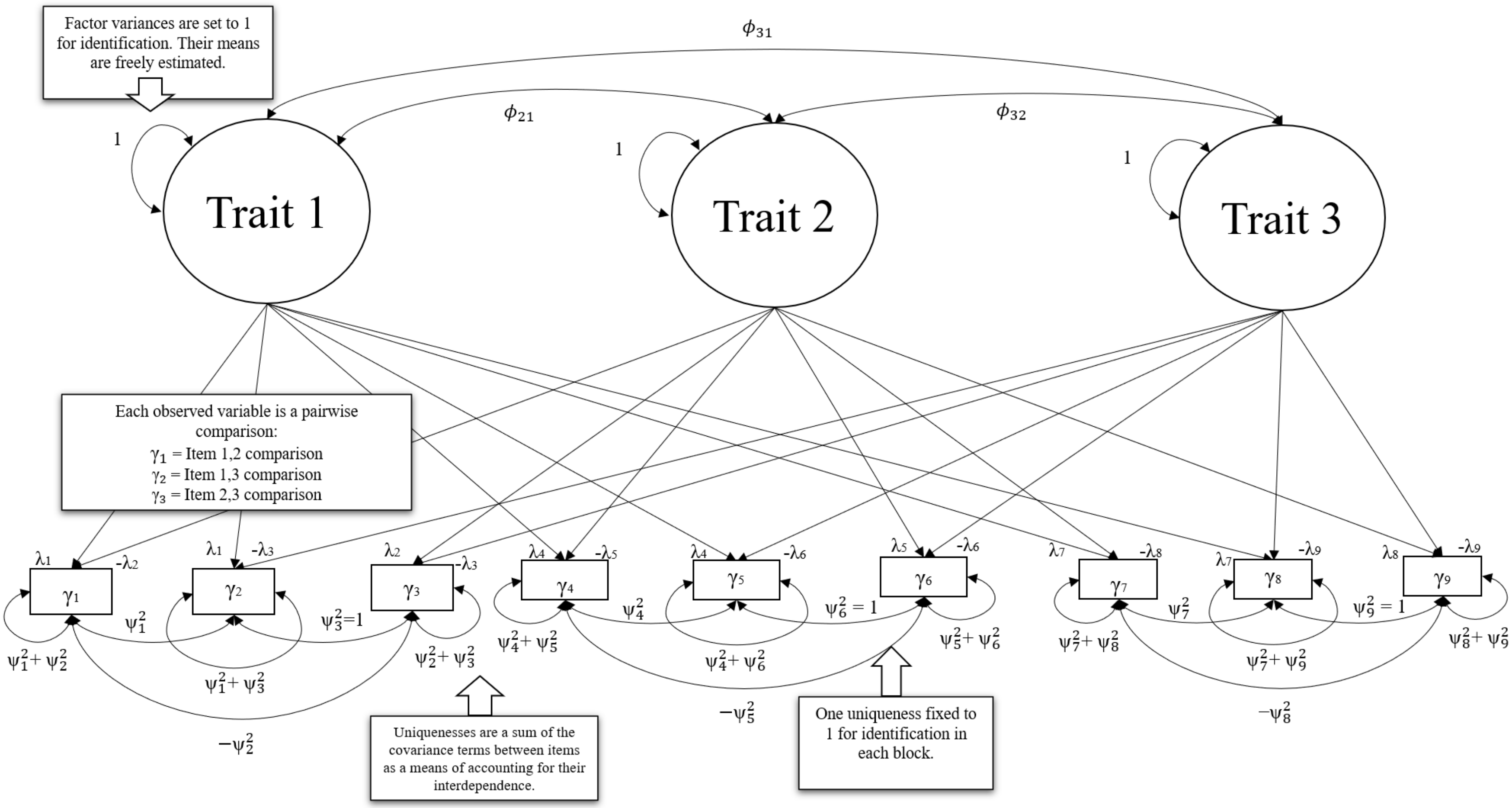
One utility mean and variance is fixed to 0 and 1 per block to set the scale for estimation.

All pairwise comparisons are estimated without error.

To identify the model, each pairwise comparison follows a patterned matrix of fixed values. When a trait utility equates to the first item in the comparison it is given a 1, when it is second it is given a -1. For example, $\boldsymbol{t_1}$ has a loading of 1 for $\boldsymbol{y_{1,2}}$ and $\boldsymbol{y_{1,3}}$

All pairwise comparison thresholds are fixed to 0 for identification.

Fig. 3 - First-Order Thurstonian-IRT Model Diagram



Factor variances are set to 1 for identification. Their means are freely estimated.

Each observed variable is a pairwise comparison:
$\gamma_1$ = Item 1,2 comparison
$\gamma_2$ = Item 1,3 comparison
$\gamma_3$ = Item 2,3 comparison

Uniquenesses are a sum of the covariance terms between items as a means of accounting for their interdependence.

One uniqueness fixed to 1 for identification in each block.

## Differential Item Functioning and Forced-Choice Data

The purpose of an assessment is to provide a valid score of examinees' abilities (AERA et al., 2014). Assessment items should not confer any advantage to specific groups based on factors unrelated to the measured construct, such as gender, race, or cultural background. When a group consistently scores higher than another group, this could be due to true higher ability (i.e., impact) or item bias (Dorans & Holland, 1992). DIF analyses can be used to identify differences between groups, other factors, or continuous data. In this study we examined DIF between groups. In this case DIF screens for potential bias by assessing if items perform differently across groups when their overall ability levels are equal (Angoff, 1993). DIF is classified into two categories: uniform and nonuniform.

Uniform DIF describes a consistent difference in the item's difficulty between two groups at all levels of the measured underlying trait (Swaminathan & Rogers, 1990). This suggests that the item is uniformly more or less difficult for one group compared to the other, irrespective of their standing on the trait. Factors such as biased item content that persistently impacts one group more than the other may cause uniform DIF (Camilli & Shepard, 1994). In the second-order model, uniform DIF is described by differences in the utility means ($t$) for two groups. Non-cognitive assessments do not have correct or incorrect items, but items can still vary in difficulty in that they are harder to endorse, i.e., more of the latent trait is needed to endorse an item. Nonuniform DIF is characterized by differences in group performance on an item only occurring at some ability levels (Swaminathan & Rogers, 1990). This means that the item's slope for different groups varies depending on their trait level. This is each utility's loading ($\lambda$) on the latent trait in the TIRT model.

With FC data, DIF does not occur for a single item in isolation. This is because the items within a block are dependent on one another and responded to in a set, creating multidimensionality. However, most methods assume unidimensionality, such as Mantel-Haenszel, Delta-Plot, standardization, and logistic regression (Angoff, 1972; Dorans & Holland, 1992; Holland & Thayer, 1988; Swaminathan & Rogers, 1990) and thus are not appropriate for FC data. Instead, latent approaches, such as IRT models, that can handle multidimensional items are better suited for FC data.

## Item Response Theory Models for DIF

The process of testing DIF in a latent-scoring approach typically involves specifying an IRT model where loading and threshold parameters are estimated for each group, except for a subset of anchor items constrained to be equal across groups. The remaining free parameters are tested to determine whether they differ significantly across groups. If they do, the item is flagged as displaying DIF. Various strategies for DIF testing have been proposed, including the free-baseline, constrained-baseline, and sequential free-baseline approaches which use model fit and/or Wald tests to determine the significance of differences (Stark et al., 2006).

The free-baseline approach requires constraining a subset of anchor indicators to be equal across groups, which establishes a comparison benchmark. The free-baseline approach is conducted over three steps. First, a set of items are constrained to be equal across groups, the anchor, while all others are freely estimated. Those freely estimated parameters are then tested for DIF using a multiple-constraint Wald test to determine if they are significantly different across groups. The multiple-constraint Wald-Test evaluates whether a set of coefficients in a model are equal, typically specified to test a difference of 0. The test statistic is calculated based on the estimated coefficients and their variance-covariance matrix, with larger values indicating more evidence of DIF. If the Wald test results in a significant difference for an item, it is flagged as a DIF item. The free-baseline approach is advantageous because it requires only a single model to be run (Stark et al., 2006). However, the optimal method for selecting the anchor set is unclear, but there is some evidence the anchor can be small if the anchor is of high quality (Lopez-Rivas et al., 2009).

This is juxtaposed against the constrained-baseline method, which follows a similar, yet opposite process: all items but one are constrained to be equal across groups, and the items are iteratively tested (Wang, 2004). The single freely estimated item is tested for DIF with a Wald test. Then, the next item is tested by constraining all remaining parameters again. This process is repeated until every item has been tested. The constrained-baseline approach performs well when the effect of DIF on the model is not severe (Stark et al., 2006) and allows for every item to be tested, in contrast with the free-baseline method where anchors cannot be tested (Chun, 2016). This approach's disadvantages include the need for multiple model testing and the potential biasing effects of model misspecification (i.e., DIF items are held equal). Finally, there is the sequential free-baseline approach where DIF testing is done in two phases. In phase one, non-DIF items are identified by using the constrained-baseline approach. Then, in phase two, the non-DIF items from phase one are used as anchors in a free-baseline run of the model. Any items from this second phase that display DIF are then flagged (Chun et al., 2016).

To modify these methods for FC data, Lee and colleagues (2021) proposed that when identifying a multi-group first-order TIRT model, entire blocks of items, the thresholds, and loadings, should be constrained to be equal across groups rather than a single item. With the first-order TIRT model it is not possible to parse out uniform vs nonuniform DIF as all parameters within a block need to be constrained. However, when using the second-order model, uniform DIF can be tested by examining constraints on the utility means (*t*) in the block. Nonuniform DIF can be analyzed by examining the loadings (λ). In the constrained-baseline approach, all blocks are constrained to be equal except for one. In the free-baseline approach, only the anchor blocks are constrained, while the remaining blocks are estimated freely in both groups. The free blocks are then tested for DIF by conducting a Wald test with degrees of freedom (*df*) equal to the number of constraints estimated in each block. For example, when testing uniform DIF in a triad block, there are two *df*, one for each freely estimated utility mean in the block. Because the utility means within each block are interconnected and the items are multidimensional, the Wald test examines both freely estimated utility means simultaneously. In Fig. 2, this means $\mu_{t_1}$ would be constrained to 0 for model identification, then $\mu_{t_2}$ and $\mu_{t_3}$would be tested for equality across groups simultaneously.  In this way, the model tests *differential block functioning* across the groups rather than differential item functioning.

## Previous Research on DIF Testing in Forced-Choice

DIF testing for FC assessments is a nascent area of research, for which only a few proposals have been made. Lin and Brown (2017) used a variation of a free-baseline method to detect DIF in two parallel FC test forms, where they estimated measurement parameters freely in each sample, and compared them after performing some scale equating. Wetzel and colleagues (2017) used a variation of the constrained-baseline method to test for DIF in an applied study of narcissism, where they constrained all TIRT measurement parameters to be equal across groups and examined the model modification indices to find items violating these constraints. Neither of these studies, however, had the formal objective to propose a method of DIF detection in FC questionnaires and thus did not systematically examine the merits of either method in various conditions. Lee and colleagues (2021) have proposed a formal method for detecting DIF, which is the free-baseline approach discussed above. In their study, they focused on block-specific factors, such as the type of DIF, the number of DIF items in each block, DIF effect size, presence of impact (differences in the means of the items outside of bias), and the number of DIF blocks on the test. They also examined sample size effects and differences between the constrained-baseline and free-baseline approaches. Overall, their results indicated that DIF blocks were consistently correctly identified (>95% of the time) across all block-specific conditions, except when the sample size was large (N=2000) and had impact (an actual difference in group ability) in the free-baseline approach. In contrast, detecting DIF blocks was far less accurate in the constrained-baseline approach across all conditions.

Lee and colleagues' (2021) work is consistent with other findings that have used the free-baseline and constrained-baseline approaches for standard IRT models. For example, Stark and colleagues (2006) tested the approaches with dichotomous data in the 2PL and polytomous models using the graded response model. Across most simulated DIF conditions, the free-baseline approach was more accurate than the constrained-baseline approach. Several others have found a similar pattern of results in various IRT models (Chun et al., 2016; Kim et al., 2012; Woods & Grimm, 2011). Other research about the approaches has indicated that increasing the number of anchors subsequently results in more accurate DIF detection (Wang, 2004; Wang & Yeh, 2003). However, Wang (2004) found that using DIF items as anchors biased the remaining items toward one group, resulting in less accurate detection rates.

## The Present Study

In the current study, we expanded Lee and colleagues' (2021) work to incorporate test and analysis features that are typical in practice. First, they examined a three and five-factor model, with 30 and 60 items (10 or 20 blocks) total, respectively. However, it is typical for personnel or educational FC assessments to measure more traits (e.g., the 32-trait Occupational Personality Questionnaire, Brown & Bartram, 2009). Second, Lee and colleagues did not test different anchor set sizes or the impact of DIF blocks being included in the anchor set. These model misspecifications are relevant to real-world testing scenarios for which anchor items are unknown. Finally, their simulation used the first-order TIRT model, which does not allow for a direct examination of utility means. Only the mean differences between pairwise comparisons can be analyzed in the first-order TIRT model. It is useful to isolate the means of each utility as this indicates uniform DIF, making the second-order TIRT model more beneficial for DIF analyses. This also allows for uniform DIF and nonuniform DIF to be assessed in an FC model, although we examined only uniform DIF in this study.

First, we tested if the results from conditions investigated by Lee and colleagues replicate when using the second-order TIRT model. Then, we generated different conditions of model misspecification to determine the effectiveness of the free-baseline and constrained-baseline approaches. Our study sought to answer eight research questions. The first five questions (RQ1-5) focus on the replication of Lee and colleagues' work in the second-order model, while the remaining questions (RQ6-8) build upon it.

> RQ1. Do Lee and colleagues’ (2021) findings that Type I error and power rates remain consistent as the number of traits increases replicate in the second-order model?
>
> RQ2. Do the findings of Lee and colleagues (2021), which indicate a pattern where an increase in DIF effect size leads to enhanced power and consistent Type I error rates in both small and large effect size conditions, replicate when examined within the second-order model?

RQ3. Do the findings of Lee et al. (2021), where Type I error rates remained consistent and power increased in increasingly larger sample sizes (500 or 1000 per group), replicate in the second-order model?

RQ4. Does the result of the free-baseline method being more accurate than the constrained-baseline method for DIF detection replicate in the second-order model?

RQ5. Do the findings of Lee and colleagues (2021),  where having a higher percentage of DIF blocks on the test does not affect DIF detection replicate in the second-order model?

RQ6. How does anchor set size influence DIF detection in the free-baseline approach when using the second-order model?

RQ7. To what extent does including DIF items in the anchor set reduce the accuracy of detection in the free-baseline approach when using the second-order model?

RQ8. What effect does unequal sample size have on detecting DIF in the presence of model misspecification?

RQs 1-5 provide needed replication research of methodological research (Loman et al., 2022), as well as an expansion to consider larger numbers of traits that are more consistent with assessment practice. RQ5 considers the effect of the different amounts of DIF for overall DIF detection, consistent with real world testing where the amount of DIF varies across tests. Lee and colleagues (2021) considered the effect of having 10-30% of the test contain DIF blocks. In our study we will examine if the trend they found replicates when the percentage is higher. RQs 6-7 test the free-baseline approach under conditions of varying anchor size (RQ6) and model misspecification (RQ7). This is more representative of real-world contexts where the anchor items are not known prior to conducting the analysis. RQs 6 and 7 do not apply to the constrained-baseline model because in those models all but one block is constrained meaning the anchor set includes all blocks, DIF and non-DIF. RQ8 examined what effect unequal sample size has on DIF detection. For an overview of our study design, the hypotheses related to each question, and the meaning of various outcomes, see Appendix 1.

## Methods

The code to reproduce the simulation can be found in the supplementary materials on this OSF page (https://osf.io/fhd9w/?view_only=cccd50cce05a4dcea4df3a31fe963f2d).

### Conditions

There are several conditions that remained constant across all replications. We describe these first before detailing the manipulated conditions. Each research question, except RQ4, has an accompanying condition that was manipulated. These can be broken down into data generation and analysis conditions.

### *Constant Conditions*

**Analysis and Sample Factors.** The number of replications for each condition is set to 500. All conditions were tested using the TIRT model with Wald tests.

**Assessment Factors.** The block size was set to three items (triads). The number of blocks and items remained constant within each number of traits condition. There were 20 triad blocks (60 items total) in the five-trait condition. In the ten-trait condition, there were 40 triad blocks (120 items total). To increase the ecological validity of the simulation some of the trait correlations were negative and others positive. To further support ecological validity these correlations were based on a meta-analysis of the correlations of the Big Five personality traits (neuroticism, extraversion, openness, agreeableness, and conscientiousness; Linden et al., 2010).

In the case of the Big Five, neuroticism is negatively correlated with the other 4 (-.36, -.17, -.36, -.43). This decision was also based on Frick and colleagues (2021) who found that parameter recovery was better when factor correlations were positive and negative. We used the matrix of correlations reported by Linden and colleagues (2010) in the five-trait condition and followed its pattern in the ten-trait condition. This meant that two traits were negatively correlated with all other traits and positively correlated with each other in the ten-trait condition. This was accomplished by randomly drawing absolute values from an inverse Wishart distribution with 100 degrees of freedom and covariances set to .3 and then making Traits 1 and 6 negatively correlated with the rest of the traits.

### ***Design Features of Blocks and Location of DIF***

We attempted to run the study, but our initial results revealed complexities that were not considered in Lee and colleagues (2021) or our  original design. In this design, we did not consider how DIF was spread across traits and the pattern of DIF was not kept consistent from five to ten traits. This also meant that the blocks being tested would relate to traits with a total number of DIF items that was not consistent. We also did not control for which blocks were being used as anchors in each trait condition. The resulted in some mixed-keyed blocks being used in the anchor for the ten-trait conditions but not in the five-trait conditions. While examining these conditions as manipulated factors will be important in future work, we chose to keep these features constant across conditions in this study.

In the five-trait conditions we simulated five blocks with negatively keyed items and ten in the ten-trait conditions. Including negatively keyed items, expressed as a negative factor loading, is important for accurately estimating the TIRT model (Brown & Maydeu-Olivares, 2011). In practice, the number of negatively keyed items would be kept low to avoid blocks having a clear 'best' (or desirable) answer. It has been shown that only 25% of blocks need to contain a negatively keyed item to accurately estimate scores, and this is the amount we used (Lee et al., 2022). We added DIF to three mixed-keyed blocks in the five-trait conditions and six in the ten-trait conditions. In these blocks, DIF was added to the negatively keyed item for all except one in both trait conditions where DIF was added to a positive item in the block. The number of DIF items on each trait was kept as equivalent as possible in each condition (see Appendix 5 for the spread of DIF across traits on the test). For example, in the 50% blocks with DIF condition, every trait in the five-factor conditions had 2 DIF items. This differs from Lee et al. (2021) who simulated DIF primarily on one trait. This was not plausible in our design due to having a higher percentage of blocks with DIF on the test than they did. Finally, in each anchor condition two or four mixed-keyed blocks were used in the five and ten-trait conditions respectively. All mixed-keyed blocks used in the anchor had DIF added to the negatively keyed-item.

**Block Factors.** We varied DIF only at the utility mean (uniform DIF) level to focus the simulation on misspecification and sample size, while keeping it feasible. We only examined uniform DIF in this study. Previous work suggests that when uniform DIF was present, nonuniform DIF also tends to be detected (Lee et al., 2021). We also chose not to include nonuniform DIF as it would substantially increase the complexity of the study. We only simulated one item per block with DIF. Lee and colleagues (2021) considered multiple items per block with DIF with varied results. This was an area of complexity we chose not to introduce into the study. These conditions are summarized in Table 1.

Table 1 - Constant Simulation Conditions

| Factor Type | Factor | Levels |
|---|---|---|

| | | |
|---|---|---|
| Analysis and Sample Factors | Replications | 500 |
| | Model for estimation | Thurstonian-IRT |
| | Parameter test | Wald Test (2 *df*) |
| Assessment Factors | Number of blocks | 5 Trait = 20<br>10 Trait = 40 |
| | Block size | 3 |
| | Negatively keyed blocks | 5 Trait = 5<br>10 Trait = 10 |
| | Number of mixed-keyed blocks in anchor | 5 Trait = 2<br>10 Trait = 4 |
| | Trait correlations | Positive and Negative Correlations |
| Block Factors | Number of items with DIF in each block | 1 |
| | Type of DIF | Uniform |

### *Manipulated Conditions: Data Generation*

**Sample Size and Equality (RQ3/RQ8).** We simulated data such that there were either 1000 or 2000 total responses, just as Lee and colleagues (2021) did. In the authors' collective experience working on FC assessments operationally, having sample sizes of 1000 or more is typical. We also expanded on Lee and colleagues by including an equal and unequal sample size condition. In real-world settings, equivalent groups are not often observed. When the sample sizes were equal the responses were evenly split (500/500 or 1000/1000 in each group. When sample sizes were unequal there were 75% more responses in one group (250/750 for the 1000 condition, 500/1500 for the 2000 condition).

**Number of Traits (RQ1).** We generated scores for respondents on five and ten traits. A five-trait analysis appeared in Lee and colleagues' (2021) original work and is also a common condition in other FC simulation studies (Frick et al., 2021; Schulte et al., 2021). Five-traits are assessed in the Big-Five FC assessment (Brown & Maydeu-Olivares, 2011). We also include a ten-trait condition to represent assessments used operationally such as the Character Skills Snapshot (seven traits; EMA, 2023) or OPAQ-32 (32-traits; Brown & Bartram, 2011). The number of items per trait was 12, resulting in 60 items (or 20 triad blocks) in the five-trait condition and 120 (40 blocks) in the 10-trait condition.

**DIF Effect Size (RQ2).** We simulated uniform DIF and manipulated the effect size of DIF by changing the amount added to the mean of items in one group that we selected to display DIF. Items were manipulated such that one group was consistently different on all DIF items (e.g., group 2 will always have effect size added to all items that display DIF). In the unequal sample size condition, the DIF effect size was always added to the smaller sample size group. For the magnitude of DIF, we rely on prior simulation research to determine the values because there has not been a practical examination of DIF for forced choice assessments. We used the same effect sizes from Lee and colleagues (2021) with a small effect size condition, .3, and a large one, .6. These effect sizes are in standardized units. These conditions are generally accepted as small and large in other simulations (Kim et al., 2016; Stark et al., 2006).

**Percent of DIF Blocks (RQ5).** The effect of the number of DIF blocks on DIF detection was examined by manipulating the percentage of items with DIF. When a block contained an item with DIF, we considered it a DIF block. We tested if there was a difference in the accuracy of DIF detection when 40%, 50%, or 60% of blocks display DIF. In the five-trait condition, this equated to 8, 10, or 12 total items with DIF. In the 10-trait condition, there were 16, 20, or 24 items with DIF.

### *Manipulated Conditions: Analysis*

In addition to considering how different data features influence DIF detection, we tested different analysis features. In practice, a researcher will not know which blocks contain DIF prior to determining an anchor, and these conditions represent the different, yet reasonable, decisions researchers can make when testing for DIF. The following conditions represent how the data were analyzed.

**Anchor Set Size (RQ6).** We examined the effect of anchor set size by manipulating the percentage of blocks specified as the anchors. Lee and colleagues (2021) did not study this beyond determining the minimum amount (20%) for model convergence in their study. We tested anchor block sets of 20% and 30%. For example, in the five-trait 20% condition this meant that four of the blocks were constrained equal across groups.

**Model-Misspecification (RQ7).** In this study a model misspecification is when a DIF block is used in the anchor set. We manipulated the amount of misspecification for the free-baseline conditions by varying the percent of DIF blocks included in the anchor set. 0%, 50%, or 100% of the total anchor blocks will have DIF. For example, in the 20% anchor set condition for five-traits there were four blocks in the anchor. In the 50% DIF in anchor set condition, two of these blocks contained DIF. There is little existing research on which we can base our decisions here, thus we tested a situation where the model was properly specified (0%), had some misspecification (50%), or was completely misspecified (100%). In all conditions, the percentage of blocks with DIF present remained constant regardless of the percent of DIF blocks in the anchor. For example, in the 40% blocks with DIF present condition for five traits, there were always eight DIF blocks on the test even as the number of DIF blocks in the anchor increased. Mixed-keyed blocks without DIF were used in the 0 and 50% misspecification condition. In the 100% misspecification condition, 2 or 4 mixed-keyed blocks with DIF were used for the five and ten-trait conditions respectively.

As the constrained-baseline approach constrains all but one block in each analysis, this condition and all other analysis factors are not applicable. This resulted in 288 conditions

(2x2x2x2x3x2x3) for the free-baseline approach and 48 (2x2x2x2x3) conditions for the constrained-baseline approach. A summary of these conditions is in Table 2.

Table 2 - Manipulated Simulation Conditions

| Factor Type | Factor | Levels |
|---|---|---|
| | **Data Generation Conditions** | |
| Sample Factors | Sample Size (Total) | 1000, 2000 |
| | Sample Size Equality | No/Yes<br>If No:<br>250/750<br>500/1500 |
| Assessment Factors | Number of Traits | 5, 10 |
| Block Factors | DIF Effect Size | Small (.3), Large (.6) |
| | Percent of Blocks With DIF | 40%, 50%, 60% |
| | **Analysis Conditions** | |
| Modeling | Anchor set size (% of blocks) | 20%, 30% |
| | DIF Included in Anchor | 0%, 50%, 100% (of total anchor set) |
| | Latent Scoring Approach | Free-Baseline or<br>Constrained- Baseline |

## Data Generation

Data was generated in two phases in R using a modified function from Frick and colleagues (2021). In phase one the parameters were generated. This included:

1. Vectors of trait scores equal to the number of traits in the condition were simulated from a multivariate normal distribution, *MVN* (0,1), for each respondent in the reference and focal group. Trait correlations followed the approach described in the constant factors.
2. Vectors of loadings (λ) equal to the number of items in the condition were sampled from a uniform distribution, *U* (0.65, 0.95).
3. Vectors of item means (μ) equal to the number of items in the condition were sampled from a uniform distribution, *U* (-1,1) respectively. These are common values to use for continuous item utilities (Brown & Maydeu-Olivares, 2011). When DIF is present, the

mean of one group was manipulated by adding the effect size to only one of the item means in the block.

