1 Introduction

Psychological measures, tests, and assessments are ubiquitous in many societies (Oakland et al., 2016; Zlatkin-Troitschanskaia et al., 2018). One widespread use has been for tracking academic progress. In the United States, scores on standardized tests contribute to progression to the next grade level and decisions about admission to college as well as to rankings of schools and evaluations of teachers (Lemann, 2000; Moss et al., 2005; Young, 2021). Similar uses in England include standardized testing in primary grades, high-stakes examinations at the end of secondary school, and publicly available rating systems (Rimfeld et al., 2019; Santori, 2020). In China, a tradition of examinations extends back centuries, and contemporary National College Entrance Examinations determine college entry (Bodenhorn et al., 2020; Rotberg, 2010). Beyond testing academic progress, school psychologists also use assessments of social, behavioral, and emotional behaviors to screen children for referrals to intervention in countries around the world (Oakland et al., 2016). And, these school-based usages intersect with clinical and organizational psychology use of tests, as a component of diagnoses of psychiatrically defined disorders and of workplace hiring and promoting (Benjamin, 2005; Rothstein & Goffin, 2006).

Measures, tests, and assessments address the challenge that many key concepts are not directly observable in the psychological sciences, and related social, health, and educational sciences (i.e., are *latent*). It is therefore common to measure latent constructs using things that are observable, such as a series of questions about knowledge, behaviors, expressions, and attitudes that individuals can report. For instance, the Aggression Questionnaire measures individuals' tendencies toward aggression through their answering on a five-point scale how characteristic of them (labeled extremely uncharacteristic, uncharacteristic, neither characteristic nor uncharacteristic, characteristic, and extremely characteristic) are a few dozen actions such as "I flare up quickly but get over it quickly" and "If somebody hits me, I hit back" (Buss & Perry, 1992; Buss & Warren, 2000). As another example, the Achenbach System of Empirically Based Assessment (ASEBA, n.d.) includes versions that ask parents, teachers, and children to report about children's behaviors. Responses to various subsets of items contribute to summary scores in relation to empirically based syndromes and psychiatric diagnostic classifications – for example, contributing to scores on an aggressive behavior syndrome for young children are statements like "Easily frustrated," "Doesn't seem to feel guilty after misbehaving" and "Physically attacks people" which are reported as being not true, somewhat/sometimes true, or very/often true of the child.

Because of the ways such measures gatekeep access to opportunities and mark individuals with prestigious or stigmatizing statuses, their use has been contested (Lemann, 2000; Moss et al., 2005; Young, 2021). Social movements in the 1960s, for instance, heightened attention to the question of whether measures fairly assessed abilities across groups, such as between those who were assigned as female versus male or as Black versus White (Byrne et al., 2009; Davidov et al., 2014). Considerations of fairness addressed questions such as: Do various groups define a construct in the same way? Do the groups view similar knowledge, behaviors, expressions, and attitudes as reflective of the construct? Do groups vary in how they interpret or report about a particular expression or behavior? Continuing the example of tendencies toward aggression introduced earlier, considerations of fairness might include asking whether some groups interpret aggressive conduct as reflecting a behavioral