4. Item errors (ε) were set to 1 in line with constraints placed on the model.
5. Measurement errors for each person were generated from a normal distribution, *N*(1, sqrt(ε)) for each item response.

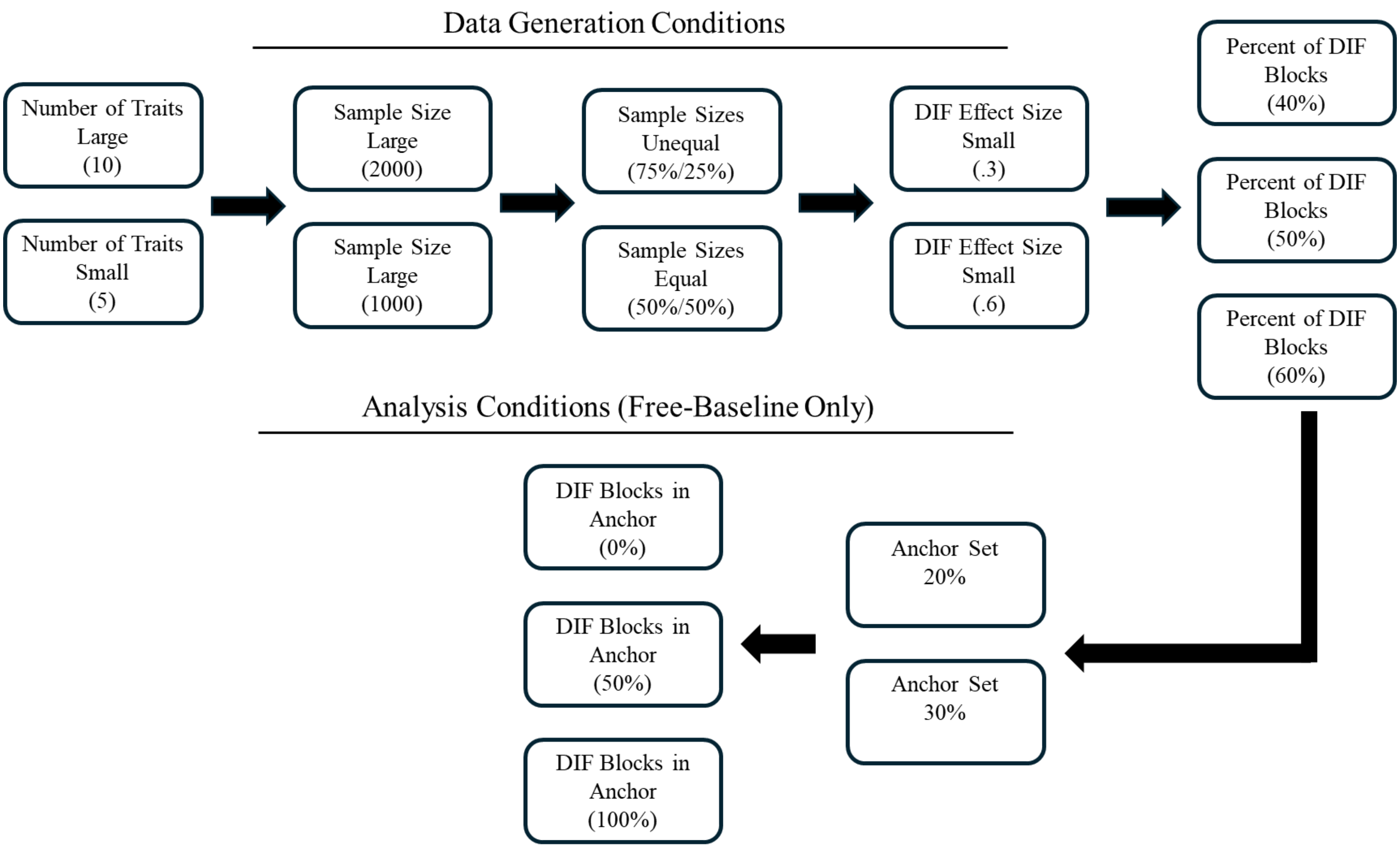
In phase two, the generated variables and parameters were used to compute each pairwise comparison between two items in a block using the expression in equation 4. For example, in block one, the value for the pairwise comparison between items 1 and 2 was 1 if the linear combination of values sampled in step one resulted in a value greater than or equal to 0 (indicating item 1 is preferred to item 2) and was 0 otherwise.

The simulation followed a two-step process (see Fig.4). First, the datasets were generated in line with the data generation conditions in Table 2. Then they were subjected to each analysis condition. For example, 500 datasets for the five-trait, N =1000 equal groups condition with a DIF effect size of .3 added to 40% of the blocks were generated. Then, we analyzed them using the free-baseline model for the 20% anchor set with 0% DIF in the anchor condition. Followed by the 20% anchor set with 50% DIF in anchor, and so on until they were subjected to each analysis condition.

## Analysis Plan

Analyses were conducted in R and Mplus. We used R to simulate the data, Mplus to analyze it (note that the TIRT model can be estimated in R using the thurstonianIRT package), and R to process the simulation results. The code to run the simulation, Mplus model files, and an analysis file to test our hypotheses are on the OSF (https://osf.io/fhd9w/?view_only=cccd50cce05a4dcea4df3a31fe963f2d).

Fig. 4 – Simulation Data and Analysis Approach


Data Generation Conditions
Number of Traits Large (10)
Number of Traits Small (5)
Sample Size Large (2000)
Sample Size Large (1000)
Sample Sizes Unequal (75%/25%)
Sample Sizes Equal (50%/50%)
DIF Effect Size Small (.3)
DIF Effect Size Small (.6)
Percent of DIF Blocks (40%)
Percent of DIF Blocks (50%)
Percent of DIF Blocks (60%)
Analysis Conditions (Free-Baseline Only)
Anchor Set 20%
Anchor Set 30%
DIF Blocks in Anchor (0%)
DIF Blocks in Anchor (50%)
DIF Blocks in Anchor (100%)

### *Convergence Rates*

We checked all replications for convergence via the calculation of the Wald Test, which will not be computed if the standard errors of parameters are not estimated. We denote these cases with ‘9999’ within the datasets in the Results folder. When non-convergence occurred, we checked those replications for errors and anomalies. In the free-baseline conditions, 0.98% of the models did not converge. All models that did not converge were from the ten-trait conditions. In the constrained-baseline models, 3.10% of the models did not converge. The non-convergence rates appeared to be due to memory limitations as indicated by error files.

### *Model Estimation*

Models were analyzed in Mplus using the *MplusAutomation* package (Hallquist, 2022) in R. The number of item constraints across groups varied based on the anchor item percentage condition. In the 20% anchor set condition, four or eight blocks were constrained equal. This means that all item parameters in the block were constrained such that the utility means and loadings were set to be equal across groups. In the 30% condition, six or twelve blocks were constrained equal. Depending on the 'DIF included in anchor condition,' some number of blocks displaying DIF were included in the anchor set.

### *Hypothesis Testing*

To answer our research questions and test our hypotheses, we examined if blocks containing a DIF item were accurately identified (indicated by a significant Wald test for each block). This was done by testing a subset of the freely estimated blocks in each condition. We tested four blocks, two with DIF and two without. This was done to reduce computation times and ensure parity in the number of statistical tests in each condition. Each Wald test was conducted with two degrees of freedom on the unconstrained utility means ($\mu_t$) in the block. Using the Wald test results, we calculated Type I error (α) rates as the proportion of non-DIF blocks incorrectly flagged as displaying DIF across replications and power (β) as 1 - the proportion of DIF blocks not flagged as DIF across replications.

## Results

The Type I error rates and power for the 288 free-baseline conditions are presented in Tables 3 and 4, respectively. Similarly, Tables 5 and 6 provide the Type I error rates and power for the 48 constrained-baseline conditions. Both tables include statistics for the data generation conditions, such as sample size equality, sample size, the percentage of blocks with DIF, DIF effect size, and the number of traits. The free-baseline tables also present statistics for the analysis conditions that did not apply to the constrained-baseline approach, namely the anchor set size and the percentage of misspecification in the anchor. The statistics shown in each table represent the average power or Type I error rates over 500 replications, with two DIF blocks (for power) or two non-DIF blocks (for Type I error rates) tested in each replication, after excluding any cases where the model did not converge. The remaining sections of the results are organized by research question. Graphs for each individual research question can be found in the supplementary materials of this manuscript.

Table 3

*Type I Error Rates in the Free-Baseline Conditions*

| | | | | 5 Factor | | 10 Factor | |
|---|---|---|---|---|---|---|---|
| Sample Size (Group 1/Group 2) | Percent DIF Blocks | Anchor Set Size (% of Blocks) | Misspecification<br>(% of DIF Items in Anchor) | Small DIF | Large DIF | Small DIF | Large DIF |
| 500/500 | 40 | 20 | 0 | **0.040** | **0.038** | **0.039** | **0.036** |
| | | | 50 | **0.046** | **0.048** | 0.054 | 0.054 |
| | | | 100 | 0.122 | 0.375 | 0.186 | 0.637 |
| | | 30 | 0 | **0.033** | **0.032** | **0.048** | **0.042** |
| | | | 50 | 0.056 | 0.063 | 0.082 | 0.202 |
| | | | 100 | 0.349 | 0.866 | 0.116 | 0.350 |
| | 50 | 20 | 0 | **0.040** | **0.038** | **0.041** | **0.037** |
| | | | 50 | **0.048** | **0.049** | 0.053 | 0.056 |
| | | | 100 | 0.121 | 0.369 | 0.181 | 0.635 |
| | | 30 | 0 | **0.033** | **0.032** | **0.048** | **0.044** |
| | | | 50 | 0.056 | 0.063 | 0.083 | 0.205 |
| | | | 100 | 0.348 | 0.863 | 0.145 | 0.365 |
| | 60 | 20 | 0 | **0.041** | **0.038** | **0.040** | **0.038** |
| | | | 50 | **0.048** | **0.050** | 0.052 | 0.054 |
| | | | 100 | 0.256 | 0.650 | 0.178 | 0.642 |
| | | 30 | 0 | **0.034** | **0.033** | **0.044** | **0.044** |
| | | | 50 | 0.052 | 0.093 | 0.083 | 0.207 |
| | | | 100 | 0.391 | 0.896 | 0.144 | 0.366 |
| 1000/1000 | 40 | 20 | 0 | **0.050** | **0.045** | **0.039** | **0.040** |
| | | | 50 | **0.047** | 0.072 | 0.066 | 0.097 |
| | | | 100 | 0.226 | 0.661 | 0.448 | 0.871 |
| | | 30 | 0 | **0.048** | **0.049** | **0.050** | 0.051 |
| | | | 50 | **0.048** | 0.077 | 0.156 | 0.429 |
| | | | 100 | 0.648 | 0.994 | 0.215 | 0.562 |
| | 50 | 20 | 0 | 0.051 | **0.047** | **0.040** | **0.039** |
| | | | 50 | **0.048** | 0.073 | 0.065 | 0.097 |
| | | | 100 | 0.226 | 0.664 | 0.458 | 0.871 |
| | | 30 | 0 | **0.048** | **0.049** | 0.054 | 0.051 |
| | | | 50 | **0.047** | 0.078 | 0.156 | 0.434 |
| | | | 100 | 0.649 | 0.993 | 0.225 | 0.574 |
| | 60 | 20 | 0 | 0.052 | **0.047** | **0.040** | **0.039** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | 50 | **0.049** | 0.078 | 0.065 | 0.097 |
| | | | 100 | 0.482 | 0.923 | 0.452 | 0.872 |
| | | 30 | 0 | **0.048** | **0.049** | 0.051 | 0.051 |
| | | | 50 | 0.071 | 0.168 | 0.152 | 0.431 |
| | | | 100 | 0.709 | 0.996 | 0.227 | 0.571 |
| 750/250 | 40 | 20 | 0 | **0.040** | **0.039** | **0.038** | **0.038** |
| | | | 50 | **0.047** | **0.050** | **0.045** | **0.046** |
| | | | 100 | 0.060 | 0.120 | 0.055 | 0.183 |
| | | 30 | 0 | **0.033** | **0.033** | **0.048** | **0.045** |
| | | | 50 | **0.050** | 0.057 | **0.048** | 0.084 |
| | | | 100 | 0.094 | 0.324 | 0.058 | 0.121 |
| | 50 | 20 | 0 | **0.040** | **0.039** | **0.040** | **0.037** |
| | | | 50 | **0.047** | **0.048** | **0.046** | **0.047** |
| | | | 100 | 0.061 | 0.124 | 0.058 | 0.179 |
| | | 30 | 0 | **0.035** | **0.034** | **0.049** | **0.045** |
| | | | 50 | **0.048** | 0.054 | 0.051 | 0.087 |
| | | | 100 | 0.092 | 0.323 | 0.068 | 0.139 |
| | 60 | 20 | 0 | **0.040** | **0.039** | **0.040** | **0.040** |
| | | | 50 | **0.047** | **0.048** | **0.046** | **0.047** |
| | | | 100 | 0.089 | 0.246 | 0.053 | 0.179 |
| | | 30 | 0 | **0.035** | **0.035** | **0.046** | **0.045** |
| | | | 50 | 0.051 | 0.061 | **0.050** | 0.084 |
| | | | 100 | 0.107 | 0.366 | 0.069 | 0.138 |
| 1500/500 | 40 | 20 | 0 | **0.048** | **0.050** | **0.043** | **0.042** |
| | | | 50 | **0.042** | 0.052 | 0.051 | 0.065 |
| | | | 100 | 0.075 | 0.218 | 0.130 | 0.446 |
| | | 30 | 0 | **0.049** | **0.048** | 0.052 | **0.048** |
| | | | 50 | **0.042** | **0.048** | 0.077 | 0.145 |
| | | | 100 | 0.183 | 0.615 | 0.089 | 0.228 |
| | 50 | 20 | 0 | **0.048** | **0.049** | **0.041** | **0.041** |
| | | | 50 | **0.042** | **0.050** | 0.053 | 0.065 |
| | | | 100 | 0.074 | 0.218 | 0.130 | 0.440 |
| | | 30 | 0 | **0.049** | **0.050** | 0.052 | 0.051 |
| | | | 50 | **0.042** | **0.047** | 0.075 | 0.147 |
| | | | 100 | 0.183 | 0.607 | 0.090 | 0.228 |
| | 60 | 20 | 0 | **0.049** | **0.047** | **0.041** | **0.040** |
| | | | 50 | **0.043** | **0.050** | 0.054 | 0.064 |
| | | | 100 | 0.163 | 0.467 | 0.129 | 0.440 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  |  | 0 | **0.048** | **0.049** | 0.051 | 0.051 |
|  |  | 30 | 50 | **0.049** | 0.072 | 0.075 | 0.150 |
|  |  |  | 100 | 0.213 | 0.683 | 0.087 | 0.226 |

Note - Table reports the Type I error rates averaged over 1000 tested blocks in each condition. Bold indicates values less than or equal to .05. Small DIF conditions had .3 added to the second group's item mean. Large DIF conditions had .6 added

Table 4
Power Rates in the Free-Baseline Conditions

|  |  |  |  | 5 Factor |  | 10 Factor |  |
|---|---|---|---|---|---|---|---|
| Sample Size (Group 1/Group 2) | Percent DIF Blocks | Anchor Set Size (% of Blocks) | Misspecification (% of DIF Items in Anchor) | Small DIF | Large DIF | Small DIF | Large DIF |
| 500/500 | 40 | 20 | 0 | 0.518 | **0.981** | 0.154 | 0.424 |
|  |  |  | 50 | 0.530 | **0.974** | 0.198 | 0.448 |
|  |  |  | 100 | 0.482 | **0.952** | 0.348 | 0.515 |
|  |  | 30 | 0 | 0.577 | **0.988** | 0.238 | 0.502 |
|  |  |  | 50 | 0.566 | **0.987** | 0.373 | 0.539 |
|  |  |  | 100 | 0.512 | 0.744 | 0.571 | **0.892** |
|  | 50 | 20 | 0 | 0.518 | **0.981** | 0.375 | **0.891** |
|  |  |  | 50 | 0.532 | **0.974** | 0.372 | **0.869** |
|  |  |  | 100 | 0.485 | **0.952** | 0.632 | **0.809** |
|  |  | 30 | 0 | 0.581 | **0.989** | 0.460 | **0.966** |
|  |  |  | 50 | 0.572 | **0.988** | 0.615 | **0.987** |
|  |  |  | 100 | 0.512 | 0.742 | 0.791 | **0.998** |
|  | 60 | 20 | 0 | 0.521 | **0.981** | 0.381 | **0.891** |
|  |  |  | 50 | 0.533 | **0.975** | 0.372 | **0.868** |
|  |  |  | 100 | 0.604 | **0.956** | 0.630 | **0.810** |
|  |  | 30 | 0 | 0.582 | **0.988** | 0.455 | **0.966** |
|  |  |  | 50 | 0.419 | **0.932** | 0.615 | **0.987** |
|  |  |  | 100 | 0.483 | 0.664 | 0.787 | **0.998** |
| 1000/1000 | 40 | 20 | 0 | 0.794 | **1.000** | 0.275 | 0.519 |
|  |  |  | 50 | **0.812** | **1.000** | 0.308 | 0.530 |
|  |  |  | 100 | 0.770 | **1.000** | 0.507 | 0.673 |
|  |  | 30 | 0 | **0.848** | **1.000** | 0.392 | 0.529 |
|  |  |  | 50 | **0.858** | **0.999** | 0.505 | 0.564 |
|  |  |  | 100 | 0.652 | **0.895** | 0.762 | **0.982** |
|  | 50 | 20 | 0 | 0.793 | **1.000** | 0.646 | **0.989** |
|  |  |  | 50 | **0.813** | **1.000** | 0.613 | **0.975** |
|  |  |  | 100 | 0.773 | **1.000** | **0.850** | **0.883** |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | 30 | 0 | **0.847** | **1.000** | 0.776 | **1.000** |
|  |  |  | 50 | **0.859** | **0.999** | **0.877** | **0.999** |
|  |  |  | 100 | 0.652 | **0.891** | **0.967** | **1.000** |
|  | 60 | 20 | 0 | 0.794 | **1.000** | 0.648 | **0.989** |
|  |  |  | 50 | **0.813** | **1.000** | 0.604 | **0.975** |
|  |  |  | 100 | **0.834** | **0.999** | **0.853** | **0.884** |
|  |  | 30 | 0 | **0.848** | **1.000** | 0.775 | **1.000** |
|  |  |  | 50 | 0.698 | **0.995** | **0.880** | **0.999** |
|  |  |  | 100 | 0.600 | **0.817** | **0.968** | **1.000** |
| 750/250 | 40 | 20 | 0 | 0.166 | 0.484 | 0.070 | 0.151 |
|  |  |  | 50 | 0.154 | 0.517 | 0.074 | 0.197 |
|  |  |  | 100 | 0.150 | 0.474 | 0.116 | 0.332 |
|  |  | 30 | 0 | 0.178 | 0.558 | 0.086 | 0.244 |
|  |  |  | 50 | 0.160 | 0.562 | 0.124 | 0.371 |
|  |  |  | 100 | 0.200 | 0.507 | 0.204 | 0.566 |
|  | 50 | 20 | 0 | 0.166 | 0.484 | 0.125 | 0.372 |
|  |  |  | 50 | 0.155 | 0.519 | 0.124 | 0.367 |
|  |  |  | 100 | 0.149 | 0.473 | 0.192 | 0.599 |
|  |  | 30 | 0 | 0.180 | 0.558 | 0.140 | 0.466 |
|  |  |  | 50 | 0.163 | 0.559 | 0.185 | 0.595 |
|  |  |  | 100 | 0.197 | 0.504 | 0.271 | 0.788 |
|  | 60 | 20 | 0 | 0.165 | 0.485 | 0.125 | 0.376 |
|  |  |  | 50 | 0.154 | 0.518 | 0.132 | 0.368 |
|  |  |  | 100 | 0.210 | 0.594 | 0.195 | 0.595 |
|  |  | 30 | 0 | 0.179 | 0.561 | 0.136 | 0.465 |
|  |  |  | 50 | 0.134 | 0.391 | 0.186 | 0.598 |
|  |  |  | 100 | 0.189 | 0.465 | 0.267 | 0.787 |
| 1500/500 | 40 | 20 | 0 | 0.278 | 0.758 | 0.111 | 0.271 |
|  |  |  | 50 | 0.275 | 0.773 | 0.111 | 0.314 |
|  |  |  | 100 | 0.228 | 0.729 | 0.234 | 0.507 |
|  |  | 30 | 0 | 0.305 | **0.815** | 0.161 | 0.393 |
|  |  |  | 50 | 0.300 | **0.825** | 0.233 | 0.511 |
|  |  |  | 100 | 0.326 | 0.641 | 0.361 | 0.760 |
|  | 50 | 20 | 0 | 0.279 | 0.761 | 0.207 | 0.619 |
|  |  |  | 50 | 0.276 | 0.774 | 0.180 | 0.586 |
|  |  |  | 100 | 0.228 | 0.729 | 0.391 | **0.843** |
|  |  | 30 | 0 | 0.304 | **0.814** | 0.257 | 0.754 |
|  |  |  | 50 | 0.301 | **0.829** | 0.338 | **0.867** |
|  |  |  | 100 | 0.323 | 0.642 | 0.540 | **0.960** |
|  | 60 | 20 | 0 | 0.277 | 0.762 | 0.204 | 0.619 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 50 | 0.273 | 0.774 | 0.179 | 0.586 |
| | | 100 | 0.359 | **0.812** | 0.396 | **0.841** |
| | 30 | 0 | 0.305 | **0.817** | 0.257 | 0.754 |
| | | 50 | 0.233 | 0.670 | 0.339 | **0.867** |
| | | 100 | 0.293 | 0.596 | 0.540 | **0.959** |

Note - Table represents the power rates averaged over 1000 tested blocks in each condition. Bold indicates values greater than or equal to .80. Small DIF conditions had .3 added to the second group's item mean. Large DIF conditions had .6 added.

Table 5

Constrained Baseline Conditions Type I Error Rates

| Sample Size (Group 1/Group 2) | Percent DIF Blocks | 5 Factor | | 10 Factor | |
|---|---|---|---|---|---|
| | | Small DIF | Large DIF | Small DIF | Large DIF |
| 500/500 | 40 | 0.065 | 0.154 | 0.068 | 0.155 |
| | 50 | **0.050** | 0.066 | 0.065 | 0.138 |
| | 60 | 0.054 | 0.092 | 0.076 | 0.199 |
| 1000/1000 | 40 | 0.098 | 0.329 | 0.112 | 0.320 |
| | 50 | 0.055 | 0.103 | 0.104 | 0.270 |
| | 60 | 0.067 | 0.179 | 0.147 | 0.385 |
| 750/250 | 40 | 0.051 | 0.071 | **0.049** | 0.064 |
| | 50 | **0.044** | 0.053 | **0.048** | 0.064 |
| | 60 | **0.048** | 0.062 | **0.049** | 0.079 |
| 1500/500 | 40 | 0.057 | 0.100 | 0.064 | 0.108 |
| | 50 | **0.041** | 0.055 | 0.061 | 0.107 |
| | 60 | **0.050** | 0.071 | 0.075 | 0.146 |

Note - Table represents the Type I rates averaged over 1000 tested blocks in each condition. Bold indicates values less than or equal to .05. Small DIF conditions had .3 added to the second group's item mean. Large DIF conditions had .6 added.

Table 6
Constrained Baseline Conditions Power

| | | 5 Factor | | 10 Factor | |
|---|---|---|---|---|---|
| Sample Size (Group 1/Group 2) | Percent DIF Blocks | Small DIF | Large DIF | Small DIF | Large DIF |
| 500/500 | 40 | **0.812** | **1.000** | 0.443 | 0.538 |
| | 50 | **0.840** | **1.000** | 0.761 | **0.998** |
| | 60 | **0.817** | **1.000** | **0.824** | **1.000** |
| 1000/1000 | 40 | **0.979** | **1.000** | 0.526 | 0.572 |
| | 50 | **0.982** | **1.000** | **0.958** | **1.000** |
| | 60 | **0.979** | **1.000** | **0.977** | **1.000** |
| 750/250 | 40 | 0.256 | 0.787 | 0.163 | 0.450 |
| | 50 | 0.268 | **0.816** | 0.221 | 0.744 |
| | 60 | 0.255 | 0.793 | 0.265 | **0.811** |
| 1500/500 | 40 | 0.466 | **0.973** | 0.312 | 0.532 |
| | 50 | 0.488 | **0.979** | 0.460 | **0.944** |
| | 60 | 0.475 | **0.974** | 0.514 | **0.968** |

Note - Table reports the power rates averaged over 1000 tested blocks in each condition. Bold indicates values greater than or equal to .80. Small DIF conditions had .3 added to the second group's item mean. Large DIF conditions had .6 added.

## RQ1. Number of Traits

Lee and colleagues (2021) found that when there was no misspecification in the anchor the number of traits did not affect Type I error or power (e.g., moving from three to five traits resulted in similar Type I error and power). We replicated their results in the 0% misspecification condition of our study. Type I error rates and power tended to decrease as the number of traits increased when averaging across all free-baseline misspecification conditions. When using the constrained-baseline approach, Type I error rates increased, and power decreased as the number of traits increased.

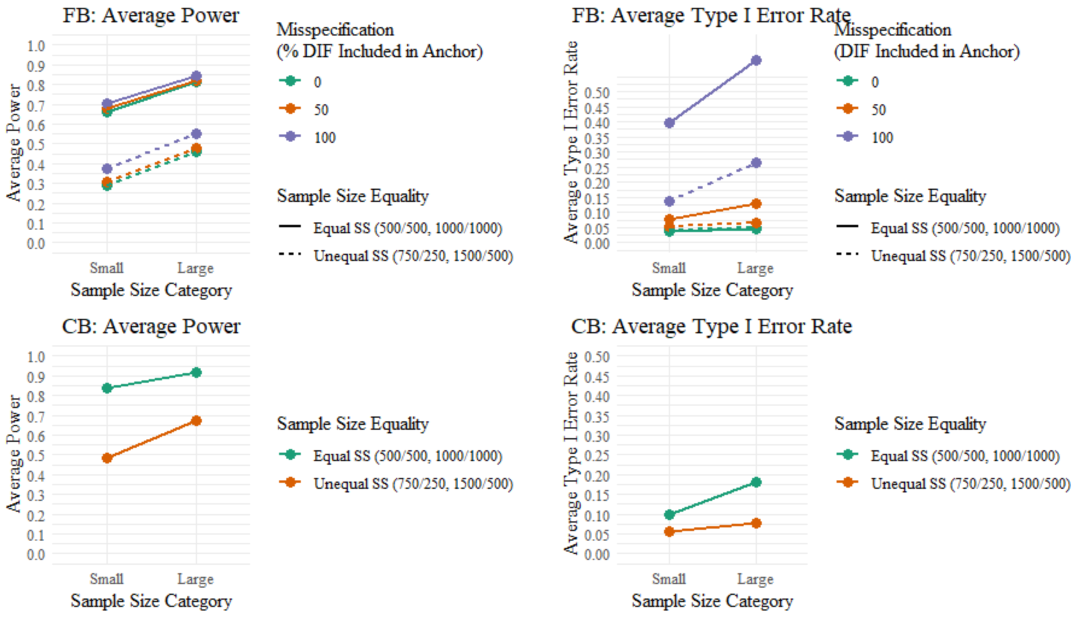
## RQ2. DIF Effect Size

Lee et al. (2021) found that as the DIF effect size increased Type I error rates remained the same and power increased when there was no misspecification. We replicated their results in the 0% misspecification free-baseline conditions. When misspecification conditions were included, power and Type I error rates increased. In the constrained-baseline conditions we only replicated Lee and colleagues' (2021) finding that power increased with the DIF effect size whereas Type I error rates increased with a greater DIF effect size instead of remaining the same.

### RQ3/8. Sample Size and Sample Size Equality

Lee et al. (2021) found that as sample size increased (e.g., from 500 to 1000 respondents in each group) power increased and Type I error rates were unaffected. In their study, sample sizes were equal and there was no misspecification. We replicated their results when sample size was equal in the constrained and free-baseline conditions (see Fig. 5). We also replicated their findings when there was no misspecification in the unequal sample size conditions. When misspecification was present, we replicated their findings that power increased, but Type I error rates increased as sample size increased in equal and unequal conditions. Power and Type I error rates were higher in equal sample size conditions than unequal ones.

Fig. 5 - Average Power and Type I Error Rates by Sample Size, Equality, and Misspecification



### RQ5. Percent of DIF Blocks

Lee et al.’s (2021) findings showed that as the percentage of DIF blocks on the test increased, power and Type I error rates were unaffected with either approach. We replicated their findings for Type I error rates and found that they remained mostly unaffected as the number of DIF blocks increased in both conditions. However, we found that that power increased when moving from 40 to 50% of blocks containing DIF and then remained constant from 50 to 60% in the free and constrained-baseline conditions.

### RQ6. Anchor Size

We investigated the effect of increasing the number of blocks used as an anchor to set the scale between groups in the free-baseline conditions. We found that as the size of the anchor set increased, average power and Type I error rates increased marginally.

### RQ7. Misspecification

We examined the effect of misspecification on power and Type I error rates in the free-baseline conditions. Misspecification was considered a case where blocks with DIF were used in the anchor. We found that as the level of misspecification increased, the average Type I error rates across all conditions increased moving from .043 when there was no misspecification (0% conditions), to .082 in the 50% conditions, and .351 in the 100% conditions. As misspecification increased, power also increased moving from .556 in the 0% conditions to, .569 in the 50% conditions, and .618 in the 100% conditions.

### Interactions

We conducted a three-way ANOVA to examine the main effects and interactions between anchor set size, percentage of DIF blocks on the test, and the level of misspecification on the Type I error and power rates from the free-baseline conditions. The mean Type I error and power rates for each condition across all replications were used as the dependent variables. Misspecification was a significant predictor of Type I error rates, *F* (1, 276) = 154.87, *p* < 0.001, $\omega^2$= .35, indicating that 35% of the variance in Type I error rates is accounted for by the level of misspecification after controlling for all other variables in the model. The percentage of items with DIF on the test was a significant predictor of differences in power, *F* (1, 276) = 5.73, *p* = 0.004, $\omega^2$ < .001, indicating that while significant it accounts for only a small amount of variance in power rates. All other main effects and their interactions in predicting Type I error or power rates were non-significant (see the supplementary materials of this manuscript for a table of all the effects).

### *Exploratory Analyses*

Our registration did not include a planned analyses for the interactions between sample size, number of traits, and DIF effect size conditions by the percent of misspecification. These conditions produced the most variation in Type I error and power rates though. As such, we conducted a four-way ANOVA to examine the main effects and interactions of these variables with misspecification in a single analysis.

Table 7 - Significant Effects for Exploratory ANOVA

| Effect | *F* | *df* (Within, Between) | *p* | $\omega^2$ |
|---|---|---|---|---|
| **Type I Error** | | | | |
| DIF Effect Size (ES) | 69.784 | (1, 256) | < .001 | 0.07 |
| Misspecification (MS) | 362.847 | (1, 256) | < .001 | 0.35 |
| Sample Size (SS) | 34.774 | (3, 256) | < .001 | 0.1 |
| MS * ES | 81.005 | (1, 256) | < .001 | 0.08 |
| MS * NT | 3.409 | (1, 256) | 0.02 | 0.007 |
| MS * SS | 11.641 | (3, 256) | < .001 | 0.01 |
| MS * ES * NT | 37.341 | (1, 256) | < .001 | 0.11 |
| **Power** | | | | |

| | | | | |
|---|---|---|---|---|
| ES | 512.838 | (1, 256) | < .001 | 0.32 |
| MS | 12.416 | (1, 256) | < .001 | 0.007 |
| SS | 229.147 | (3, 256) | < .001 | 0.43 |
| NT | 25.144 | (1, 256) | < .001 | 0.02 |
| ES *SS | 12.156 | (1, 256) | < .001 | 0.02 |
| MS * NT | 46.273 | (1, 256) | < .001 | 0.03 |

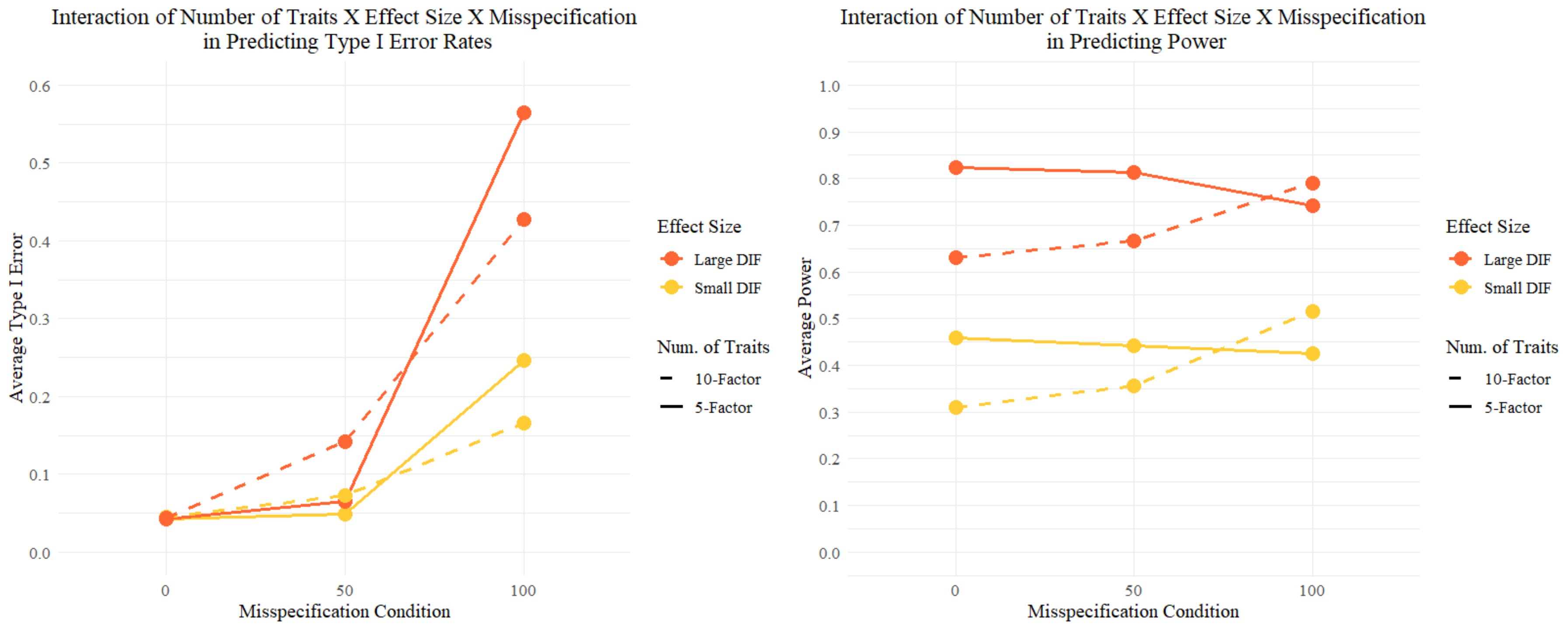
Note - $\omega^2$= Proportion of variance accounted for by independent variables.

**Type I Error Rates.** The main effects for DIF effect size, the level of misspecification, and sample size were all significant at the *p* < .001 level, indicating that all factors but the number of traits contribute to Type I error rates. There was a significant two-way interaction effect between the level of misspecification and the DIF effect size (see Table 7). The impact of misspecification on Type I error rates increased as DIF effect sizes increased depending on whether the DIF effect size is small or large (see Fig. 6). There was also an interaction with the number of traits and misspecification, such that the effect of misspecification on Type I error rates varied by the number of traits. Fig. 6 shows that Type I error rates increased with the percent of misspecification in the ten-trait conditions, whereas the five-factor conditions remained approximately equal on average moving from 0% to 50% misspecification in the small DIF conditions before increasing in the 100% misspecification conditions. There was also a two-way interaction between sample size and misspecification. While all conditions followed a similar pattern of results with an increase in Type I error rates from 0% to 50% misspecification, there is a much steeper rise in the 1000/1000 conditions than the others indicating a highly misspecified model produces a substantial amount of false positives (see Fig. 7).

**Power.** The main effects for DIF effect size, the level of misspecification, the number of traits, and sample size in predicting power were all significant at the *p* < .001 level. Effect size and sample size were the biggest predictors of differences in power accounting for 32% and 43%

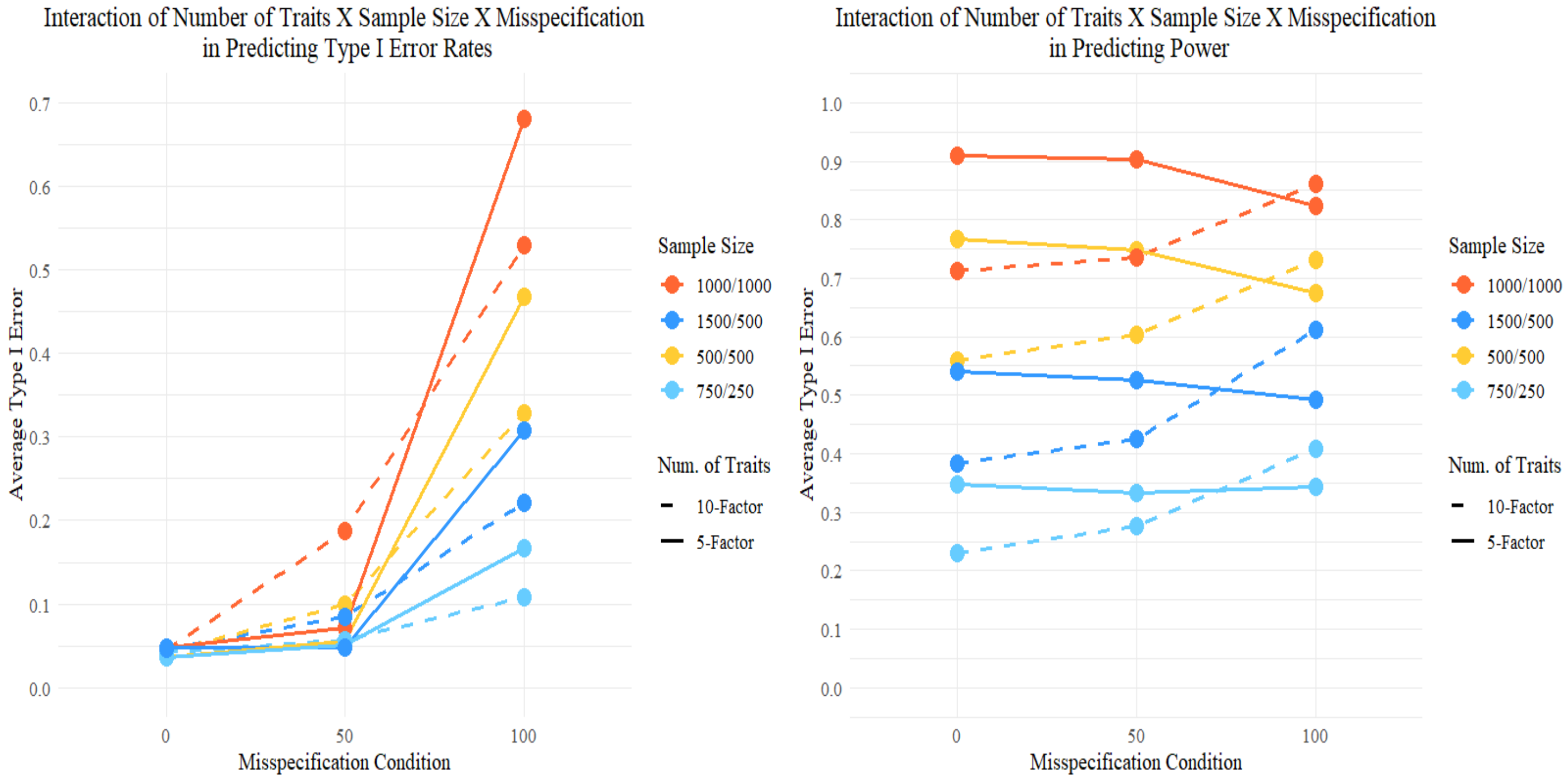
of the variance respectively. While the two-way interactions between effect size and sample size as well as misspecification and trait size were significant, they accounted for only a small portion of the variance in power (3% and 2% respectively). This can be seen in Fig. 6 where power increases for the ten-trait conditions as misspecification increases and power decreases in the five-factor conditions. However, power is generally unacceptable and only reaches .7 in the 0% misspecification condition for ten-factors. This may indicate that a sample size of 2000 is not sufficient to accurately detect DIF items for the ten-trait conditions.

Fig. 6 - Interaction Between the Number of Traits, Effect Size, and Misspecification


Interaction of Number of Traits X Effect Size X Misspecification in Predicting Type I Error Rates

Interaction of Number of Traits X Effect Size X Misspecification in Predicting Power

Fig. 7 - Interaction Between The Number of Traits, Sample Size, and Misspecification



# Discussion

Our findings suggest that the free-baseline method for detecting DIF is not accurate when DIF blocks are included in the anchor set, a possible model misspecification in practice. Our results suggest the constrained-baseline approach may provide more accurate results when the anchor set is contaminated. In particular, when sample sizes are equal and there are five traits, the ability of the method to detect DIF items is high. When there is no misspecification, the free-baseline method performs well and better than the constrained-baseline approach, replicating the results of Lee and colleagues (2021). The rest of the discussion section will focus on how our results relate to the work of Lee et al. (2021), the differences between the constrained and free-baseline approaches, and how other design features of the assessment impact the ability to detect DIF.

## Replication of Key Findings

Lee and colleagues (2021) reported on the number of traits, sample and effect size, the percent of DIF blocks on the test, and differences between the free-baseline and constrained-baseline approaches using a first-order TIRT model. We sought to replicate their results using the second-order model which allows flexibility in testing uniform and nonuniform DIF separately. We examined the replicability of their work in the context of our 0% misspecification condition, where there were no DIF blocks included in the anchor. When this was the case, we completely replicated their findings on the free-baseline and constrained-baseline approach with exception to the effect size of DIF and percentage of blocks with DIF on the test. In the constrained-baseline approach it appeared that as the effect size of DIF grew, it became harder to accurately detect DIF blocks. Based on the replication of these selected conditions, this may indicate generalizability of other findings that Lee et al. (2021) examined in the first-order model (such as the number of DIF items included in a block having little effect on DIF detection) to the second-order model.

## The Free-Baseline vs. Constrained-Baseline Approach

The free-baseline approach requires one to identify an anchor set of blocks to set the scale between the two groups being tested. Then all remaining blocks can be tested for DIF. The constrained-baseline approach constrains all blocks except one across groups and tests just that block. The free-baseline approach is the better approach when DIF blocks are not contaminating the anchor and the effect size is large as it results in more accurate detection of blocks without DIF for each comparable constrained-baseline condition. However, in conditions where 50% of the anchor is composed of DIF blocks, accurate detection of blocks without DIF is equal to or worse than the constrained-baseline approach. It becomes substantially worse as misspecification in the anchor increases to 100%. In the constrained-baseline method, all DIF and non-DIF blocks are constrained. The constrained-baseline method likely has lower Type I error rates than a misspecified free-baseline model because more non-DIF blocks are being constrained than DIF blocks, which creates a less contaminated anchor. For example, in the five-factors condition where 40% of blocks contain DIF, there were at least 11 non-DIF blocks constrained. In the 60% condition, there were at least seven. All constrained-baseline conditions have more non-DIF blocks being constrained than in any misspecification condition for the free-baseline approach. In terms of accurately detecting a DIF block, the free-baseline method is equivalent to the constrained-baseline approach when the model is correctly specified. However, this becomes unpredictable when misspecification is present. This is in line with previous findings that the power of the free-baseline approach can be greatly reduced when the anchor set is tainted

(Guenole, 2018; Rivas et al., 2009). Meanwhile, the power of the constrained-baseline approach remains consistent across conditions.

Our results suggest if one can be reasonably certain their anchor set is free of any DIF blocks and the effect size of DIF is large, they should use the free-baseline approach. In these cases, the approach will also be robust against sample size inequality. Meanwhile, the constrained-baseline approach is difficult to recommend because while it does result in higher power, but still unacceptable power, than using the free-baseline approach with a tainted anchor set, it results in unacceptable Type I error rates in most conditions. It is possible that smaller levels of misspecification, only 25% DIF blocks in the anchor for example, may result in more accurate DIF detection in the free-baseline model approach.

In practice it would require in-depth validation work ahead of DIF testing to hypothesize that the anchor is not tainted, but one cannot be confident. The literature suggests the use of the sequential free-baseline approach where the constrained-baseline method is used to identify non-DIF blocks that are then used as the anchor in the free-baseline approach (Chun et al., 2016; Lee et al., 2021; Stark et al., 2006). This seems like the best approach. The constrained-baseline approach could first be used, with certainty that it will correctly flag DIF items in most cases where sample sizes are equal or the effect size of DIF is large. Though the approach may incorrectly flag non-DIF items, any that remain after initial testing can be presumed to be items without DIF. This would allow a correctly specified anchor to be chosen for the free-baseline method. The analysis would then allow the research to make more accurate conclusions about the blocks being identified as containing and not containing DIF.

### The Influence of DIF and Assessment Design

The accuracy of non-DIF block detection was reduced moving from 0 to 50% misspecification. It was dramatically reduced when the entire anchor set was tainted. As the DIF effect size increased we found accurate DIF detection also increased. We also found that there was a marginal benefit to using a larger anchor set in the free-baseline method for detecting items, however, this also came with increased Type I error rates. These findings are consistent with those of other researchers (Rivas et al., 2009; Stark et al., 2006), who found that the size of an untainted anchor set did not significantly affect detection of DIF or non-DIF items in the free-baseline method, implying one can use an anchor set that is relatively small. This reduces the opportunity for misspecification by limiting the number of anchor blocks needed.

We also found a slight increase in power when moving from 40 to 50% of the test containing blocks with DIF in both the free and constrained-baseline approaches. Lee et al. (2021) found that there were no differences as the amount of DIF blocks increased. Our conflicting findings may in part be due to differences in the spread of DIF across traits in our simulations. While Lee and colleagues simulated DIF onto mostly a single trait, we equalized the spread of DIF due to other design factors of our simulation, meaning multiple traits included a DIF block. These included the need to maintain enough free blocks to test for DIF while having a higher percentage of total DIF blocks. The potential effect of this is most apparent in comparing the 40% and 50% blocks with DIF conditions. While the five-factor conditions were relatively constant, the ten-trait conditions had lower power in the 40% conditions compared to the 50% and 60% conditions. This indicates that there are potential differences in the effect of detecting DIF based on how it is spread out across more traits.

The effect size, number of traits, and level of misspecification also interacted in complex ways that make other broad generalizations about the efficacy of the methods difficult. While the two trait conditions generally follow the same pattern of Type I error results there was an inconsistent relationship between the number of traits and misspecification in relation to power. There was a consistent increase in power as the effect size and misspecification increased in the large conditions. Meanwhile, power consistently decreased in the small conditions. Practically, it appears that as misspecification increases all Wald tests are significant more often for the ten-factor conditions. This results in more Type I errors and what seems like greater power. Huang et al. (2024) found in their simulation that misspecification resulted in greater detection of DIF even when it was not present. This aligns with the results of the ten-trait condition. It is possible that in simpler models, like those in the five-trait conditions, setting the scale by use of a completely tainted anchor shifts the scale of both groups such that Type I errors can still be detected while DIF blocks remain hard to detect. This is a novel area of research and to our knowledge there has not been substantial research examining the interaction between the number of traits on a test and misspecification.

## Sample Size

As expected, as sample size increased, power increased, and Type I error rates decreased when there was no misspecification. Power was higher when sample sizes were equal. Type I error rates were lower when sample sizes were unequal. Others have also found the same results when examining sample size imbalance (Stark et al., 2006; Yoon & Lai, 2016). In the context of this work, the smaller group in the unequal conditions always had the DIF effect such that all of the DIF was systematically in the smaller group. The larger group is providing more information to the model. This may be resulting in a decrease in Type I error rates, as the parameter estimates of both groups are pulled towards those of the larger group, overwhelming the DIF effect. However, this does not allow the model to accurately identify DIF items. Taken together, the results indicate that a large, approximately equal sample size would give the most accurate results in detecting DIF and non-DIF blocks.

## Limitations and Future Directions

This work has several limitations. First, our conclusions are limited to the specific simulation conditions we tested. Secondly, we focused on DIF in the context of a two-group comparison (e.g., male/female) comparisons. These results may not extend to continuous variables or more than two groups. Another limitation of this work is that we did not examine nonuniform DIF. Using the latent-scoring approaches, it still needs to be determined if nonuniform DIF can be accurately detected in the second-order TIRT model. Additionally, our selection of simulated conditions is not based on real-world analyses because no available ones exist. We based our decisions on other similar simulation studies and our experience using FC assessments operationally. As more work on DIF for FC enters the literature, expanding simulation research to represent data and effects seen in practice will be important.

Our replication of Lee and colleagues (2021) work is also limited. While we attempted to replicate their methodology exactly, we found complexities that we had not considered in our original simulation design. These features had to be accounted for and required us to diverge in some unique ways. For example, they had significantly more mixed-keyed blocks on their test. We chose to pursue a realistic balance of mixed-keyed blocks. They also simulated DIF onto primarily one trait. This was not possible with the percentage of DIF blocks on our test. These differences make our replication more conceptual than exact. Finally, we did not have the same proportion of DIF added to mixed-keyed block positive items in the five-trait and ten-trait conditions. Both conditions had only DIF added to one positive item in a mixed-keyed block. This may reduce comparability of the results.

In future work, we will expand the simulation to examine more conditions related to mixed-keyed blocks. This includes examining how the quantity of mixed-keyed blocks in the anchor effects Type I error and power. We also will examine how adding DIF to a positive or negative item within the block changes the results. Finally, we will ensure complete parity across the different trait conditions.

There are also several areas that are still open questions with regard to the latent variable modeling approach. In our study, we did not vary the number of items on each trait. It is probable that in real-world circumstances, traits will not have an equal number of items. In the design of our simulation, we also did not keep the number of items used as an anchor consistent across traits. Future work should examine the effect of holding this condition equal, although it is not likely to occur in real-world scenarios either.

## Conclusion

There is an increasing interest in using FC assessments in high-stakes scenarios. In this study, we replicated and extended the work on DIF testing in FC assessments introduced by Lee and colleagues (2021). Our approach used the second-order TIRT model to allow for a separate examination of uniform and non-uniform DIF if desired. When no misspecification was present, we replicated almost all of their findings. However, when misspecification was present the free-baseline approach may not provide reliable or accurate enough results to use it with certainty. Meanwhile, the constrained-baseline approach provided more consistent results and had good power when sample sizes were equal. However, its ability to correctly identify non-DIF items was limited to a few cases such as when the sample size was small and unequal. These results indicate that the sequential free-baseline approach may be the most efficacious for detecting DIF and non-DIF items. Our results also underscore the importance of rigorous and in-depth validation work of the test and items ahead of DIF testing. This may require qualitative investigation to ensure that there are blocks that operate equally across groups such that a correctly specified anchor set can be selected.

## References

ACT. (2022). Mosaic by ACT Social Emotional Learning Assessment Technical Manual: Elementary School Assessment. https://www.act.org/content/dam/act/unsecured/documents/2022/R2138-Mosaic-SEL-ES-Technical-Manual-03-2022.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56. https://doi.org/10.1111/j.2044-8325.1996.tb00599.x

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, 15(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x

Brown, A., & Bartram, D. (2009). Doing less but getting more: Improving forced-choice measures with IRT. Society for Industrial & Organizational Psychology annual conference, New Orleans. https://kar.kent.ac.uk/44788/

Brown, A., & Bartram, D. (2011). The occupational personality questionnaire revolution: Applying Item Response Theory. *Organisationalpsychology.nz*.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71(3), 460–502.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *The Journal of Applied Psychology*, 104(11), 1347–1368. https://doi.org/10.1037/apl0000414

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement*, 40(7), 486-499.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, 1992(1), i-40.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*(4), 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Enrollment Management Association. (2023). Summary of the Snapshot research and findings. https://assets-global.website-files.com/62b5ff6837362e413f2bfed4/64e26e4744c35d763506450b_summary-of-the-snapshot-research-and-findings.pdf

Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*, 1-29.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Jurgensen, C. E. (1944). Report on the "Classification Inventory," a personality test for industrial use. *The Journal of Applied Psychology*, *28*(6), 445–460. https://doi.org/10.1037/h0053595

Kim, E. S., Joo, S. H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 870-887.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, 72(3), 469-492.

Kirkpatrick, J. J. (1951). Cross-validation of a forced-choice personality inventory. *The Journal of Applied Psychology*, *35*(6), 413–417. https://doi.org/10.1037/h0061581

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22-56.

Lee, P., Joo, S.-H., & Stark, S. (2021). Detecting DIF in multidimensional forced choice measures using the Thurstonian Item Response Theory Model. *Organizational Research Methods*, *24*(4), 739–771.

Lee, P., Joo, S. H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods. *Personality and Individual Differences*, 191, 111555. https://doi.org/10.1016/j.paid.2022.111555

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced-choice measures with different scoring approaches. *Personality and Individual Differences*, 123, 229–235. https://doi.org/10.1016/j.paid.2017.11.031

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, 77(3), 389-414. https://doi.org/10.1177/0013164416646162

Lohmann, A., Astivia, O. L., Morris, T. P., & Groenwold, R. H. (2022). It's time! Ten reasons to start replicating simulation studies. *Frontiers in Epidemiology*, 2, 973470.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251-265.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*(6), 935–974. https://doi.org/10.1080/00273171.2010.531231

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284–291.

Richardson, M. W., & Kuder, G. F. (1933). Making a rating scale that measures. *The Personnel Journal*, 12, 36–40. https://psycnet.apa.org/fulltext/1933-04735-001.pdf

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *The Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thurstone, L. L. (2017). A law of comparative judgment. In Scaling (pp. 81–92). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315128948-7/law-comparative-judgment-louis-thurstone

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Exploring the impact of negative and positive faking on multidimensional forced-choice personality measures. *Human Performance*, *19*(3), 175–199. https://doi.org/10.1207/s15327043hup1903_1

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, 72(3), 221-261.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479-498.

Wetzel, E., Brown, A., Hill, P., Chung, J., Robins, R., & Roberts, B. (2017). The narcissism epidemic is dead; long live the narcissism epidemic. *Psychological Science*, 28(12), 1833-1847.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339-361.

Yoon, M., & Lai, M. H. C. (2017). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 201–213. https://doi.org/10.1080/10705511.2017.1387859

# Appendicies

## Appendix 1 – Simulation Code R Script

```
#---------------------------------- Table of Contents --------------------------------------- #

# Note. Scripts were broken up by trait, effect size, and the free-baseline/constrained baseline conditions.

# All code is maintained script to script and the only thing that changes is the simulation conditions.

# Note to fully run the simulation you must have a valid copy of Mplus in the source file location.

# Section I. Load dependencies

# Section II. Local Function

# Section III. Define Constant Parameters

# Section VI. Simulation

##############################################################################
#############################################################################

####################### Section I. Load dependencies #############################

##############################################################################
##############################################################################


# Load each package

library(clubSandwich)

library(doParallel)

library(future)

library(furrr)

library(mvtnorm)

library(MplusAutomation)
```

```
# Replace following with the directory of Scripts and Models
#source <- ('/example/Simulation/Script and Models')

setwd(source)
k.start <- 1

# number of replications

k.range = 500
#------------------------------------------------------#
# Change the following based on conditions
#------------------------------------------------------#
# Sample size, 1 = 500/500, 2 = 1000/1000, unequal: 3=750/250, 4= 1500/500
Samplesize = 1

# Trait size, 1 = 5trait, 2 = 10 trait
r = 1

# DIF effect Size, 1 = small (.3), 2 = large (.6)
d = 1
# Percent of items with DIF, 1 = 40%, 2 = 50%, 3 = 60%
b = 1
# Approach, 1 is FB, 2 is CB
a = 1

# AS for 10f =
size = 1
```

```
# Misspecification folder
ldif = 1

approach1 = c('Free-Baseline')
approach2 = c('Constrained-Baseline')
approach_list = list(approach1,approach2)
approach = approach_list[[a]]
AS1 = c('30Anchor')
AS2 = c('20Anchor')
ancsize = list(AS1,AS2)
as = ancsize[[size]]

ld1 = ('0DIF')
ld2 = ('50DIF')
ld3 = ('100DIF')

LD = list(ld1,ld2,ld3)
ld = LD[[ldif]]
# Meta-info
# Set folder paths. These are change based on free or constrained approach
folder5 = file.path(source,"5factor",approach)
folder10 = file.path(source,"10factor",approach)
folder_list = list(folder5,folder10)
folder = folder_list[[r]]

set.seed(123)
seeds <- sample.int(10000, k.range, replace = TRUE)
################### Section II. Local functions (Frick et al., 2021)#####################
```

```
#--------------------------------------------------------------------------------#
# The following function is used to simulate forced response data. It incorporates the
manipulated
# DIF items from and outside object, itemsg2.
#--------------------------------------------------------------------------#
tirt.sim <- function(traits, items,itemsg2, design.load, design.mat, nblocks) {



  #load.mat: matrix of factor loadings: rows=items, columns=traits
  load.mat <- items[,2]*design.load
  g2_items<- itemsg2
  #sample error from N(0,uniqueness)
  #error: matrix: rows=persons, cols=items
  error <- matrix(nrow=nrow(traits), ncol=nrow(items))
  for (i in 1:nrow(items)) {
    error[,i] <- rnorm(n=nrow(traits), mean=0, sd=sqrt(items[i,3]))
  }

  ##compute utilities for each item and each participant

  utilG1 <- t(items[,1] + load.mat %*% t(traits[1:g1,])) + error[1:g1,]
  utilG2 <- t(g2_items[,1] + load.mat %*% t(traits[g2:N,])) + error[g2:N,]

  util <- rbind(utilG1,utilG2)

  ##utility differences (compute the difference in latent utilities for each participant and the
related item.)
  util.diff <- util %*% t(design.mat)
  #dichtotomise
  outcomes <- ifelse(util.diff<0,0,1)

  group <- c(rep(1, nrow(traits) / 2), rep(0, nrow(traits) / 2))

  result <- cbind(group,outcomes)

  return(result)
}

#####  Simulation function to be passed to parallel processing function
run_simulation <- function(k,seed) {

  #### Error coding for failure of seed to produce results.
  ### Errors are stored, then new seed is selected and it is ran again

    set.seed(seed + k  + k.range)
  session_folders <- sapply(1:ncores, function(i) {
    session_folder <- file.path("root_directory",
paste0('traitcon_',r,'_ss_',Samplesize,"_ES_",d,"_difCon_",b,"_","session_", i))
    if (!dir.exists(session_folder)) {
      dir.create(session_folder)
    }
    return(session_folder)
  })
  pid <- as.numeric(Sys.getpid())
  session_folder <- file.path(folder,
paste0('traitcon_',r,'_ss_',Samplesize,"_ES_",d,"_difCon_",b,"_","session_", pid, "_k_", k))
  if (!dir.exists(session_folder)) {
    dir.create(session_folder)
  }
  copy_folder_contents(folder, session_folder)
  subdirs <- list.dirs(folder, full.names = TRUE, recursive = FALSE)



  #-----------------------------------------------------#
  # draw traits from multivariate normal distribution (0,1),
  # with Cholesky decomposition of covariance matrix
  #-----------------------------------------------------#



  traits <- rmvnorm(n=N, mean=rep(0, factor.ntraits[r]), sigma = trait.cov[[r]],
                    method="chol")
  #-----------------------------------------------------#
  # calculate total score, sums and differences of true traits
  # Used to calculate Bias further on
  #-----------------------------------------------------#

  traits.total <- rowSums(traits)/factor.ntraits[r]

  traits.diff <- traits %*% trait.comp[[r]] #+++++++++ CHANGE Based on condition ++++++++#

  traits.sum <- traits %*% abs(trait.comp[[r]]) #+++++++++ CHANGE Based on condition
++++++++#

  traits.cor.mat <- cor(traits)
```

```
# as vector
traits.cor <- traits.cor.mat[lower.tri(traits.cor.mat)]


#design.mat: design matrix of mfc: rows=pairwise comparisons, cols=items
ni <- ni_x[[r]]
nblocks <- ni/3
design.mat <- design.mat.all[[r]]


#------------------------------------------------------#
# Calculate loadings
# check rank of comparison matrix
# draw loadings until comparison matrix is of
# full rank in all loading shares
#------------------------------------------------------#
full.rank <- NULL
comp.mat <- NULL
load.mat <- NULL

loads <- runif(ni, min=.65, max=.95)


design.load <- design.load.all[[r]]
load.mat[[r]] <- loads*design.load
comp.mat[[r]] <- design.mat %*% load.mat[[r]]
```

```
#-------------------------------------------------------#
#check if rank of comp.mat = number of columns of comp.mat (5)
#save in full rank with index a
#-------------------------------------------------------#

full.rank[r] <- qr(comp.mat[[r]])$rank==ncol(comp.mat[[r]])

#-------------------------------------------------------#
# draw other item parameters, combine to table items
# table: mean, loading, uniqueness
#-------------------------------------------------------#

#-------------------------------------------------------#
# Begin: b is the DIF conditions, Currently modified to only run one condition at a time
#-------------------------------------------------------#

#cond.name <- paste0('_traitCon',r,'_difCon',d)

items <- data.frame(matrix(nrow=ni, ncol=3))
colnames(items) <- c("u.mean","loads","uni")
items$u.mean <- runif(ni, min=-1, max=1)
items$loads <- loads

#uniqueness = 1-load^2
items$uni <- rep(1,ni)
DIF_size = EF_size[d]
g2_items <- items
```

```
# Add DIF
g2_items[items_DIF[[r]][[b]],1] <- g2_items[items_DIF[[r]][[b]],1] + DIF_size

design.load <- design.load.all[[r]]

#----------------------------------------------------#
# simulate data #
#----------------------------------------------------#

data <- tirt.sim(traits, items,itemsg2 = g2_items, design.load, design.mat, nblocks)

#----------------------------------------------------#
# Analyze data using mplus automation
# The models run switch based on trait condition.
#----------------------------------------------------#

#----------------------------------------------------#
# START
# The following function runs all models in the folders set above, then extracts the wald stats
#----------------------------------------------------#



filepath <- file.path(session_folder, "exdat.csv")
write.table(data, filepath, col.names = FALSE, row.names = FALSE, sep = ",")

runModels(session_folder, showOutput = TRUE, recursive = F, Mplus_command = mplusPath)

assign(paste0("sums_", pid), readModels(session_folder, what = "summaries", recursive = F))
```

```
for (i in 1:length(get(paste0("sums_", pid)))) {
  # Get the relevant data
  filename <- get(paste0("sums_", pid))[[i]][["summaries"]][["Filename"]]

  if (grepl("_0DIF", filename, ignore.case = TRUE)) {
    anchorDIFCon <- "0"
  } else if (grepl("50DIF", filename, ignore.case = TRUE)) {
    anchorDIFCon <- "50"
  } else if (grepl("100dif", filename, ignore.case = TRUE)) {
    anchorDIFCon <- "100"
  } else {
    anchorDIFCon = '99'
  }



  blockNumber <- as.integer(gsub(".*block(\\d+).*", "\\1", filename, ignore.case = TRUE))



  if (blockNumber %in% c(8, 12,14,32)) {
    blockType <- "DIF"
  } else {
    blockType <- "NoDIF"
  }

  if (grepl("20A", filename)) {
    anchorCon <- "20"
  } else if (grepl("30A", filename)) {
      anchorCon <- "30"
    }else {
      anchorCon <- '99'
    }



    blockNumber <- gsub(".*block(\\d+).*", "\\1", filename, ignore.case = TRUE)


    waldchisq_value <- get(paste0("sums_", pid))[[i]][[2]][["WaldChiSq_Value"]]
    waldchisq_pvalue <- get(paste0("sums_", pid))[[i]][[2]][["WaldChiSq_PValue"]]


    if ("WaldChiSq_Value" %in% names(get(paste0("sums_", pid))[[i]][[2]])) {
      waldchisq_value <- get(paste0("sums_", pid))[[i]][[2]][["WaldChiSq_Value"]]
    } else {
      cat("WaldChiSq_Value not found for iteration:", i, "\n")
      waldchisq_value <- 9999  # Set to 9999 when not found
    }


    if (waldchisq_value != 9999 && "WaldChiSq_PValue" %in% names(get(paste0("sums_",
pid))[[i]][[2]])) {
      waldchisq_pvalue <- get(paste0("sums_", pid))[[i]][[2]][["WaldChiSq_PValue"]]
    } else {
      waldchisq_pvalue <- 9999  # Set to 9999 when not found or when WaldChiSq_Value is 9999
    }


    # Add this data as a new row to the dataframe
    new_row <- c(Rep = k, TraitCon = r, DifCon = d, PerDIFCon = b, Samplesize = Samplesize,
                Anchorcon = anchorCon, anchorDIFCon = anchorDIFCon, blockNum =
blockNumber, blockType = blockType,
                 WaldChiSq_Value = waldchisq_value, WaldChiSq_PValue = waldchisq_pvalue)


    # Append the new row to the data frame
    rows_list[[i]] <- new_row



  }

  df <- do.call(rbind,rows_list)
  return(df)
}


###### ----- Functions for parallel processing of folders ---- ####

# This function copies the contents of the root folder 'folder' into each process folder
copy_folder_contents <- function(source_folder, destination_folder) {
  # List all files in the source folder
  all_contents <- list.files(source_folder, full.names = TRUE)

  # Filter out directories
  files <- all_contents[!sapply(all_contents, function(path) { isTRUE(file.info(path)$isdir) })]

  # Copy each file to the destination folder
  sapply(files, function(file) {
    file.copy(file, file.path(destination_folder, basename(file)), overwrite = TRUE)
  })
```

```
}


# This function replaces the number of cores used in the mplus analyis
# Function to replace the desired line
replace_line <- function(file_path) {
  lines <- readLines(file_path)
  replacement_line <- paste0("PROCESSORS=", ceiling(ncores/2), ";")
  lines <- gsub("^\\s*PROCESSORS=2;\\s*$", replacement_line, lines)
  writeLines(lines, file_path)
}


######################################################################################
######################################################################################
#


##### Section III. Define Constant Parameters, portions of code from Frick et al., 2021 #######


######################################################################################
######################################################################################
#
#-------------------------------------------------------------------------------------------------------------------------------
------------------------------------------#
# The following parameters are used consistently throughout the simulation.
#-------------------------------------------------------------------------------------------------------------------------------
------------------------------------------#


rows_list <- list()
failure_list <- list()


# Sample size
```

```
Ncon= matrix(c(500,500,
          1000,1000,
          750,250,
          1500,500),
          ncol=4)

# Number of items - 5 trait
nitems.5 = 60
# Number of items - 10 trait
nitems.10 = 120

ni_x = list(nitems.5,nitems.10)

# Group Size Information
g1 = Ncon[1,Samplesize]
g2 = g1 + 1
N = (Ncon[1,Samplesize]+Ncon[2,Samplesize])
#--------------------------------------------------------------------------------------------------#
# Pairs combination (for comparisons. This matrix tells which items to compares when
computing util differences)
#--------------------------------------------------------------------------------#


# For 5 factors
pairs.5 <- combn(5,2)
trait.comp.5 <- matrix(0, nrow=5,ncol=ncol(pairs.5))
for (i in 1:ncol(trait.comp.5)) {
  trait.comp.5[pairs.5[1,i],i] <- 1
  trait.comp.5[pairs.5[2,i],i] <- -1
}

set.seed(123)

# For 10 factors
pairs.10 <- combn(10,2)
trait.comp.10 <- matrix(0, nrow=10,ncol=ncol(pairs.10))
for (i in 1:ncol(trait.comp.10)) {
  trait.comp.10[pairs.10[1,i],i] <- 1
  trait.comp.10[pairs.10[2,i],i] <- -1
}


trait.comp <- list(trait.comp.5,trait.comp.10)

#### simulation design for traits ###

#factor number of traits
factor.ntraits <- c(5,10)

#simulation design for DIF

EF_size <- c(.3,.6)

### Simulation conditions

## 5 Factor
```

```
### 40 %

Blocks_DIF_5f_40 <- c(1,12,15,22,36,44,49,57)

### 50 %

Blocks_DIF_5f_50 <- c(1,12,15,22,36,44,49,57,60,53)

### 60 %

Blocks_DIF_5f_60 <- c(1,12,15,22,36,44,49,57,60,53,29,48)

Blocks_DIF_5f <- list(Blocks_DIF_5f_40,Blocks_DIF_5f_50,Blocks_DIF_5f_60)


## 10 Factor

### 40 %

Blocks_DIF_10f_40 <- c(1,4,7,10,38,40,44,47,26,54,19,59,62,66,83,106)

###  50 %

Blocks_DIF_10f_50 <- c(1,4,7,10,38,40,44,47,26,54,19,59,62,66,83,106,70,73,80,94)

### 60 %

Blocks_DIF_10f_60 <-
c(1,4,7,10,38,40,44,47,26,54,19,59,62,66,83,106,70,73,80,94,86,97,104,114

)


Blocks_DIF_10f <- list(Blocks_DIF_10f_40,Blocks_DIF_10f_50,Blocks_DIF_10f_60)


items_DIF <- list(Blocks_DIF_5f,Blocks_DIF_10f)


####################################### Trait Data ########################


############ Covariance Matricies ############


## The covariance matrix for the 5 trait condition


trait5.cov<- matrix(c(
  1,-.36,-.17,-.36,-.43,
  -.36,1,.43,.26,.29,
  -.17,.43,1,.21,.20,
  -.36,.26,.21,1,.43,
  -.43,.29,.20,.43,1),
  nrow=5,ncol=5)

#-----------------------------------------------------------------------------------------#
# The covariance matrix for the 10 trait condition, generated using code from Frick et al, 2021
# covariance matrix for inverse Wishart: all covariances set to .3
#---------------------------------------------------------------------------------------------------------------
#

# Done once, kept consistent with set.seed(123) generation
#cov10 <- diag(10)
#cov10[cov10==0] <- .3

#draw from inverse Wishart with df=100

#trait10.cov <- cov2cor(SimDesign::rinvWishart(1, 100, cov10))

#reverse traits 1, 6, 10
#trait10.cov[,1][trait10.cov[,1]!=1] <- trait10.cov[,1][trait10.cov[,1]!=1]*-1
#trait10.cov[1,][trait10.cov[1,]!=1] <- trait10.cov[1,][trait10.cov[1,]!=1]*-1
#trait10.cov[,6][trait10.cov[,6]!=1] <- trait10.cov[,6][trait10.cov[,6]!=1]*-1
#trait10.cov[6,][trait10.cov[6,]!=1] <- trait10.cov[6,][trait10.cov[6,]!=1]*-1

trait10.cov <- matrix(
  c(
    1.00, -0.44, -0.31, -0.16, -0.06,  0.24, -0.40, -0.31, -0.34, -0.41,
    -0.44,  1.00,  0.35,  0.42,  0.28, -0.33,  0.41,  0.34,  0.42,  0.40,
    -0.31,  0.35,  1.00,  0.16,  0.26, -0.36,  0.45,  0.22,  0.30,  0.28,
    -0.16,  0.42,  0.16,  1.00,  0.32, -0.29,  0.22,  0.21,  0.24,  0.30,
    -0.06,  0.28,  0.26,  0.32,  1.00, -0.26,  0.28,  0.19,  0.29,  0.16,
    0.24, -0.33, -0.36, -0.29, -0.26,  1.00, -0.37, -0.22, -0.22, -0.23,
    -0.40,  0.41,  0.45,  0.22,  0.28, -0.37,  1.00,  0.27,  0.38,  0.29,
    -0.31,  0.34,  0.22,  0.21,  0.19, -0.22,  0.27,  1.00,  0.32,  0.41,
    -0.34,  0.42,  0.30,  0.24,  0.29, -0.22,  0.38,  0.32,  1.00,  0.34,
    -0.41,  0.40,  0.28,  0.30,  0.16, -0.23,  0.29,  0.41,  0.34,  1.00
  ),
  nrow = 10,
  byrow = TRUE
)


### Combine to a single list


trait.cov <- list(trait5.cov,trait10.cov)


#design.block: part of design matrix for one block
design.block <- matrix(c(1,-1,0,1,0,-1,0,1,-1),
                       nrow=3,ncol=3,byrow=TRUE)
#design matrix of mfc: rows=pairwise comparisons, cols=items
design.mat.5 <- NULL
design.mat.10 <- NULL
  design.mat.5 <- matrix(rep(0,nitems.5^2),nrow=nitems.5,ncol=nitems.5)
  #fill blocks in design.mat with block design
  for (i in 1:(nitems.5/3)) {
    design.mat.5[(3*i-2):(3*i),(3*i-2):(3*i)] <- design.block
  }
  design.mat.10 <- matrix(rep(0,nitems.10^2),nrow=nitems.10,ncol=nitems.10)
  #fill blocks in design.mat with block design
  for (i in 1:(nitems.10/3)) {
    design.mat.10[(3*i-2):(3*i),(3*i-2):(3*i)] <- design.block
  }

design.mat.all <- list(design.mat.5,design.mat.10)
############## Design matrix for loadings ####################


#-----------------------------------------------------------------------------------------------------------#
# This is an object that sets the direction of the loading, -1, 0, and 1s indicating a negative,
# no loading or positive loading.  25% of the blocks are negative, 1 loading per trait in 5 factor 2 loadings per trait in 10
#--------------------------------------------------------------------------------------------------------------#
design_load_5_factor <-matrix(c(
  -1,0,0,0,0,
  0,1,0,0,0,
  0,0,1,0,0,
  1,0,0,0,0,
  0,-1,0,0,0,
  0,0,0,1,0,
  1,0,0,0,0,
  0,1,0,0,0,
  0,0,0,0,-1,
  0,0,0,0,1,
  0,0,1,0,0,
  0,0,0,-1,0,
  0,0,0,1,0,
  0,0,-1,0,0,
  0,0,0,0,1,
  1,0,0,0,0,
  0,0,0,1,0,
  0,0,0,0,1,
  0,1,0,0,0,
  0,0,1,0,0,
  0,0,0,1,0,
  0,1,0,0,0,
  0,0,1,0,0,
  0,0,0,0,1,
  0,1,0,0,0,
  1,0,0,0,0,
  0,0,0,0,1,
  0,0,1,0,0,
  0,0,0,1,0,
  0,0,0,0,1,
  1,0,0,0,0,
  0,1,0,0,0,
  0,0,1,0,0,
  1,0,0,0,0,
  0,1,0,0,0,
  0,0,0,1,0,
  1,0,0,0,0,
  0,1,0,0,0,
  0,0,0,0,1,
    1,0,0,0,0,
    0,0,1,0,0,
    0,0,0,1,0,
    1,0,0,0,0,
    0,0,1,0,0,
    0,0,0,0,1,
    1,0,0,0,0,
    0,0,0,1,0,
    0,0,0,0,1,
    0,1,0,0,0,
    0,0,1,0,0,
    0,0,0,1,0,
    0,1,0,0,0,
    0,0,1,0,0,
    0,0,0,0,1,
    0,1,0,0,0,
    0,0,0,1,0,
    0,0,0,0,1,
    0,0,1,0,0,
    0,0,0,1,0,
    1,0,0,0,0
  ),nrow=60, byrow=TRUE)

  design_load_10_factor <-matrix(c(
    -1,0,0,0,0,0,0,0,0,0,   0,1,0,0,0,0,0,0,0,0,     0,0,1,0,0,0,0,0,0,0,
    0,0,0,-1,0,0,0,0,0,0,   0,0,0,0,1,0,0,0,0,0,     0,0,0,0,0,1,0,0,0,0,
    0,0,0,0,0,0,-1,0,0,0,   0,0,0,0,0,0,0,1,0,0,     0,0,0,0,0,0,0,0,1,0,
    0,0,0,0,0,0,0,0,0,-1,   1,0,0,0,0,0,0,0,0,0,     0,0,0,1,0,0,0,0,0,0,
  0,-1,0,0,0,0,0,0,0,0,   0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,0,1,0,0,0,
  0,0,1,0,0,0,0,0,0,0,    0,0,0,0,0,1,0,0,0,0,   0,0,0,0,0,0,0,-1,0,0,
  0,0,0,0,0,0,0,0,-1,0,   0,0,0,0,0,0,0,0,0,1,   0,1,0,0,0,0,0,0,0,0,
  0,0,-1,0,0,0,0,0,0,0,   0,0,0,1,0,0,0,0,0,0,   0,0,0,0,0,0,1,0,0,0,
  0,0,0,0,-1,0,0,0,0,0,   0,0,0,0,0,0,0,1,0,0,   0,0,0,0,0,0,0,0,0,1,
  0,0,0,0,0,-1,0,0,0,0,   0,0,0,0,0,0,0,0,1,0,   1,0,0,0,0,0,0,0,0,0,
  1,0,0,0,0,0,0,0,0,0,    0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,1,0,0,0,0,
  1,0,0,0,0,0,0,0,0,0,    0,0,0,0,0,0,1,0,0,0,   0,0,0,0,0,0,0,1,0,0,
  0,1,0,0,0,0,0,0,0,0,    0,0,1,0,0,0,0,0,0,0,   0,0,0,1,0,0,0,0,0,0,
  0,1,0,0,0,0,0,0,0,0,    0,0,0,0,0,1,0,0,0,0,   0,0,0,0,0,0,0,1,0,0,
  0,0,1,0,0,0,0,0,0,0,    0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,0,0,0,1,0,
  0,0,1,0,0,0,0,0,0,0,    0,0,0,0,0,1,0,0,0,0,   0,0,0,0,0,0,0,0,0,1,
  0,0,0,1,0,0,0,0,0,0,    0,0,0,0,0,0,0,1,0,0,   0,0,0,0,0,0,0,0,1,0,
  0,0,0,0,0,1,0,0,0,0,    0,0,0,0,0,0,1,0,0,0,   0,0,0,0,0,0,0,0,0,1,
  1,0,0,0,0,0,0,0,0,0,    0,0,0,0,0,0,0,0,1,0,   0,1,0,0,0,0,0,0,0,0,
  1,0,0,0,0,0,0,0,0,0,    0,0,0,1,0,0,0,0,0,0,   0,0,0,0,0,0,1,0,0,0,
  0,1,0,0,0,0,0,0,0,0,    0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,0,0,0,0,1,
  0,0,1,0,0,0,0,0,0,0,    0,0,0,0,0,0,1,0,0,0,   0,0,0,0,0,0,0,1,0,0,
  0,0,0,1,0,0,0,0,0,0,    0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,1,0,0,0,0,
  1,0,0,0,0,0,0,0,0,0,    0,0,1,0,0,0,0,0,0,0,   0,0,0,0,0,0,0,0,0,1,
  0,1,0,0,0,0,0,0,0,0,    0,0,0,0,0,0,1,0,0,0,   0,0,0,0,0,0,0,1,0,0,
  0,0,0,1,0,0,0,0,0,0,    0,0,1,0,0,0,0,0,0,0,   0,0,0,0,0,0,0,0,0,1,
  0,0,0,0,1,0,0,0,0,0,    0,0,0,0,0,0,1,0,0,0,   0,0,0,0,0,0,0,0,1,0,
  1,0,0,0,0,0,0,0,0,0,    0,0,0,0,0,1,0,0,0,0,   0,0,0,0,0,0,0,0,1,0,
  0,1,0,0,0,0,0,0,0,0,    0,0,0,0,0,0,0,0,1,0,   0,0,0,0,0,0,1,0,0,0,
  0,0,0,1,0,0,0,0,0,0,    0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,0,0,1,0,0,
  0,0,0,0,1,0,0,0,0,0,    0,1,0,0,0,0,0,0,0,0,   0,0,0,1,0,0,0,0,0,0,
  0,0,1,0,0,0,0,0,0,0,    0,0,0,0,0,1,0,0,0,0,   0,0,0,0,0,0,0,1,0,0,
  0,0,0,0,1,0,0,0,0,0,   0,0,0,0,0,0,0,1,0,0,     0,0,0,0,0,0,0,0,0,1,
  1,0,0,0,0,0,0,0,0,0,   0,0,0,0,0,0,1,0,0,0,     0,0,0,0,0,0,0,0,0,1,
  0,1,0,0,0,0,0,0,0,0,   0,0,0,0,0,1,0,0,0,0,     0,0,0,0,0,0,0,0,1,0,
  0,0,0,0,0,0,0,0,1,0,   0,0,0,1,0,0,0,0,0,0,     0,0,0,0,0,0,0,1,0,0,
  1,0,0,0,0,0,0,0,0,0,   0,0,1,0,0,0,0,0,0,0,     0,0,0,0,0,0,0,0,0,1,
  1,0,0,0,0,0,0,0,0,0,   0,0,1,0,0,0,0,0,0,0,     0,0,0,0,0,0,0,0,0,1,
  0,1,0,0,0,0,0,0,0,0,   0,0,0,1,0,0,0,0,0,0,     0,0,0,0,0,0,0,0,1,0,
  0,0,0,0,0,1,0,0,0,0,   0,0,0,0,1,0,0,0,0,0,     0,0,0,0,0,0,1,0,0,0

),
                        nrow=120, byrow=TRUE)



design.load.all = list('5'= design_load_5_factor,'10'=design_load_10_factor)


#################  Parallel processing #############
ncores <- detectCores()
ncores = ncores - 1
### Changing the cores used by each .inp file to make use of locally available maximums (each analysis file uses ncores/2)
# Path to the directory containing the .inp files
path_to_files <- folder


# List all .inp files in the directory
files <- list.files(path_to_files, pattern = "\\.inp$", full.names = TRUE)


# Apply the function to each .inp file
#lapply(files, replace_line)
```

```
#set seed and ranges for factors
plan(multicore)


####################### Section IV.  Simulation ###########
#----------------------------------------------------#
# Begin Simulation
#----------------------------------------------------#
start.time = Sys.time()
df.all <- future_map2(1:k.range,seeds, run_simulation, .progress = T)
end.time = Sys.time()
print(end.time-start.time)
#### Save the DF to Results
df.final <- as.data.frame(do.call(rbind, df.all))
df.final$Rep <- as.numeric(df.final$Rep)
df.final$TraitCon <- as.numeric(df.final$TraitCon)
df.final$DifCon <- as.numeric(df.final$DifCon)
df.final$PerDIFCon <- as.numeric(df.final$PerDIFCon)
df.final$WaldChiSq_Value <- as.numeric(df.final$WaldChiSq_Value)
df.final$WaldChiSq_PValue <- as.numeric(df.final$WaldChiSq_PValue)
df.final$Anchorcon <- as.character(df.final$Anchorcon)
df.final$anchorDIFCon <- as.character(df.final$anchorDIFCon)
df.final$blockNum <- as.character(df.final$blockNum)
df.final$blockType <- as.character(df.final$blockType)
cond.name = paste0('samplesize_',Samplesize,'_traitsize_',r,'_ES_',d,'_DIFItems_',b,'_',approach,'_anc_',size,'msfold',ld)
results_path <- file.path(source,'Results//')
write.csv(df.final, file = paste0(results_path, cond.name, ".csv"))
```

## Appendix 2 – Example Mplus 5-Factor Free-Baseline Model (20 Anchor, 0DIF)

```
TITLE: TEST

 DATA: FILE IS 'exdat.csv';

 VARIABLE:
 Names ARE
 group
 i1i2 i1i3 i2i3
 i4i5 i4i6 i5i6
 i7i8 i7i9 i8i9
 i10i11 i10i12 i11i12
 i13i14 i13i15 i14i15
 i16i17 i16i18 i17i18
 i19i20 i19i21 i20i21
 i22i23 i22i24 i23i24
 i25i26 i25i27 i26i27
 i28i29 i28i30 i29i30
 i31i32 i31i33 i32i33
 i34i35 i34i36 i35i36
 i37i38 i37i39 i38i39
 i40i41 i40i42 i41i42
 i43i44 i43i45 i44i45
 i46i47 i46i48 i47i48
 i49i50 i49i51 i50i51
 i52i53 i52i54 i53i54
 i55i56 i55i57 i56i57
 i58i59 i58i60 i59i60;
 !USEOBS = (group == 1);
 USEVARIABLES ARE i1i2-i59i60;
 CATEGORICAL ARE ALL;
 GROUPING IS group(0=G1, 1=G2)

 ANALYSIS:
 ESTIMATOR = ulsmv;
 PARAMETERIZATION = theta;
 PROCESSORS=2;

 MODEL:
 ! Common to both groups
 !The utilities
 ! Binary outcomes are determined by the item utilities (first-order factors)
t1  BY  i1i2@1; t2  BY  i1i2@-1;
t1  BY  i1i3@1; t3  BY  i1i3@-1;
t2  BY  i2i3@1; t3  BY  i2i3@-1;
t4  BY  i4i5@1; t5  BY  i4i5@-1;
t4  BY  i4i6@1; t6  BY  i4i6@-1;
t5  BY  i5i6@1; t6  BY  i5i6@-1;
t7  BY  i7i8@1; t8  BY  i7i8@-1;
t7  BY  i7i9@1; t9  BY  i7i9@-1;
t8  BY  i8i9@1; t9  BY  i8i9@-1;
t10  BY  i10i11@1; t11  BY  i10i11@-1;
t10  BY  i10i12@1; t12  BY  i10i12@-1;
t11  BY  i11i12@1; t12  BY  i11i12@-1;
```

```
t13  BY  i13i14@1; t14  BY  i13i14@-1;
t13  BY  i13i15@1; t15  BY  i13i15@-1;
t14  BY  i14i15@1; t15  BY  i14i15@-1;
t16  BY  i16i17@1; t17  BY  i16i17@-1;
t16  BY  i16i18@1; t18  BY  i16i18@-1;
t17  BY  i17i18@1; t18  BY  i17i18@-1;
t19  BY  i19i20@1; t20  BY  i19i20@-1;
t19  BY  i19i21@1; t21  BY  i19i21@-1;
t20  BY  i20i21@1; t21  BY  i20i21@-1;
t22  BY  i22i23@1; t23  BY  i22i23@-1;
t22  BY  i22i24@1; t24  BY  i22i24@-1;
t23  BY  i23i24@1; t24  BY  i23i24@-1;
t25  BY  i25i26@1; t26  BY  i25i26@-1;
t25  BY  i25i27@1; t27  BY  i25i27@-1;
t26  BY  i26i27@1; t27  BY  i26i27@-1;
t28  BY  i28i29@1; t29  BY  i28i29@-1;
t28  BY  i28i30@1; t30  BY  i28i30@-1;
t29  BY  i29i30@1; t30  BY  i29i30@-1;
t31  BY  i31i32@1; t32  BY  i31i32@-1;
t31  BY  i31i33@1; t33  BY  i31i33@-1;
t32  BY  i32i33@1; t33  BY  i32i33@-1;
t34  BY  i34i35@1; t35  BY  i34i35@-1;
t34  BY  i34i36@1; t36  BY  i34i36@-1;
t35  BY  i35i36@1; t36  BY  i35i36@-1;
t37  BY  i37i38@1; t38  BY  i37i38@-1;
t37  BY  i37i39@1; t39  BY  i37i39@-1;
t38  BY  i38i39@1; t39  BY  i38i39@-1;
t40  BY  i40i41@1; t41  BY  i40i41@-1;
```

```
t40  BY  i40i42@1; t42  BY  i40i42@-1;
t41  BY  i41i42@1; t42  BY  i41i42@-1;
t43  BY  i43i44@1; t44  BY  i43i44@-1;
t43  BY  i43i45@1; t45  BY  i43i45@-1;
t44  BY  i44i45@1; t45  BY  i44i45@-1;
t46  BY  i46i47@1; t47  BY  i46i47@-1;
t46  BY  i46i48@1; t48  BY  i46i48@-1;
t47  BY  i47i48@1; t48  BY  i47i48@-1;
t49  BY  i49i50@1; t50  BY  i49i50@-1;
t49  BY  i49i51@1; t51  BY  i49i51@-1;
t50  BY  i50i51@1; t51  BY  i50i51@-1;
t52  BY  i52i53@1; t53  BY  i52i53@-1;
t52  BY  i52i54@1; t54  BY  i52i54@-1;
t53  BY  i53i54@1; t54  BY  i53i54@-1;
t55  BY  i55i56@1; t56  BY  i55i56@-1;
t55  BY  i55i57@1; t57  BY  i55i57@-1;
t56  BY  i56i57@1; t57  BY  i56i57@-1;
t58  BY  i58i59@1; t59  BY  i58i59@-1;
t58  BY  i58i60@1; t60  BY  i58i60@-1;
t59  BY  i59i60@1; t60  BY  i59i60@-1;

! Errors of binary outcomes are zero
i1i2-i59i60@0;

! Utilities are caused by attributes (second-order factors)

Trait1  BY
t1*-1
```

```
t4*1 (L4)
t7*1 (L7)
t16*1 (L16)
t26*1
t31*1
t34*1
t37*1
t40*1
t43*1
t46*1
t60*1;

Trait2  BY
t2*1
t5*-1 (L5)
t8*1 (L8)
t19*1 (L19)
t22*1
t25*1
t32*1
t35*1
t38*1
t49*1
t52*1
t55*1;

Trait3  BY
t3*1
```

```
t11*1
t14*-1
t20*1 (L20)
t23*1
t28*1
t33*1
t41*1
t44*1
t50*1
t53*1
t58*1;

Trait4  BY
t6*1 (L6)
t12*-1
t13*1
t17*1 (L17)
t21*1 (L21)
t29*1
t36*1
t42*1
t47*1
t51*1
t56*1
t59*1;

Trait5  BY
t9*-1 (L9)
```

```
t10*1
t15*1
t18*1 (L18)
t24*1
t27*1
t30*1
t39*1
t45*1
t48*1
t54*1
t57*1;



MODEL G1:
! variances for all traits are set to 1
Trait1-Trait5@1;
[Trait1-Trait5@0];

!Binary outcomes' errors are zero since this is ranking
i1i2-i59i60@0;

! Utility errors are set to 1
t1-t60@1;
! pairwise thresholds are fixed to 0
[i1i2$1-i59i60$1@0];
```

```
! and utilities means are free except anchor items and one item per block for identification
[t1*] (M1);  [t2*] (M2);  [t3@0] (M3);
[t4*] (M4);  [t5*] (M5);  [t6@0] (M6);
[t7*] (M7);  [t8*] (M8);  [t9@0] (M9);
[t10*] (M10);  [t11@0] (M11);  [t12*] (M12);
[t13*] (M13);  [t14@0] (M14);  [t15*] (M15);
[t16*] (M16);  [t17*] (M17);  [t18@0] (M18);
[t19*] (M19);  [t20*] (M20);  [t21@0] (M21);
[t22*] (M22);  [t23*] (M23);  [t24@0] (M24);
[t25*] (M25);  [t26*] (M26);  [t27@0] (M27);
[t28*] (M28);  [t29*] (M29);  [t30@0] (M30);
[t31*] (M31);  [t32*] (M32);  [t33@0] (M33);
[t34*] (M34);  [t35@0] (M35);  [t36*] (M36);
[t37*] (M37);  [t38*] (M38);  [t39@0] (M39);
[t40*] (M40);  [t41*] (M41);  [t42@0] (M42);
[t43*] (M43);  [t44*] (M44);  [t45@0] (M45);
[t46*] (M46);  [t47@0] (M47);  [t48*] (M48);
[t49*] (M49);  [t50*] (M50);  [t51@0] (M51);
[t52*] (M52);  [t53*] (M53);  [t54@0] (M54);
[t55*] (M55);  [t56@0] (M56);  [t57*] (M57);
[t58*] (M58);  [t59@0] (M59);  [t60*] (M60);


MODEL G2:


Trait1-Trait5*1;
[Trait1-Trait5*0];
```

```
!Binary outcomes' errors are zero since this is ranking
i1i2-i59i60@0;


! Utility errors are set to 1
t1-t60@1;
! pairwise thresholds are fixed to 0
[i1i2$1-i59i60$1@0];


! and utilities means are free except anchor items and one item per block for identification
(specifcied such it is never the DIF item)
[t1*] (N1);  [t2*] (N2);  [t3@0] (N3);


[t4*] (M4);  [t5*] (M5);  [t6@0] (M6);
[t7*] (M7);  [t8*] (M8);  [t9@0] (M9);


[t10*] (N10);  [t11@0] (N11);  [t12*] (N12);
[t13*] (N13);  [t14@0] (N14);  [t15*] (N15);


[t16*] (M16);  [t17*] (M17);  [t18@0] (M18);
[t19*] (M19);  [t20*] (M20);  [t21@0] (M21);


[t22*] (N22);  [t23*] (N23);  [t24@0] (N24);
[t25*] (N25);  [t26*] (N26);  [t27@0] (N27);
[t28*] (N28);  [t29*] (N29);  [t30@0] (N30);
[t31*] (N31);  [t32*] (N32);  [t33@0] (N33);
[t34*] (N34);  [t35@0] (N35);  [t36*] (N36);
[t37*] (N37);  [t38*] (N38);  [t39@0] (N39);
[t40*] (N40);  [t41*] (N41);  [t42@0] (N42);
[t43*] (N43);  [t44*] (N44);  [t45@0] (N45);
```

```
[t46*] (N46);  [t47@0] (N47);  [t48*] (N48);
[t49*] (N49);  [t50*] (N50);  [t51@0] (N51);
[t52*] (N52);  [t53*] (N53);  [t54@0] (N54);
[t55*] (N55);  [t56@0] (N56);  [t57*] (N57);
[t58*] (N58);  [t59@0] (N59);  [t60*] (N60);

MODEL TEST:
! Block 8DIF
!Testing UNIFORM
0=M22-N22;
0=M23-N23;
```

## Appendix 3 – Example Mplus 10-Factor Free-Baseline Model (20 Anchor, 0DIF)

```
 DATA: FILE IS 'exdat.csv';

 VARIABLE:
 Names ARE
 group
 i1i2 i1i3 i2i3
 i4i5 i4i6 i5i6
 i7i8 i7i9 i8i9
 i10i11 i10i12 i11i12
 i13i14 i13i15 i14i15
 i16i17 i16i18 i17i18
 i19i20 i19i21 i20i21
 i22i23 i22i24 i23i24
 i25i26 i25i27 i26i27
 i28i29 i28i30 i29i30
 i31i32 i31i33 i32i33
 i34i35 i34i36 i35i36
 i37i38 i37i39 i38i39
 i40i41 i40i42 i41i42
 i43i44 i43i45 i44i45
 i46i47 i46i48 i47i48
 i49i50 i49i51 i50i51
 i52i53 i52i54 i53i54
 i55i56 i55i57 i56i57
 i58i59 i58i60 i59i60
 i61i62 i61i63 i62i63
i64i65 i64i66 i65i66
```

```
i67i68 i67i69 i68i69
i70i71 i70i72 i71i72
i73i74 i73i75 i74i75
i76i77 i76i78 i77i78
i79i80 i79i81 i80i81
i82i83 i82i84 i83i84
i85i86 i85i87 i86i87
i88i89 i88i90 i89i90
i91i92 i91i93 i92i93
i94i95 i94i96 i95i96
i97i98 i97i99 i98i99
i100i101 i100i102 i101i102
i103i104 i103i105 i104i105
i106i107 i106i108 i107i108
i109i110 i109i111 i110i111
i112i113 i112i114 i113i114
i115i116 i115i117 i116i117
i118i119 i118i120 i119i120;


  !USEOBS = (group == 1);
  USEVARIABLES ARE i1i2-i119i120;
  CATEGORICAL ARE ALL;
  GROUPING IS group(0=G1, 1=G2)

  ANALYSIS:
  ESTIMATOR = ulsmv;
  PARAMETERIZATION = theta;
```

```
PROCESSORS=2;

OUTPUT:

STANDARDIZED

MODEL:
! Common to both groups
!The utilities
t1  BY  i1i2@1; t2  BY  i1i2@-1;
t1  BY  i1i3@1; t3  BY  i1i3@-1;
t2  BY  i2i3@1; t3  BY  i2i3@-1;
t4  BY  i4i5@1; t5  BY  i4i5@-1;
t4  BY  i4i6@1; t6  BY  i4i6@-1;
t5  BY  i5i6@1; t6  BY  i5i6@-1;
t7  BY  i7i8@1; t8  BY  i7i8@-1;
t7  BY  i7i9@1; t9  BY  i7i9@-1;
t8  BY  i8i9@1; t9  BY  i8i9@-1;
t10  BY  i10i11@1; t11  BY  i10i11@-1;
t10  BY  i10i12@1; t12  BY  i10i12@-1;
t11  BY  i11i12@1; t12  BY  i11i12@-1;
t13  BY  i13i14@1; t14  BY  i13i14@-1;
t13  BY  i13i15@1; t15  BY  i13i15@-1;
t14  BY  i14i15@1; t15  BY  i14i15@-1;
t16  BY  i16i17@1; t17  BY  i16i17@-1;
t16  BY  i16i18@1; t18  BY  i16i18@-1;
t17  BY  i17i18@1; t18  BY  i17i18@-1;
t19  BY  i19i20@1; t20  BY  i19i20@-1;
```

```
t19  BY  i19i21@1; t21  BY  i19i21@-1;
t20  BY  i20i21@1; t21  BY  i20i21@-1;
t22  BY  i22i23@1; t23  BY  i22i23@-1;
t22  BY  i22i24@1; t24  BY  i22i24@-1;
t23  BY  i23i24@1; t24  BY  i23i24@-1;
t25  BY  i25i26@1; t26  BY  i25i26@-1;
t25  BY  i25i27@1; t27  BY  i25i27@-1;
t26  BY  i26i27@1; t27  BY  i26i27@-1;
t28  BY  i28i29@1; t29  BY  i28i29@-1;
t28  BY  i28i30@1; t30  BY  i28i30@-1;
t29  BY  i29i30@1; t30  BY  i29i30@-1;
t31  BY  i31i32@1; t32  BY  i31i32@-1;
t31  BY  i31i33@1; t33  BY  i31i33@-1;
t32  BY  i32i33@1; t33  BY  i32i33@-1;
t34  BY  i34i35@1; t35  BY  i34i35@-1;
t34  BY  i34i36@1; t36  BY  i34i36@-1;
t35  BY  i35i36@1; t36  BY  i35i36@-1;
t37  BY  i37i38@1; t38  BY  i37i38@-1;
t37  BY  i37i39@1; t39  BY  i37i39@-1;
t38  BY  i38i39@1; t39  BY  i38i39@-1;
t40  BY  i40i41@1; t41  BY  i40i41@-1;
t40  BY  i40i42@1; t42  BY  i40i42@-1;
t41  BY  i41i42@1; t42  BY  i41i42@-1;
t43  BY  i43i44@1; t44  BY  i43i44@-1;
t43  BY  i43i45@1; t45  BY  i43i45@-1;
t44  BY  i44i45@1; t45  BY  i44i45@-1;
t46  BY  i46i47@1; t47  BY  i46i47@-1;
t46  BY  i46i48@1; t48  BY  i46i48@-1;
```

```
 t47  BY  i47i48@1; t48  BY  i47i48@-1;
 t49  BY  i49i50@1; t50  BY  i49i50@-1;
 t49  BY  i49i51@1; t51  BY  i49i51@-1;
 t50  BY  i50i51@1; t51  BY  i50i51@-1;
 t52  BY  i52i53@1; t53  BY  i52i53@-1;
 t52  BY  i52i54@1; t54  BY  i52i54@-1;
 t53  BY  i53i54@1; t54  BY  i53i54@-1;
 t55  BY  i55i56@1; t56  BY  i55i56@-1;
 t55  BY  i55i57@1; t57  BY  i55i57@-1;
 t56  BY  i56i57@1; t57  BY  i56i57@-1;
 t58  BY  i58i59@1; t59  BY  i58i59@-1;
 t58  BY  i58i60@1; t60  BY  i58i60@-1;
 t59  BY  i59i60@1; t60  BY  i59i60@-1;
t61  BY  i61i62@1; t62  BY  i61i62@-1;
t61  BY  i61i63@1; t63  BY  i61i63@-1;
t62  BY  i62i63@1; t63  BY  i62i63@-1;
t64  BY  i64i65@1; t65  BY  i64i65@-1;
t64  BY  i64i66@1; t66  BY  i64i66@-1;
t65  BY  i65i66@1; t66  BY  i65i66@-1;
t67  BY  i67i68@1; t68  BY  i67i68@-1;
t67  BY  i67i69@1; t69  BY  i67i69@-1;
t68  BY  i68i69@1; t69  BY  i68i69@-1;
t70  BY  i70i71@1; t71  BY  i70i71@-1;
t70  BY  i70i72@1; t72  BY  i70i72@-1;
t71  BY  i71i72@1; t72  BY  i71i72@-1;
t73  BY  i73i74@1; t74  BY  i73i74@-1;
t73  BY  i73i75@1; t75  BY  i73i75@-1;
t74  BY  i74i75@1; t75  BY  i74i75@-1;
```

```
t76  BY  i76i77@1; t77  BY  i76i77@-1;
t76  BY  i76i78@1; t78  BY  i76i78@-1;
t77  BY  i77i78@1; t78  BY  i77i78@-1;
t79  BY  i79i80@1; t80  BY  i79i80@-1;
t79  BY  i79i81@1; t81  BY  i79i81@-1;
t80  BY  i80i81@1; t81  BY  i80i81@-1;
t82  BY  i82i83@1; t83  BY  i82i83@-1;
t82  BY  i82i84@1; t84  BY  i82i84@-1;
t83  BY  i83i84@1; t84  BY  i83i84@-1;
t85  BY  i85i86@1; t86  BY  i85i86@-1;
t85  BY  i85i87@1; t87  BY  i85i87@-1;
t86  BY  i86i87@1; t87  BY  i86i87@-1;
t88  BY  i88i89@1; t89  BY  i88i89@-1;
t88  BY  i88i90@1; t90  BY  i88i90@-1;
t89  BY  i89i90@1; t90  BY  i89i90@-1;
t91  BY  i91i92@1; t92  BY  i91i92@-1;
t91  BY  i91i93@1; t93  BY  i91i93@-1;
t92  BY  i92i93@1; t93  BY  i92i93@-1;
t94  BY  i94i95@1; t95  BY  i94i95@-1;
t94  BY  i94i96@1; t96  BY  i94i96@-1;
t95  BY  i95i96@1; t96  BY  i95i96@-1;
t97  BY  i97i98@1; t98  BY  i97i98@-1;
t97  BY  i97i99@1; t99  BY  i97i99@-1;
t98  BY  i98i99@1; t99  BY  i98i99@-1;
t100  BY  i100i101@1; t101  BY  i100i101@-1;
t100  BY  i100i102@1; t102  BY  i100i102@-1;
t101  BY  i101i102@1; t102  BY  i101i102@-1;
t103  BY  i103i104@1; t104  BY  i103i104@-1;
```

```
t103  BY  i103i105@1; t105  BY  i103i105@-1;
t104  BY  i104i105@1; t105  BY  i104i105@-1;
t106  BY  i106i107@1; t107  BY  i106i107@-1;
t106  BY  i106i108@1; t108  BY  i106i108@-1;
t107  BY  i107i108@1; t108  BY  i107i108@-1;
t109  BY  i109i110@1; t110  BY  i109i110@-1;
t109  BY  i109i111@1; t111  BY  i109i111@-1;
t110  BY  i110i111@1; t111  BY  i110i111@-1;
t112  BY  i112i113@1; t113  BY  i112i113@-1;
t112  BY  i112i114@1; t114  BY  i112i114@-1;
t113  BY  i113i114@1; t114  BY  i113i114@-1;
t115  BY  i115i116@1; t116  BY  i115i116@-1;
t115  BY  i115i117@1; t117  BY  i115i117@-1;
t116  BY  i116i117@1; t117  BY  i116i117@-1;
t118  BY  i118i119@1; t119  BY  i118i119@-1;
t118  BY  i118i120@1; t120  BY  i118i120@-1;
t119  BY  i119i120@1; t120  BY  i119i120@-1;

! Errors of binary outcomes are zero
i1i2-i119i120@0;

! Utilities are caused by attributes (second-order factors)

Trait1  BY
t1*-1
t11*1
t30*1 (L30)
t31*1 (L31)
```

```
t34*1 (L34)
t55*1
t58*1
t70*1
t82*1
t100*1
t109*1
t112*1;

Trait2  BY
t2*1
t13*-1 (L13)
t21*1
t37*1
t40*1
t57*1
t61*1
t73*1
t85*1
t92*1
t103*1
t115*1;

Trait3  BY
t3*1
t16*1 (L16)
t22*-1 (L22)
t38*1
```

```
t43*1
t46*1
t64*1
t71*1
t77*1 (L77)
t94*1
t110*1
t113*1;

Trait4  BY
t4*-1
t12*1
t23*1 (L23)
t39*1
t49*1
t59*1
t67*1 (L67)
t76*1 (L76)
t88*1
t93*1
t107*1
t116*1;

Trait5  BY
t5*1
t14*1 (L14)
t25*-1
t32*1 (L32)
```

```
t44*1
t62*1
t68*1 (L68)
t79*1
t89*1
t91*1
t97*1
t119*1;

Trait6  BY
t6*1
t17*1 (L17)
t28*-1 (L28)
t33*1 (L33)
t41*1
t47*1
t52*1
t69*1 (L69)
t83*1
t95*1
t104*1
t118*1;

Trait7  BY
t7*-1
t15*1 (L15)
t24*1 (L24)
t35*1 (L35)
```

```
t53*1
t60*1
t65*1
t74*1
t80*1
t87*1
t101*1
t120*1;

Trait8  BY
t8*1
t18*-1 (L18)
t26*1
t36*1 (L36)
t42*1
t50*1
t66*1
t75*1
t90*1
t96*1
t98*1
t108*1;

Trait9  BY
t9*1
t19*-1
t29*1 (L29)
t45*1
```

```
t51*1
t56*1
t81*1
t84*1
t86*1
t105*1
t106*1
t117*1;

Trait10  BY
t10*-1
t20*1
t27*1
t48*1
t54*1
t63*1
t72*1
t78*1 (L78)
t99*1
t102*1
t111*1
t114*1;


 MODEL G1:
 ! variances for all traits are set to 1
 Trait1-Trait10@1;
 ! and means to 0
```

```
  [Trait1-Trait10@0];

  !Binary outcomes' errors are zero since this is ranking
  i1i2-i119i120@0;


  ! pairwise thresholds are fixed to 0
  [i1i2$1-i119i120$1@0];

! Utility errors are set to 1
t1-t120@1;

! and utilities means are free except anchor items and one item per block for identification

[t1*] (M1);  [t2*] (M2);  [t3@0] (M3);
[t4*] (M4);  [t5*] (M5);  [t6@0] (M6);
[t7*] (M7);  [t8*] (M8);  [t9@0] (M9);
[t10*] (M10);  [t11*] (M11);  [t12@0] (M12);
[t13*] (M13);  [t14*] (M14);  [t15@0] (M15);
[t16*] (M16);  [t17*] (M17);  [t18@0] (M18);
[t19*] (M19);  [t20*] (M20);  [t21@0] (M21);
[t22*] (M22);  [t23*] (M23);  [t24@0] (M24);
[t25*] (M25);  [t26*] (M26);  [t27@0] (M27);
[t28*] (M28);  [t29*] (M29);  [t30@0] (M30);
[t31*] (M31);  [t32*] (M32);  [t33@0] (M33);
[t34*] (M34);  [t35*] (M35);  [t36@0] (M36);
[t37*] (M37);  [t38*] (M38);  [t39@0] (M39);
[t40*] (M40);  [t41*] (M41);  [t42@0] (M42);
```

```
[t43*] (M43);  [t44*] (M44);  [t45@0] (M45);
[t46*] (M46);  [t47*] (M47);  [t48@0] (M48);
[t49*] (M49);  [t50*] (M50);  [t51@0] (M51);
[t52*] (M52);  [t53@0] (M53);  [t54*] (M54);
[t55*] (M55);  [t56*] (M56);  [t57@0] (M57);
[t58*] (M58);  [t59*] (M59);  [t60@0] (M60);
[t61*] (M61);  [t62*] (M62);  [t63@0] (M63);
[t64*] (M64);  [t65@0] (M65);  [t66*] (M66);
[t67*] (M67);  [t68*] (M68);  [t69@0] (M69);
[t70*] (M70);  [t71*] (M71);  [t72@0] (M72);
[t73*] (M73);  [t74*] (M74);  [t75@0] (M75);
[t76*] (M76);  [t77*] (M77);  [t78@0] (M78);
[t79*] (M79);  [t80*] (M80);  [t81@0] (M81);
[t82*] (M82);  [t83*] (M83);  [t84@0] (M84);
[t85*] (M85);  [t86*] (M86);  [t87@0] (M87);
[t88*] (M88);  [t89*] (M89);  [t90@0] (M90);
[t91*] (M91);  [t92*] (M92);  [t93@0] (M93);
[t94*] (M94);  [t95*] (M95);  [t96@0] (M96);
[t97*] (M97);  [t98*] (M98);  [t99@0] (M99);
[t100*] (M100);  [t101*] (M101);  [t102@0] (M102);
[t103*] (M103);  [t104*] (M104);  [t105@0] (M105);
[t106*] (M106);  [t107*] (M107);  [t108@0] (M108);
[t109*] (M109);  [t110*] (M110);  [t111@0] (M111);
[t112*] (M112);  [t113@0] (M113);  [t114*] (M114);
[t115*] (M115);  [t116*] (M116);  [t117@0] (M117);
[t118*] (M118);  [t119*] (M119);  [t120@0] (M120);
```

```
MODEL G2:
  Trait1-Trait10*1;
  [Trait1-Trait10*0];

!Binary outcomes' errors are zero since this is ranking
i1i2-i119i120@0;

! Utility errors are set to 1
t1-t120@1;
  ! pairwise thresholds are fixed to 0
  [i1i2$1-i119i120$1@0];

! and utilities means are free except anchor items and one item per block for identification

[t1*] (N1);  [t2*] (N2);  [t3@0] (N3);
[t4*] (N4);  [t5*] (N5);  [t6@0] (N6);
[t7*] (N7);  [t8*] (N8);  [t9@0] (N9);
[t10*] (N10);  [t11*] (N11);  [t12@0] (N12);

[t13*] (M13);  [t14*] (M14);  [t15@0] (M15);
[t16*] (M16);  [t17*] (M17);  [t18@0] (M18);

[t19*] (N19);  [t20*] (N20);  [t21@0] (N21);

[t22*] (M22);  [t23*] (M23);  [t24@0] (M24);

[t25*] (N25);  [t26*] (N26);  [t27@0] (N27);
```

[t28*] (M28);  [t29*] (M29);  [t30@0] (M30);

[t31*] (M31);  [t32*] (M32);  [t33@0] (M33);

[t34*] (M34);  [t35*] (M35);  [t36@0] (M36);

[t37*] (N37);  [t38*] (N38);  [t39@0] (N39);

[t40*] (N40);  [t41*] (N41);  [t42@0] (N42);

[t43*] (N43);  [t44*] (N44);  [t45@0] (N45);

[t46*] (N46);  [t47*] (N47);  [t48@0] (N48);

[t49*] (N49);  [t50*] (N50);  [t51@0] (N51);

[t52*] (N52);  [t53@0] (N53);  [t54*] (N54);

[t55*] (N55);  [t56*] (N56);  [t57@0] (N57);

[t58*] (N58);  [t59*] (N59);  [t60@0] (N60);

[t61*] (N61);  [t62*] (N62);  [t63@0] (N63);

[t64*] (N64);  [t65@0] (N65);  [t66*] (N66);

[t67*] (M67);  [t68*] (M68);  [t69@0] (M69);

[t70*] (N70);  [t71*] (N71);  [t72@0] (N72);

[t73*] (N73);  [t74*] (N74);  [t75@0] (N75);

[t76*] (M76);  [t77*] (M77);  [t78@0] (M78);

[t79*] (N79);  [t80*] (N80);  [t81@0] (N81);

[t82*] (N82);  [t83*] (N83);  [t84@0] (N84);

[t85*] (N85);  [t86*] (N86);  [t87@0] (N87);

[t88*] (N88);  [t89*] (N89);  [t90@0] (N90);

[t91*] (N91);  [t92*] (N92);  [t93@0] (N93);

```
[t94*] (N94);  [t95*] (N95);  [t96@0] (N96);

[t97*] (N97);  [t98*] (N98);  [t99@0] (N99);

[t100*] (N100);  [t101*] (N101);  [t102@0] (N102);

[t103*] (N103);  [t104*] (N104);  [t105@0] (N105);

[t106*] (N106);  [t107*] (N107);  [t108@0] (N108);

[t109*] (N109);  [t110*] (N110);  [t111@0] (N111);

[t112*] (N112);  [t113@0] (N113);  [t114*] (N114);

[t115*] (N115);  [t116*] (N116);  [t117@0] (N117);

[t118*] (N118);  [t119*] (N119);  [t120@0] (N120);


MODEL TEST:
```

## Appendix 4 – Analysis R Script

```
# Before running the script:
# set your working directory to the location of the data files.
#setwd('Simulation/Script and Models/Results')
# Load Libraries (install if needed)
library(dplyr)
library(gridExtra)
library(tidyr)
library(ggplot2)
library(effectsize)
library(extrafont)

# Loading Data
FB_Data <- read.csv('FB_Data.csv')
CB_Data <- read.csv('CB_Data.csv')

#Remove ID column
FB_Data <- FB_Data[-1]
CB_Data <- CB_Data[-1]



####---------------------------- Convergence and Data Preparation -----------------------####

#####----- Checking Missing values for convergence -----###
Cb_data_noMiss <- filter(CB_Data, CB_Data$WaldChiSq_PValue != 9999)
FB_DataNomiss <- filter(FB_Data, FB_Data$WaldChiSq_PValue != 9999)
```

```
FB_DataMiss <- filter(FB_Data, FB_Data$WaldChiSq_PValue == 9999)
num_99_values_FB <- sum(FB_Data$WaldChiSq_PValue == 9999, na.rm = TRUE)
num_99_values_CB <- sum(CB_Data$WaldChiSq_PValue == 9999, na.rm = TRUE)

# Checking Conditions where FB Data was missing
FB_DataMiss$anchorDIFCon <- as.factor(FB_DataMiss$anchorDIFCon)
fb_missStats<- summary(FB_DataMiss$anchorDIFCon)
fb_missStats/nrow(FB_DataMiss)

#####------ Function to compute statistics -----####
compute_stats <- function(df_subset) {
  type_1_error_df <- df_subset %>% filter(blockType == "NoDIF")
  power_df <- df_subset %>% filter(blockType == "DIF")

  type_1_error_rate <- if (nrow(type_1_error_df) > 0) {
    correct_non_identification <- ifelse(type_1_error_df$WaldChiSq_PValue > 0.05, 1, 0)
    round(1 - sum(correct_non_identification) / nrow(type_1_error_df), 5)
  } else {
    NA
  }

  power_rate <- if (nrow(power_df) > 0) {
    correct_identification <- ifelse(power_df$WaldChiSq_PValue <= 0.05, 1, 0)
    round(sum(correct_identification) / nrow(power_df), 5)
  } else {
    NA
  }
  return(data.frame(type_1_error_rate = type_1_error_rate, power_rate = power_rate))
}



#####------- Free-baseline analysis, table creation ------#####
variables <- c("samplesize", "TraitCon", "DifCon", "PerDIFCon", "Anchorcon",
"anchorDIFCon")
values_list <- list(1:4, 1:2, 1:2, 1:3, c(20, 30), c(0, 50, 100))
results_fb <- list()
results_rep <- list()
for (samplesize in 1:4) {
  for (trait_size in 1:2) {
    for (difcon in 1:2) {
      for (perdifcon in 1:3) {
        for (anchorcon in c(20, 30)) {
          for (anchordifcon in  c(0, 50, 100)) {
            df_subset <- FB_DataNomiss %>%
              filter(
                Samplesize == samplesize,
                TraitCon == trait_size,
                DifCon == difcon,
                PerDIFCon == perdifcon,
                Anchorcon == anchorcon,
                anchorDIFCon == anchordifcon
              )
            print(paste("Samplesize", samplesize, "Trait:", trait_size, "DifCon:", difcon,
"PerDIFCon:", perdifcon, "Anchorcon:", anchorcon, "AnchorDIFCon:", anchordifcon, "Rows:",
nrow(df_subset)))


            stats <- compute_stats(df_subset)


            stats$Samplesize <- samplesize
            stats$Trait_Size <- trait_size
            stats$DifCon <- difcon
            stats$PerDIFCon <- perdifcon
            stats$Anchorcon <- anchorcon
            stats$AnchorDIFCon <- anchordifcon


            results_fb[[length(results_fb) + 1]] <- stats
          }
        }
      }
    }
  }
}



resultsFreeBaseline <- bind_rows(results_fb)


resultsFreeBaseline <- resultsFreeBaseline %>%
  mutate(
    `Percent DIF Blocks` = recode(PerDIFCon, `1` = 40, `2` = 50, `3` = 60),
    `Anchor Set Size` = recode(Anchorcon, `20` = 20, `30` = 30),
    `DIF Included in Anchor` = recode(AnchorDIFCon, `0` = 0, `50` = 50, `100` = 100)
  )



resultsFreeBaseline <- resultsFreeBaseline %>%
  mutate(
    `5 Factor Small DIF Type I Error` = ifelse(Trait_Size == 1 & DifCon == 1, type_1_error_rate, NA),
    `5 Factor Small DIF Power` = ifelse(Trait_Size == 1 & DifCon == 1, power_rate, NA),
    `5 Factor Large DIF Type I Error` = ifelse(Trait_Size == 1 & DifCon == 2, type_1_error_rate, NA),
    `5 Factor Large DIF Power` = ifelse(Trait_Size == 1 & DifCon == 2, power_rate, NA),
    `10 Factor Small DIF Type I Error` = ifelse(Trait_Size == 2 & DifCon == 1, type_1_error_rate, NA),
    `10 Factor Small DIF Power` = ifelse(Trait_Size == 2 & DifCon == 1, power_rate, NA),
    `10 Factor Large DIF Type I Error` = ifelse(Trait_Size == 2 & DifCon == 2, type_1_error_rate, NA),
    `10 Factor Large DIF Power` = ifelse(Trait_Size == 2 & DifCon == 2, power_rate, NA)
  )



# Separate Type I Error and Power tables
final_FB_TypeIError <- resultsFreeBaseline %>%
  group_by(`Samplesize`,`Percent DIF Blocks`, `Anchor Set Size`, `DIF Included in Anchor`) %>%
  summarise(
    `5 Factor Small DIF Type I Error` = round(mean(`5 Factor Small DIF Type I Error`, na.rm = TRUE), 5),
    `5 Factor Large DIF Type I Error` = round(mean(`5 Factor Large DIF Type I Error`, na.rm = TRUE), 5),
    `10 Factor Small DIF Type I Error` = round(mean(`10 Factor Small DIF Type I Error`, na.rm = TRUE), 5),
    `10 Factor Large DIF Type I Error` = round(mean(`10 Factor Large DIF Type I Error`, na.rm = TRUE), 5)
  )


final_FB_Power <- resultsFreeBaseline %>%
  group_by(`Samplesize`,`Percent DIF Blocks`, `Anchor Set Size`, `DIF Included in Anchor`) %>%
  summarise(
    `5 Factor Small DIF Power` = round(mean(`5 Factor Small DIF Power`, na.rm = TRUE), 5),
    `5 Factor Large DIF Power` = round(mean(`5 Factor Large DIF Power`, na.rm = TRUE), 5),
    `10 Factor Small DIF Power` = round(mean(`10 Factor Small DIF Power`, na.rm = TRUE), 5),
    `10 Factor Large DIF Power` = round(mean(`10 Factor Large DIF Power`, na.rm = TRUE), 5)
  )



# Uncomment to create .csv tables
#write.csv(final_FB_TypeIError, 'final_FB_TypeIError.csv')
#write.csv(final_FB_Power, 'final_FB_Power.csv')


##### ---------------- Constrained-Baseline Analysis and Table Creation --------- #
results_cb <- list()


# Calculate Stats
for (SampleSize in c(1, 2,3,4)) {
  for (trait_size in c(1:2)) {
    for (difcon in c(1, 2)) {
      for (perdifcon in c(1, 2, 3)) {
        df_subset <- Cb_data_noMiss %>%
         filter(Samplesize == SampleSize,
            TraitCon == trait_size,
            DifCon == difcon,
            PerDIFCon == perdifcon)

        # Print out the number of rows after filtering
        print(paste("Samplesize",SampleSize,"Trait:", trait_size, "DifCon:", difcon, "PerDIFCon:", perdifcon, "Anchorcon:", anchorcon, "AnchorDIFCon:", anchordifcon, "Rows:", nrow(df_subset)))

        # Compute stats
        stats <- compute_stats(df_subset) # Assuming TraitCon is the right variable, else replace

        # Add the iterated variables to the results
        stats$Samplesize <- SampleSize
        stats$Trait_Size <- trait_size
        stats$DifCon <- difcon
        stats$PerDIFCon <- perdifcon

        results_cb[[length(results_cb) + 1]] <- stats
      }
    }
  }
}
# Combine all results
results_constrained <- bind_rows(results_cb)

###----------- Table Creation -------------###
# Needs Expanded for Sample Size
```

```
results_constrained <- results_constrained %>%
  mutate(
    `Percent DIF Blocks` = recode(PerDIFCon, `1` = 40, `2` = 50, `3` = 60)
  )

results_constrained <- results_constrained %>%
  mutate(
    `5 Factor Small DIF Type I Error` = ifelse(Trait_Size == 1 & DifCon == 1, type_1_error_rate, NA),
    `5 Factor Small DIF Power` = ifelse(Trait_Size == 1 & DifCon == 1, power_rate, NA),
    `5 Factor Large DIF Type I Error` = ifelse(Trait_Size == 1 & DifCon == 2, type_1_error_rate, NA),
    `5 Factor Large DIF Power` = ifelse(Trait_Size == 1 & DifCon == 2, power_rate, NA),
    `10 Factor Small DIF Type I Error` = ifelse(Trait_Size == 2 & DifCon == 1, type_1_error_rate, NA),
    `10 Factor Small DIF Power` = ifelse(Trait_Size == 2 & DifCon == 1, power_rate, NA),
    `10 Factor Large DIF Type I Error` = ifelse(Trait_Size == 2 & DifCon == 2, type_1_error_rate, NA),
    `10 Factor Large DIF Power` = ifelse(Trait_Size == 2 & DifCon == 2, power_rate, NA)
  )



# Separate Type I Error and Power tables
final_CB_TypeIError <- results_constrained %>%
  group_by(`Samplesize`,`Percent DIF Blocks`) %>%
  summarise(
    `5 Factor Small DIF Type I Error` = round(mean(`5 Factor Small DIF Type I Error`, na.rm = TRUE), 4),
    `5 Factor Large DIF Type I Error` = round(mean(`5 Factor Large DIF Type I Error`, na.rm = TRUE), 4),
    `10 Factor Small DIF Type I Error` = round(mean(`10 Factor Small DIF Type I Error`, na.rm = TRUE), 5),
    `10 Factor Large DIF Type I Error` = round(mean(`10 Factor Large DIF Type I Error`, na.rm = TRUE), 5)
  )


final_CB_Power <- results_constrained %>%
  group_by(`Samplesize`,`Percent DIF Blocks`) %>%
  summarise(
    `5 Factor Small DIF Power` = round(mean(`5 Factor Small DIF Power`, na.rm = TRUE), 4),
    `5 Factor Large DIF Power` = round(mean(`5 Factor Large DIF Power`, na.rm = TRUE), 4),
    `10 Factor Small DIF Power` = round(mean(`10 Factor Small DIF Power`, na.rm = TRUE), 5),
    `10 Factor Large DIF Power` = round(mean(`10 Factor Large DIF Power`, na.rm = TRUE), 5)
  )



#write.csv(final_CB_TypeIError,'final_CB_TypeIError.csv')

#write.csv(final_CB_Power,'final_CB_Power.csv')



##### ------ Data frame Ungrouping -----------------#######
final_CB_Power <- final_CB_Power %>%
  ungroup() %>%
  mutate(
    Marginal_Small_DIF_5F = `5 Factor Small DIF Power`,
    Marginal_Large_DIF_5F = `5 Factor Large DIF Power`,
    Marginal_Small_DIF_10F = `10 Factor Small DIF Power`,
    Marginal_Large_DIF_10F = `10 Factor Large DIF Power`
  )
final_CB_TypeIError <- final_CB_TypeIError %>%
  ungroup() %>%
  mutate(
    Marginal_Small_DIF_5F = `5 Factor Small DIF Type I Error`,
    Marginal_Large_DIF_5F = `5 Factor Large DIF Type I Error`,
    Marginal_Small_DIF_10F = `10 Factor Small DIF Type I Error`,
    Marginal_Large_DIF_10F = `10 Factor Large DIF Type I Error`
  )
final_FB_Power <- final_FB_Power %>%
  ungroup() %>%
  mutate(
    Marginal_Small_DIF_5F = `5 Factor Small DIF Power`,
    Marginal_Large_DIF_5F = `5 Factor Large DIF Power`,
    Marginal_Small_DIF_10F = `10 Factor Small DIF Power`,
    Marginal_Large_DIF_10F = `10 Factor Large DIF Power`
  )
# Ungroup the data frame before calculating marginal values
final_FB_TypeIError <- final_FB_TypeIError %>%
  ungroup() %>%
  mutate(
    Marginal_Small_DIF_5F = `5 Factor Small DIF Type I Error`,
    Marginal_Large_DIF_5F = `5 Factor Large DIF Type I Error`,
    Marginal_Small_DIF_10F = `10 Factor Small DIF Type I Error`,
    Marginal_Large_DIF_10F = `10 Factor Large DIF Type I Error`
  )
```

```
##### ------- ANOVAS ----- ####

# Compute statistics function
compute_stats <- function(df_subset) {
  type_1_error_df <- df_subset %>% filter(blockType == "NoDIF")
  power_df <- df_subset %>% filter(blockType == "DIF")

  type_1_error_rate <- if (nrow(type_1_error_df) > 0) {
    correct_non_identification <- ifelse(type_1_error_df$WaldChiSq_PValue > 0.05, 1, 0)
    round(1 - sum(correct_non_identification) / nrow(type_1_error_df), 5)
  } else {
    NA
  }

  power_rate <- if (nrow(power_df) > 0) {
    correct_identification <- ifelse(power_df$WaldChiSq_PValue <= 0.05, 1, 0)
    round(sum(correct_identification) / nrow(power_df), 5)
  } else {
    NA
  }

  return(data.frame(type_1_error_rate = type_1_error_rate, power_rate = power_rate))
}

# Generate the results
variables <- c("samplesize", "TraitCon", "DifCon", "PerDIFCon", "Anchorcon",
"anchorDIFCon")
```

```
values_list <- list(1:4, 1:2, 1:2, 1:3, c(20, 30), c(0, 50, 100))
results_AOV_FB <- list()

for (samplesize in 1:4) {
  for (trait_size in 1:2) {
    for (difcon in 1:2) {
      for (perdifcon in 1:3) {
        for (anchorcon in c(20, 30)) {
          for (anchordifcon in  c(0, 50, 100)) {
            df_subset <- FB_DataNomiss %>%
              filter(
                Samplesize == samplesize,
                TraitCon == trait_size,
                DifCon == difcon,
                PerDIFCon == perdifcon,
                Anchorcon == anchorcon,
                anchorDIFCon == anchordifcon
              )

            print(paste("Samplesize", samplesize, "Trait:", trait_size, "DifCon:", difcon,
"PerDIFCon:", perdifcon, "Anchorcon:", anchorcon, "AnchorDIFCon:", anchordifcon, "Rows:",
nrow(df_subset)))

            stats <- compute_stats(df_subset)

            stats$Samplesize <- samplesize
            stats$Trait_Size <- trait_size
            stats$DifCon <- difcon
            stats$PerDIFCon <- perdifcon
            stats$Anchorcon <- anchorcon
            stats$AnchorDIFCon <- anchordifcon

            results_AOV_FB[[length(results_AOV_FB) + 1]] <- stats
          }
        }
      }
    }
  }
}

# Combine all results into a single data frame
AOV_dataframe_FB <- do.call(rbind, results_fb)

# Convert factors if necessary
AOV_dataframe_FB$Anchorcon <- as.factor(AOV_dataframe_FB$Anchorcon)
AOV_dataframe_FB$anchorDIFCon <- as.factor(AOV_dataframe_FB$AnchorDIFCon)
AOV_dataframe_FB$PerDIFCon <- as.factor(AOV_dataframe_FB$PerDIFCon)
AOV_dataframe_FB$Samplesize <- as.factor(AOV_dataframe_FB$Samplesize)
AOV_dataframe_FB$DifCon <- as.factor(AOV_dataframe_FB$DifCon)
AOV_dataframe_FB$Trait_Size <- as.factor(AOV_dataframe_FB$Trait_Size)

# Perform ANOVA for Power Rate
anova_power_FB <- aov(power_rate ~  Anchorcon * AnchorDIFCon * PerDIFCon, data = AOV_dataframe_FB)

# Perform ANOVA for Type I Error Rate
anova_type1_FB <- aov(type_1_error_rate ~ Anchorcon * AnchorDIFCon * PerDIFCon, data = AOV_dataframe_FB)
```

```
# Exploratory Power

anova_power_explor <- aov(power_rate ~  DifCon * AnchorDIFCon * Samplesize *Trait_Size,
data = AOV_dataframe_FB)


# Exploratory Type I Error Rate

anova_type1_explor <- aov(type_1_error_rate ~ DifCon * AnchorDIFCon * Samplesize
*Trait_Size, data = AOV_dataframe_FB)


####------------ JUST RESULTS -----------####

# Convergence

num_99_values_CB/nrow(CB_Data)

num_99_values_FB/nrow(FB_Data)


# Planned ANOVA

summary(anova_power_FB)

omega_squared(anova_power_FB, partial =F)

summary(anova_type1_FB)

omega_squared(anova_type1_FB, partial =F)


# Exploratory ANOVA

summary(anova_power_explor)

omega_squared(anova_power_explor, partial =F )

summary(anova_type1_explor)

omega_squared(anova_type1_explor, partial =F )
```

## Appendix 5 – DIF Design Table

| Trait | DIF Items in Each Percent of DIF Condition | | |
|---|---|---|---|
| | Five-Trait | | |
| | 40 | 50 | 60 |
| 1 | 1 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 3 | 1 | 2 | 2 |
| 4 | 2 | 2 | 3 |
| 5 | 2 | 2 | 3 |
| Total | 8 | 10 | 12 |
| | Ten-Trait | | |
| 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 2 |
| 3 | 1 | 2 | 2 |
| 4 | 2 | 2 | 2 |
| 5 | 2 | 2 | 3 |
| 6 | 2 | 2 | 3 |
| 7 | 1 | 2 | 2 |
| 8 | 2 | 2 | 2 |
| 9 | 2 | 2 | 3 |
| 10 | 2 | 2 | 3 |
| Total | 16 | 20 | 24 |

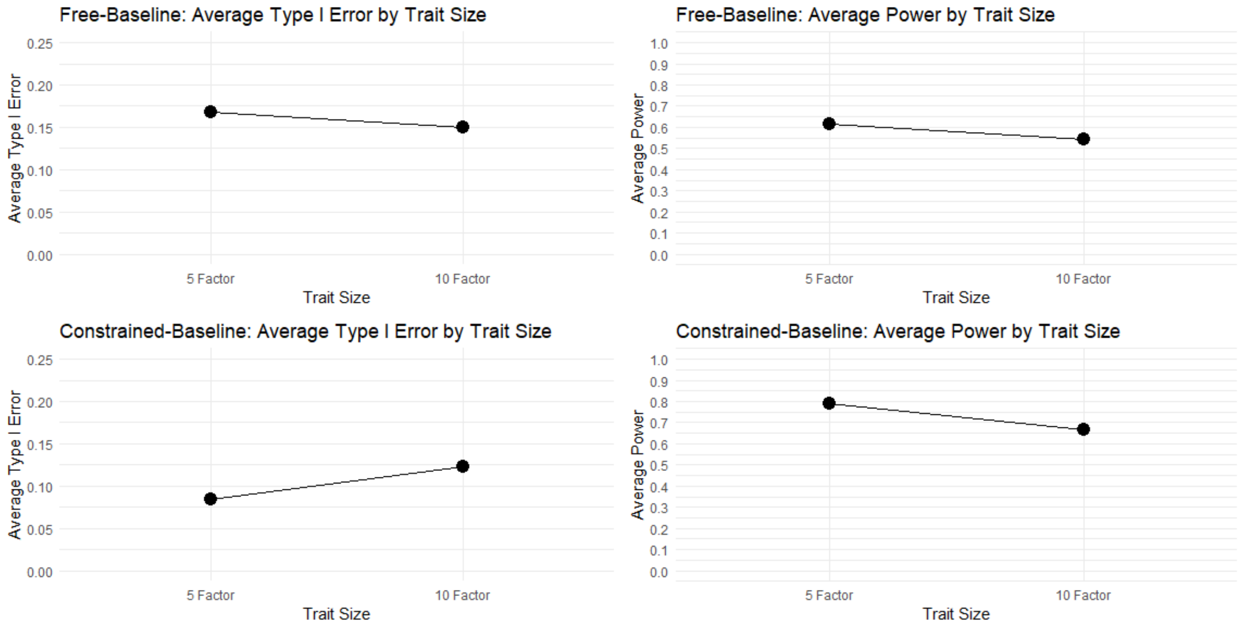
## Supplementary Materials

### Extra Tables and Graphs

RQ1. Number of Traits

Fig. 1

*Average Power and Type I Error Rates for Number of Traits*



RQ2. DIF Effect Size

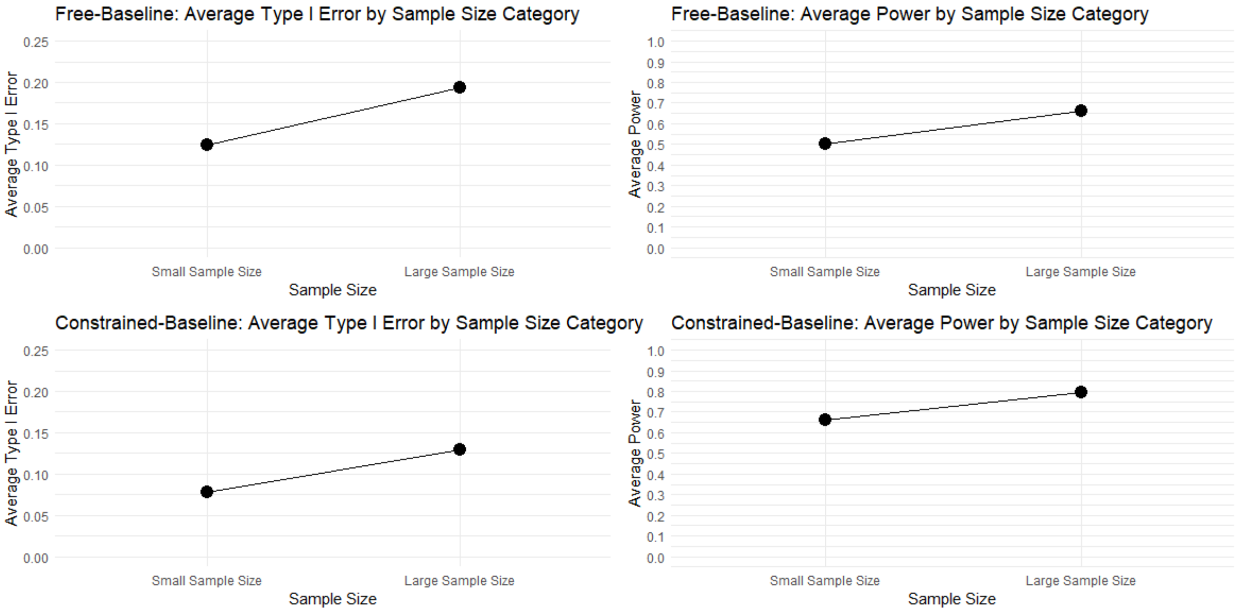
Fig. 2

*Average Power and Type I Error Rates for Effect Size*

RQ3. Sample Size

Fig. 3

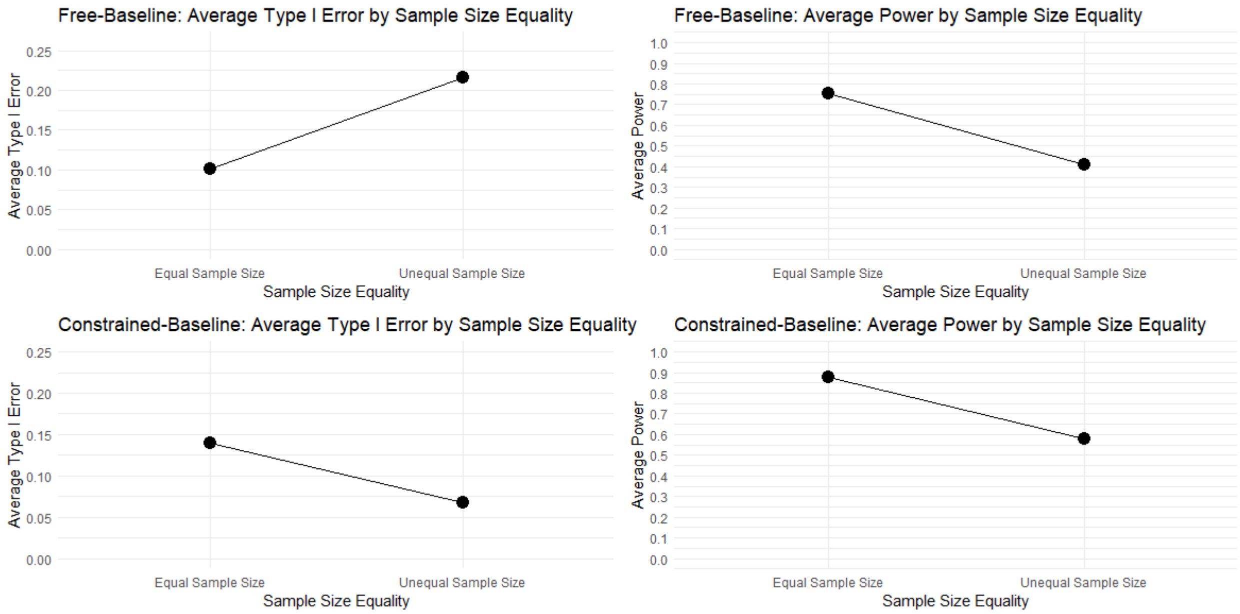
*Average Power and Type I Error Rates for Sample Size*



RQ8. Sample Size Equality

Fig. 4

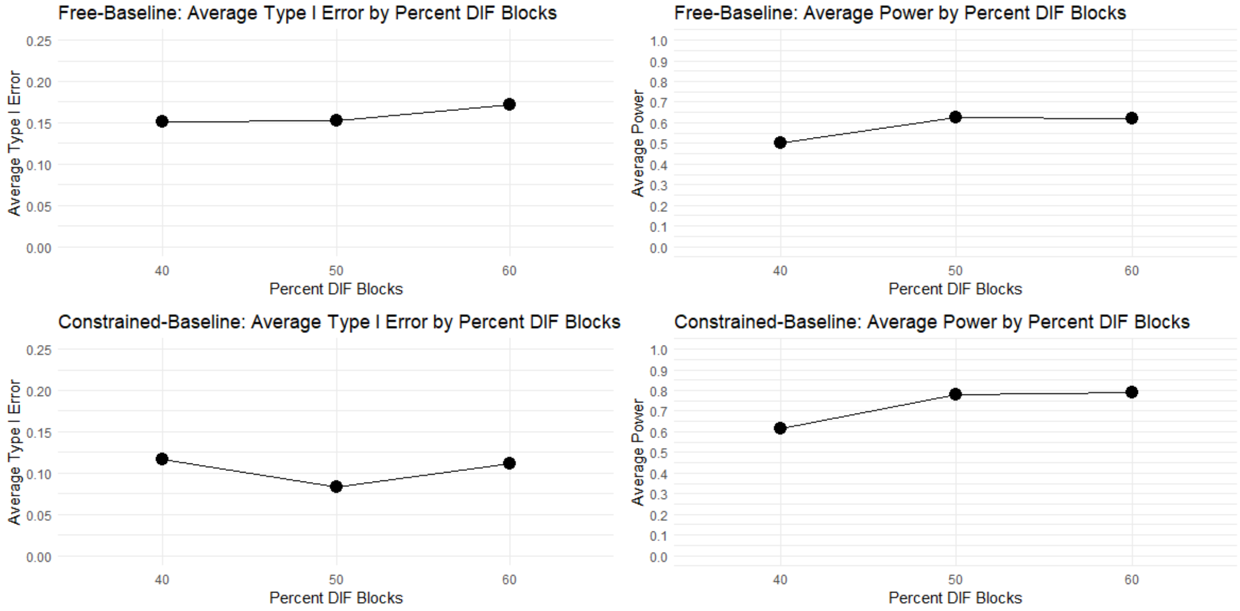
*Average Power and Type I Error Rates for Sample Size Equality*



RQ5. Percent of DIF Blocks

Fig. 5

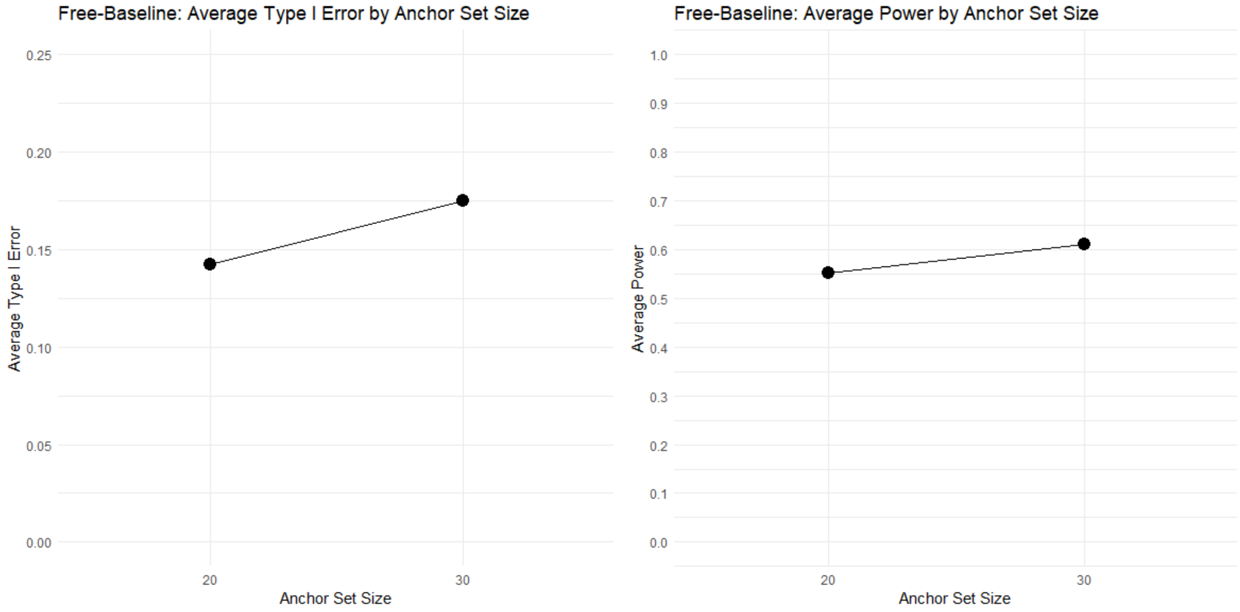
*Average Power and Type I Error Rates for Percent of DIF Blocks*



RQ6. Anchor Size

Fig. 6

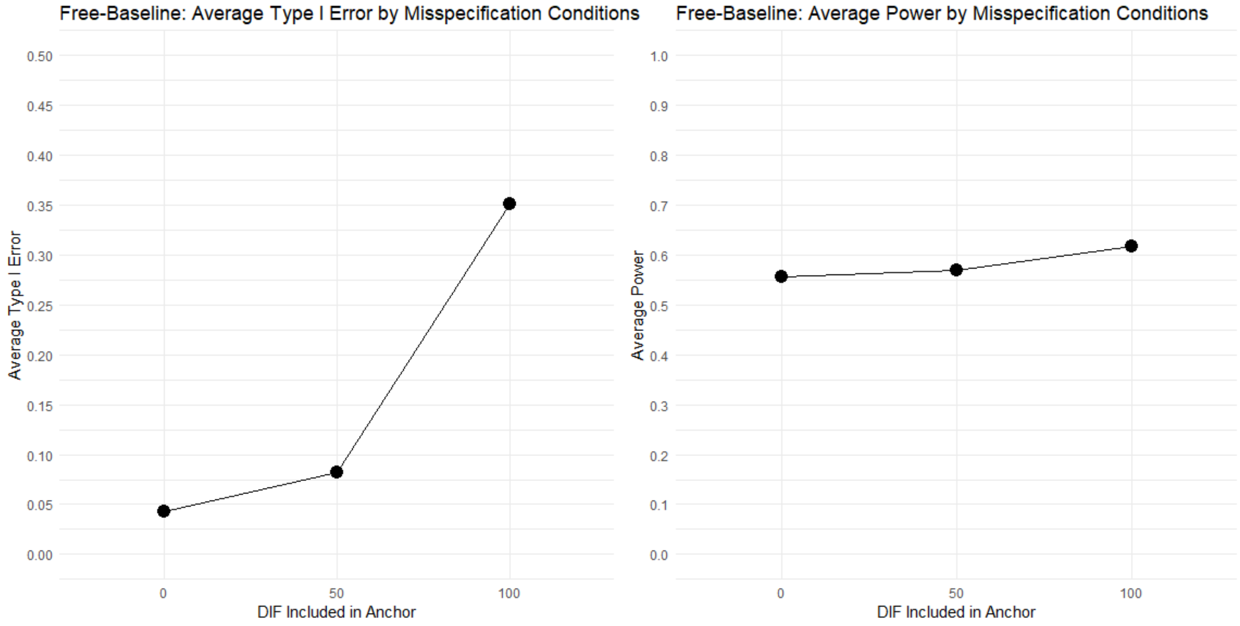
*Average Power and Type I Error Rates for Anchor Set Size*



RQ7. Misspecification

Fig. 7

*Average Power and Type I Error Rates for Misspecification*



Planned Analysis ANOVA Table
Type I Error and Power Effects

| Effect | *F* | *df*<br>*(Within, Between)* | *p* | $\omega^2$ |
|---|---|---|---|---|
| Type I Error | | | | |
| Anchor (Anc) | 2.620 | (1, 276 ) | 0.107 | < .001 |
| Misspecification (MS) | 154.867 | (1, 276) | < .001 | 0.35 |
| Percent of DIF Items (PDIF) | 0.444 | (2, 276) | 0.642 | 0 |
| Anc x MS | 0.645 | (1, 276) | 0.423 | 0 |
| Anc x PDIF | 0.097 | (2, 276) | 0.907 | 0 |
| PDIF x MS | 0.539 | (2, 276) | 0.584 | 0 |
| Power | | | | |
| Anc | 3.064 | (1, 276) | 0.081 | < .001 |
| MS | 2.276 | (1, 276) | 0.133 | < .001 |
| PDIF | 5.725 | (2, 276) | 0.004 | 0.03 |
| Anc x MS | 0.012 | (1, 276) | 0.913 | 0 |
| Anc x PDIF | 0.167 | (2, 276) | 0.847 | 0 |
| PDIF x MS | 0.102 | (2, 276) | 0.903 | 0 |

# Chapter 5. General Discussion

# General Discussion

Non-cognitive assessments are increasingly used to inform high-stakes educational and occupational decisions. When a test is used to make life-altering decisions a strong validity argument is needed. While the value of non-cognitive assessments has been increasingly recognized for their potential to predict individual success (Duckworth & Yaeger, 2015), their application in high-stakes contexts comes with challenges. The biggest challenge is that people can misrepresent themselves. They may do this by answering non-cognitive items in a way they think they should rather than answering honestly. This is known as faking (Brown & Maydeu-Olivares, 2011). The forced-choice (FC) test format is a promising alternative to rating scales that can reduce faking, but like any other methodological decision, it introduces its own unique challenges (Bürkner et al, 2019; Lin & Brown, 2017; Ng et al., 2021). In my dissertation we first established what types of validity evidence non-cognitive tests tend to have in Chapter 2 by expanding on prior reviews (Cordier et al., 2016; Cox et al., 2019; Halle & Darling-Churchill, 2016). I then focused on the FC test format as a method that can improve the validity of test scores by reducing faking in Chapter 3. In this chapter we comprehensively reviewed the state of FC methodology from the first time it was mentioned in the 1940s into the modern-day methods used for their construction. Our work advanced prior reviews by thoroughly laying out all stages of FC development and their relevant methodology rather than focusing on a single phase of the development process (Li et al., 2024; Wetzel et al., 2020). This serves as the first step toward creating a standardized procedure for their construction in the future. I also identified methods that needed further investigation in FC. This led me to my work in Chapter 4 where we built upon Lee and colleagues (2021) work by examining a method for detecting differential item functioning (DIF) under realistic operational conditions. In the remainder of this chapter, I provide a summary of each chapter, describe their implications, and discuss their limitations. I then draw connections across the body of work to discuss its implications and limitations as a whole.

### Chapter 2 – Non-Cognitive Mapping Review

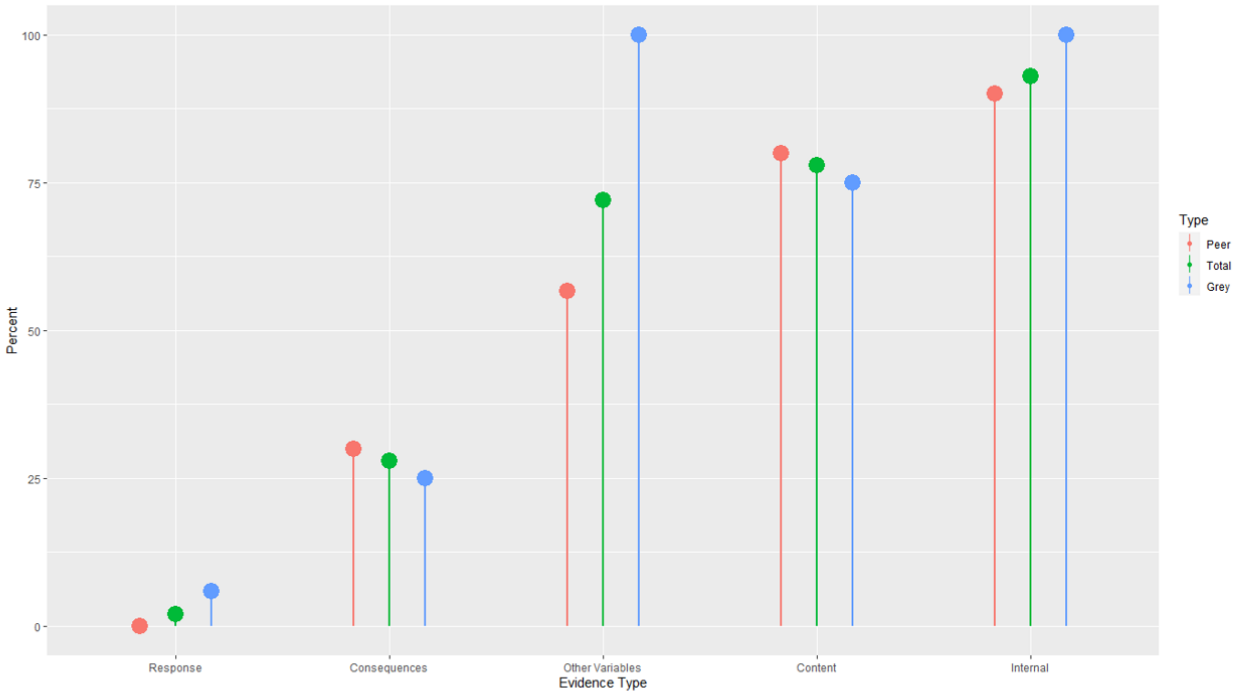
In Chapter 2 of this dissertation, I critically evaluated the types of validity evidence commonly reported for non-cognitive assessments and identified gaps in current reporting standards. This chapter began with a review of the various frameworks and terms that fall under the umbrella term of non-cognitive. We then outlined the types of validity evidence psychological instruments are expected to possess when they are aligned with the Standards (AERA et al., 2024). These include the types shown in Table 1.

Table 1 - Types of Validity Evidence

| Evidence Type | Definition | Evidence Considerations |
|---|---|---|
| Test Content | The construct has been identified and defined, and content experts were consulted. | Was a construct identified, and if yes, was it defined?<br>Were content experts consulted? |
| Response Process | Whether the theory was examined or individual responses were systematically tested. | Were response processes to items tested with individuals (e.g., think-alouds)? |
| Internal Structure | Any statistical technique to determine if the model reflects the construct it proposes to measure (e.g., factor analyses) and that the scores are reliable.<br>Psychometric properties are equal across groups. | Were any statistics that test for internal structure reported or measured?<br>Was IRT or factor analysis used to test for equality across groups? |
| Relationships with Other Variables | Evidence for how the construct is related to other variables. | This may include correlations, testing for group differences, and predictive testing of the measure with other variables. |
| Consequences of Testing | Includes positive or negative consequences. Evidence that pertained to how the score was interpreted or other evidence of the score’s applied purpose and utility was noted (Messick, 1975). | Were consequences considered? |

Afterwards, we conducted a mapping review of 46 validity studies from the peer-reviewed and grey literature. This work expanded upon reviews that had been conducted previously by focusing on non-cognitive skills generally and their entire validity argument rather than a few select skills or types of validity evidence (Cordier et al., 2016; Cox et al., 2019; Halle & Darling-Churchill, 2016). Our results indicated  that there were several gaps in the types of validity evidence reported for the peer-review and grey literature measures examined. First, it was common practice for the tests to be reported without demographic data. There was also a lack of consideration for the consequences of testing or response process evidence (see Fig. 1). These two results are linked as consequences of testing are often related to the differential impact a test has on different groups of people. This can give early signs of potential group differences in how people interpret the test items as well as provide evidence on if the construct is being adequately measured (Padilla, & Benitez, 2014). Finally, while internal structure evidence was reported frequently it largely related to the reporting of reliability coefficients with minimal examination of DIF or other statistical analyses to examine group differences.

Fig. 1 – Differences in evidence reported between grey and peer-reviewed literature



The findings from this work indicated that test developers need to report more information, especially on demographic data, response processes, and the consequences of using the test. These are important features that must be considered to ensure the fairness of the test. While other reviews (e.g., Cordier et al., 2015; Cox et al., 2019; Halle & Darling-Churchill, 2016) examined subsets of non-cognitive skills and found mixed results, they did not provide recommendations on how the reporting of validity evidence could be improved. To this end we made several recommendations on how the reporting standards of non-cognitive assessment can improve. This work also served as a needed review on the state of non-cognitive validity evidence in the framework of the Standards (AERA et al., 2014).

This work has several limitations. The first was raised by a reviewer of the associated manuscript of this chapter who rightly pointed out that often times, validity evidence for an instrument is spread across multiple manuscripts and articles. In my mapping review I only searched for readily accessible information and did not conduct a thorough review instrument by instrument. This means pieces of validity evidence may have been missed. This will be addressed in the revisions of our manuscript. Secondly, this work is positioned in the framework of the Standards. There are other frameworks for considering validity (see Borsboom, 2004) that may have shifted or changed the recommendations made. Finally, this work is limited to educational testing and the tests we examined. While there may be generalizability of our recommendations to other tests or fields, our results are only directly applicable to the 46 measures we examined.

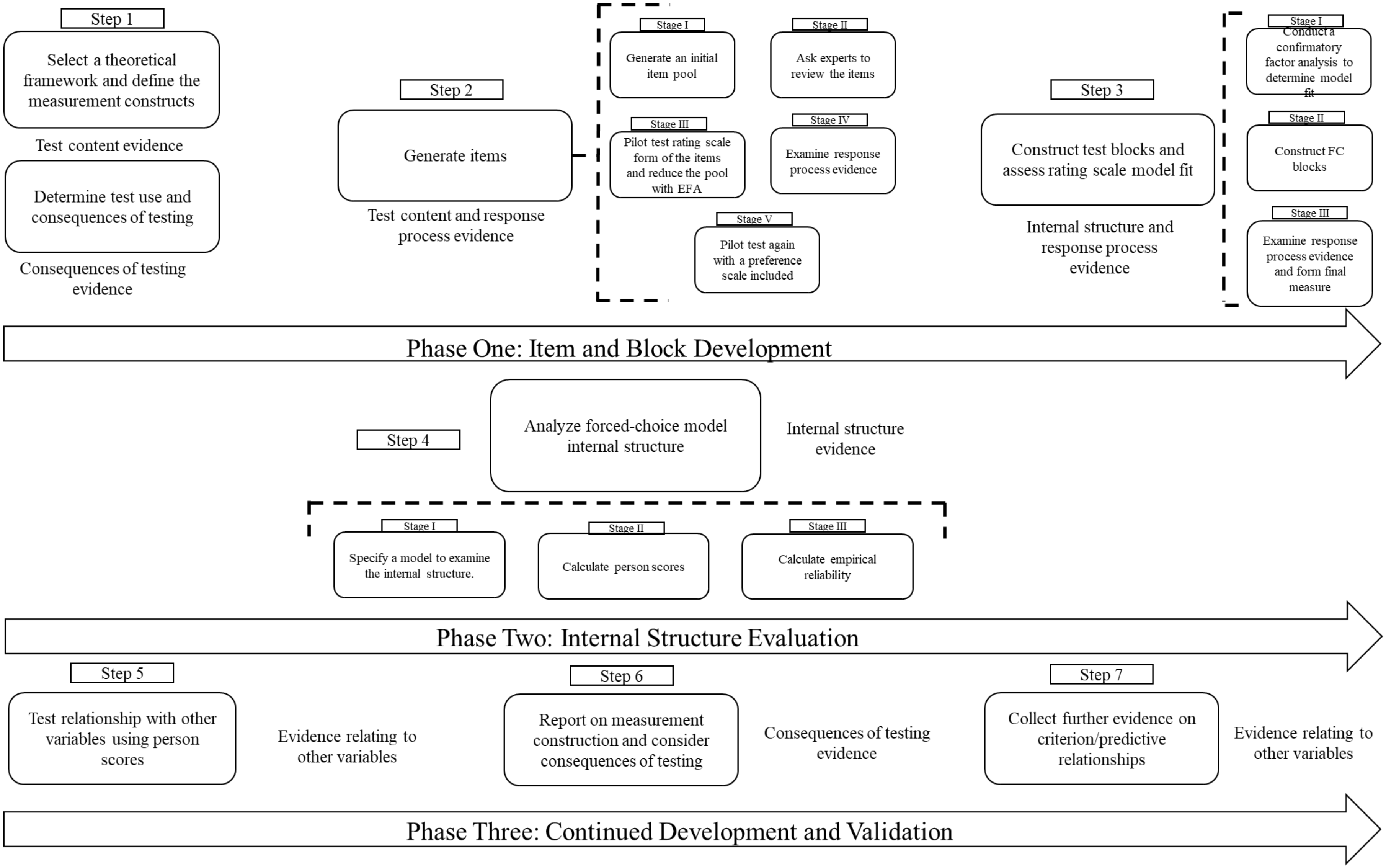
### Chapter 3 – Forced-Choice Review

In Chapter 3, I chronicled the history of FC from its first mention in the 1940s into a review on the modern-day methods being used to construct these types of tests. Our goal was to lay out the state of the field as a first step toward creating a more structured construction procedure in the future once best-practices are established. This involved reviewing issues in item development, block construction, scoring approaches, reliability and measurement bias. We also created an annotated bibliography covering the more than eight decades of literature we reviewed in this chapter. This chapter placed all of the FC methodology reviewed in three phases of development that were aligned with the Standards (AERA et al., 2014; see Fig. 2). While others had described or focused on a few parts of the construction process (see Li et al., 2024; Wetzel et al., 2020), ours is the first to bring all parts of it together into a single review. Finally, we detailed gaps in the current methodology for constructing FC tests such as a lack of methodology for DIF testing, a need for further examination of response process evidence, and continued development of block construction procedures. The primary limitation of this chapter is the same as FC itself. The format is still developing methodological best-practices. In the future the contribution of the chapter may increase as the process described in Fig. 2 is supported.

## Chapter 4 – Simulation Study

In Chapter 4, I examined a latent-scoring based approach for testing DIF in FC measures. I found this to be an under-researched area in Chapter 3, with only Lee and colleagues (2021) formally investigating this approach for FC, in conditions that were optimal. In their paper, Lee et al. investigated two latent-scoring based approaches for testing DIF in the Thurstonian-IRT model. The first is the free-baseline approach where an anchor set of blocks is constrained across the two groups of interest (Stark et al., 2006). The remaining blocks can then be tested for DIF all at once. When a DIF block is included in the anchor, this is considered a misspecification. Second is the constrained-baseline approach where all but one block is constrained equal across the groups. That block is then tested. It is then constrained, and the next block is freed and tested. This procedure continues until all blocks have been examined. In this chapter we modified Lee et al.'s approach from a first-order to a second-order TIRT model (see Brown & Maydeu-Olivares, 2011). This specification allows for the examination of uniform and non-uniform DIF separately. In the first-order model, they must be analyzed together.

Fig. 2 – Forced-Choice Measurement Construction Process


Step 1
Select a theoretical framework and define the measurement constructs
Test content evidence
Determine test use and consequences of testing
Consequences of testing evidence
Step 2
Generate items
Test content and response process evidence
Stage I
Generate an initial item pool
Stage II
Ask experts to review the items
Stage III
Pilot test rating scale form of the items and reduce the pool with EFA
Stage IV
Examine response process evidence
Stage V
Pilot test again with a preference scale included
Step 3
Construct test blocks and assess rating scale model fit
Internal structure and response process evidence
Stage I
Conduct a confirmatory factor analysis to determine model fit
Stage II
Construct FC blocks
Stage III
Examine response process evidence and form final measure
Phase One: Item and Block Development
Step 4
Analyze forced-choice model internal structure
Internal structure evidence
Stage I
Specify a model to examine the internal structure.
Stage II
Calculate person scores
Stage III
Calculate empirical reliability
Phase Two: Internal Structure Evaluation
Step 5
Test relationship with other variables using person scores
Evidence relating to other variables
Step 6
Report on measurement construction and consider consequences of testing
Consequences of testing evidence
Step 7
Collect further evidence on criterion/predictive relationships
Evidence relating to other variables
Phase Three: Continued Development and Validation

We then conducted a simulation study with 336 conditions focused on determining the effect of model-misspecification in the free-baseline approach, examining differences in the constrained and free-baseline approaches, and replicating some results from Lee and colleagues using a second-order TIRT model. In our simulation we examined power and Type I error rates. Power for a condition was defined as the proportion of DIF blocks correctly flagged using a multiple-constraint Wald test. Type I error rates were defined as 1 – the proportion of incorrectly flagged non-DIF blocks. A correct flag for power and an incorrect flag for Type I error means the Wald test is significant ($p \leq .05$).

Our simulation was able to replicate many of Lee et al.'s  results when there was no misspecification. When misspecification was present, meaning 50% or 100% of the anchor was composed of DIF blocks, the free-baseline approach did not accurately detect DIF or non-DIF blocks. The constrained-baseline approach was better compared to a misspecified free-baseline model but also did not have acceptable detection rates of DIF and non-DIF blocks. These results have implications for the use of the proposed DIF detection method. It does not appear to be accurate when misspecification is present, which is likely in practice. These results indicate that further research is needed to identify a reliable DIF testing method for FC tests.

Our study had several limitations. The primary limitation was that we could not keep the items we constrained across groups equal in all conditions due to the simulation design. At times there were more items constrained on a trait than other traits. This may have led to inflated Type I error or power rates due to factors not controlled in the simulation. This feature may have affected the results of the tested blocks depending on which traits they related to. For example, there may be an effect of testing a block where the related traits have 12 items constrained equal across groups as opposed to nine. Additionally, we did not examine conditions related to testing nonuniform DIF. This is an important future area of research that will need to be tested. Finally, our simulation only examined the second-order TIRT model. It is possible this approach may work better in ideal point models that were not examined here, even when there is misspecification.

## Implications

Together, my work has provided insights into the field of FC and non-cognitive assessment. While each study contributes to the field in its own way, there are two themes that emerged. First, there is a need for further consideration of fairness in testing and the methodology available to do so in FC. This also includes further consideration of how the evidence for fairness is presented and what is reported. In Chapter 2, we found that consideration of the consequences of testing was one of the least reported types of evidence. This coincides with a lack of demographic reporting. Over the course of our review in Chapter 3, it was determined that there was a lack of methodology for FC tests to test for measurement bias in items. Testing for measurement bias is a crucial aspect of ensuring the fairness of the test. When measurement bias is present, the scores of some groups may be over or underestimated, which could jeopardize an individual from that group's chance of success. While a method investigated by Lee and colleagues (2021) seemed promising, we found that it did not work in conditions representative of the real world, where misspecification may be present. This leaves FC testing without a clear method to examine potential measurement bias in items. While it is possible that think-aloud interviews and other forms of methods for collecting response process evidence (Fuechtenhans & Brown, 2022) may be used to examine group differences, this has yet to be tested and will need further investigation in the future.

Secondly, my research suggests that while the FC test format has many methods available for each phase of test construction, there is a need to identify the circumstances in which they work best. For example, there is a proliferation of modeling techniques (see Table 2). While dominance models such as TIRT (Brown & Maydeu-Olivares, 2011) appear to be the most well-researched, there are a variety of ideal point models, such as the GGUM-RANK (Hontengas et al., 2016) and MUPP (Stark et al., 2005) models, which can be used almost interchangeably (Brown, 2016). Each may have its specific purpose and rationale for use as methodologists try to correct issues they see in the current modeling landscape. However, the field would certainly be served by coming to an agreement on which select few are used. This focus would allow best practices, as discussed in Chapter 3, to be established. Furthermore, my dissertation indicates there needs to be a thorough consideration of how each method can be used to support the validity argument of the test. For example, we should be able to confidently answer how proper block development supports the use of the test. It is clear that reducing response bias through this process may support the internal structure and content of the test, but this has not been explicitly examined. In the future, work will be needed to establish best practices in FC testing and how they fit into forming a strong validity argument in line with the Standards (AERA et al., 2014).

Table 1 - Forced-Choice

Models

| | Model Name | Acronym | Citation | Measurement Model | Decision model |
|---|---|---|---|---|---|
| 1. | Zinnes-Griggs' Unfolding Preference Model | | Zinnes & Griggs, 1974 | LFA | Thurstonian |
| 2. | Simple Squared Difference Model for Pairwise Preferences | SSDMPP | Andrich, 1989 | IP | Bradley-Terry |
| 3. | Simple Hyperbolic Cosine Model | SHCMPP | Andrich, 1995 | IP | Bradley-Terry |
| 4. | Multi-Unidimensional Pairwise Preference Model | MUPP | Stark et al., 2005 | IP | Bradley-Terry |
| 5. | Bayesian Random Block – IRT | BRB-IRT | Lee & Smith, 2020 | IP | Bradley-Terry |
| 6. | Multi-Unidimensional Pairwise Preference Model – 2PL | MUPP-2PL | Morillo et al., 2016 | LFA | Bradley-Terry |
| 7. | Generalized-Graded Unfolding Model | GGUM - RANK | Hontengas et al., 2016 | IP | Bradley-Terry |
| 8. | Confirmatory Multidimensional Generalized-Graded Unfolding Model | CCGGUM | Wang & Wu, 2016 | IP | Bradley-Terry |
| 9. | Zinnes-Griggs Pairwise Preference Item Response Theory Model | ZG-MUPP | Joo et al., 2023 | IP | Bradley-Terry |
| 10. | Forced-Choice Ranking Models | FCRM's | Hung & Huang, 2022 | IP or LFA | Bradley-Terry |
| 11. | Generalized Thurstonian Unfolding Model | GTUM | Zhang et al. 2023 | LFA | Bradley-Terry |
| 12. | Joint-Response-Time Thurstonian-IRT | JRT-TIRT | Guo et al., 2023 | LFA | Thurstonian |
| 13. | Thurstonian-IRT | TIRT | Brown & Maydeu-Olivares, 2011 | LFA | Thurstonian |

*Note*. IP = Ideal-Point. LFA = Factor Analysis. This table does not describe more advanced applications such as faking-mixture models (see Frick, 2021; Brown & Böckenholt, 2023).

## Limitations and Open Questions

The work in this dissertation was motivated by an interest in understanding, translating, and advancing the state of FC methodology in non-cognitive testing. I aimed to bring together decades of research and communicate it in an easily-digestible way for a broad audience. I also contributed a new understanding of current FC methodology and the state of non-cognitive testing to the field. Although this body of work advances the field of FC non-cognitive assessment, it has several important limitations.

First and foremost, this work presumes there is utility in the FC format. This is backed by several pieces of research showing that FC can reduce response bias (Bartlett et al., 1960; Brogden, 1954; Brown & Maydeu-Olivares, 2011; Cao & Drasgow, 2019; Christiansen et al., 2005; Edwards, 1957; Goffin et al., 2011; Jackson et al., 2000; Joubert et al., 2015; Lee & Smith, 2020; Wetzel et al., 2020; Wetzel & Frick, 2020) and increase the validity of test scores (Bartram, 2007; Lee et al., 2018; Wetzel & Frick, 2020; Vasilopoulos et al., 2006). However, others have shown this is questionable when there are mixed-keyed blocks on the test (Bürkner et al., 2019; Pavlov et al., 2019; Schulte et al., 2021). It is also the case that there are more steps in developing an FC assessment than other formats (see Chapter 3). These are necessary to construct an FC test with a strong validity argument. Additionally, methodology for the FC format is still developing and there are not readily available methods for testing DIF. This results in a need for test makers to carefully weigh the pros and cons of choosing to design an FC test and what decisions it should support. While the current approach to DIF is applicable in only some scenarios, think-aloud interviews and analysis of the rating scale items may provide insight into potential group differences. However, it may still be best to caution against the use of these types of tests for admissions decisions, where the respondents tend to be diverse, before a full array of methodology and practice has been developed.

Secondly, all but Chapter 4 of this dissertation were focused on reviewing the state of the field. While I believe this is necessary work that was foundational in informing Chapter 4, it does limit the potential impact of the dissertation. For example, if Chapter 3 also empirically designed a new assessment using the process described in Fig. 2 to show its efficacy, the impact of the chapter would be increased.

This dissertation leads to several questions that still need work to address them. First, in Chapter 2, a reviewer noted the need for a more thorough investigation of each instrument that was examined. I plan to conduct a more in-depth mapping review that identifies all relevant literature for each assessment. This will improve the utility of the review and provide a more accurate picture on the strength of the validity arguments presented for non-cognitive assessments. It would also be interesting to focus this review further on FC assessments exclusively. In line with this I have begun a meta-analysis that will provide a picture of the types of validity evidence reported by FC assessments. Secondly, I would like to conduct a follow-up to Chapter 4 where I examine non-uniform DIF and explore new methods for DIF detection. This may involve reformulation of observed score techniques to handle ipsative data or working toward a new technique that is built for FC assessment. Finally, as I build FC assessments in my future work and continue to research this area, I will return to Chapter 3 and assemble a structured construction procedure for FC assessments.

## Conclusion

My dissertation has made significant contributions to our understanding of how FC tests are developed and their methodology. Across three chapters I detailed how the current validity arguments of non-cognitive tests could be improved, how the FC format can be used to do so, and focused on methodology that could help improve their fairness. Overall, I showed how FC non-cognitive tests are reaching a point where best-practices are being established but that there is still work to be done to ensure they can be built with a strong validity argument that considers all evidentiary types.

## References

Algan, Y., Beasley, E., Vitaro, F., & Tremblay, R. (2014). The impact of non-cognitive skills training on academic and non-academic trajectories: From childhood to early adulthood.

Almlund, Duckworth, & Heckman. (2011). Personality psychology and economics. *Handbook of the Economics of Art and Culture*.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56.

Bartlett, C. J., Quay, L. C., & Wrightsman, L. S. (1960). A Comparison of two methods of attitude measurement: likert-type and forced choice. *Educational and Psychological Measurement*, *20*(4), 699–704.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*(3), 263–272.

Brogden, H. E. (1954). A simple proof of a personnel classification theorem. *Psychometrika, 19*(3), 205–208.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061.

Brown, A., & Bartram, D. (2011). *OPQ32r Technical Manual*. https://kar.kent.ac.uk/44780/

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502

Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827-854.

CASEL. (2020). Collaborative for academic, social, and emotional learning. https://casel.org.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *The Journal of Applied Psychology*, *104*(11), 1347–1368.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*(3), 267–307.

Cordier, R., Speyer, R., Chen, Y.-W., Wilkes-Gillan, S., Brown, T., Bourke-Taylor, H., Doma, K., & Leicht, A. (2015). Evaluating the psychometric quality of social skills measures: a systematic review. PLOS ONE, *10*(7), e0132299.

Cox, J., Foster, B., & Bamat, D. (2019). A review of instruments for measuring social and emotional learning skills among secondary school students. REL 2020-010. Regional Educational Laboratory Northeast & Islands.

Cronbach, L. J. (1956). Assessment of individual differences. *Annual Review of Psychology*, *7*(1), 173–196. https://doi.org/10.1146/annurev.ps.07.020156.001133

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher* , *44*(4), 237–251.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. *108*. https://psycnet.apa.org/fulltext/1958-00464-000.pdf

Egalite, A. J., Mills, J. N., & Greene, J. P. (2016). The softer side of learning: Measuring students' non-cognitive skills. *Improving Schools*, *19*(1), 27–40.

Fuechtenhans, M., & Brown, A. (2023). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment*, *31*(1), 105-119.

García, E. (2016). The need to address non-cognitive skills in the education policy agenda. *Non-Cognitive Skills and Factors in Educational*, 31–64.

Goffin, R. D., Jang, I., & Skinner, E. (2011). Forced-choice and conventional personality assessment: Each may have unique value in pre-employment testing. *Personality and Individual Differences*, *51*(7), 840–844.

Halle, T. G., & Darling-Churchill, K. E. (2016). Review of measures of social and emotional development. Jo*urnal of Applied Developmental Psychology*, *45*, 8–18.

Heckman, J. J., & Kautz, T. (2014). *Achievement tests and the role of character in American life*. https://tkautz.github.io/documents/Heckman_Kautz_2014_Achievement%20Tests.pdf

Heckman, James J., & Kautz, T. (2013). *Fostering and Measuring Skills: Interventions That Improve Character and Cognition* (No. 19656). National Bureau of Economic Research. https://doi.org/10.3386/w19656

Heckman, James J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482.

Hontangas, P. M., Leenen, I., de la Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema, 28*(1), 76–82. https://doi.org/10.7334/psicothema2015.204

Humphries, J. E., & Kosse, F. (2017). On the interpretation of non-cognitive skills–What is being measured and why it matters. *Journal of Economic Behavior & Organization, 136,* 174-185. Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution? *Human Performance, 13*(4), 371–388.

Jones, P. R., Moore, D. R., Shub, D. E., & Amitay, S. (2015). The role of response bias in perceptual learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(5), 1456–1470.

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, *23*(1), 92–97.

Kautz, T., & Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. onegoalgraduation.org. https://www.onegoalgraduation.org/wp-content/uploads/2018/04/Measuring-and-Fostering-Non-Cognitive-Skills-in-Adolescence-Evidence-from-Chicago-Public-Schools-and-the-OneGoal-Program.pdf

Kautz, Tim, Heckman, J. J., Diris, R., ter Weel, B., & Borghans, L. (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success* (No. 20749). National Bureau of Economic Research. https://doi.org/10.3386/w20749

Lee, H., & Smith, W. Z. (2020). A Bayesian random block item response theory model for forced-choice formats. *Educational and Psychological Measurement*, *80*(3), 578–603.

Lee, P., Joo, S.-H., & Stark, S. (2021). Detecting DIF in multidimensional forced choice measures using the Thurstonian Item Response Theory model. *Organizational Research Methods*, *24*(4), 739–771.

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*, 229–235.

Li, M., Zhang, B., Li, L., Sun, T., & Brown, A. (2024). Mixed-Keying or Desirability-Matching in the construction of Forced-Choice measures? An empirical investigation and practical recommendations. *Organizational Research Methods*, 10944281241229784.

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389-414.

Martins, P. S. (2010). *Can Targeted, Non-Cognitive Skills Programs Improve Achievement? Evidence from Epis*. https://doi.org/10.2139/ssrn.1696890

Melnick, H., Cook-Harvey, C. M., & Darling-Hammond, L. (2017). Encouraging social and emotional learning in the context of new accountability. Learning Policy Institute.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269, W64.

Ng, V., Lee, P., Ho, M.-H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The Development and Validation of a Multidimensional Forced-Choice Format Character Measure: Testing the Thurstonian IRT Approach. *Journal of Personality Assessment*, *103*(2), 224–237. https://doi.org/10.1080/00223891.2020.1739056

Öhman, M. (2015). *"Be Smart, Live Long: The Relationship Between Cognitive and Non-Cognitive Abilities and Mortality" and "Exogenous Health Information and Individuals' Subjective Well-Being: RD Evidence from a Screening-Program for an Asymptomatic Disease."* https://www.econstor.eu/bitstream/10419/129387/1/837890918.pdf

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and Likert scores. *Organizational Research Methods*, *22*(3), 710-739.

Schulte, N., Holling, H., & Bürkner, P. C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats?. *Educational and Psychological Measurement*, *81*(2), 262-289.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, *70*(1), 747-770.

Smithers, L. G., Sawyer, A. C. P., Chittleborough, C. R., Davies, N. M., Davey Smith, G., & Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour, 2*(11), 867–880.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement*, *29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of applied psychology, 91*(6), 1292.

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, *88*(4), 1156-1171.

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review*, *35*(3), 175–197.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-Choice personality tests: A measure of personality and cognitive ability? *Human Performance*, *19*(3), 175–199.

Walton, K. E., Burrus, J., Murano, D., Anguiano-Carrasco, C., Way, J., & Roberts, R. D. (2022). A Big Five-based multimethod social and emotional skills assessment: The Mosaic$^{TM}$ by ACT® social emotional learning assessment. *Journal of Intelligence*, *10*(4). https://doi.org/10.3390/jintelligence10040072

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253.

Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales: Current state of the research. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *36*(4), 511–515.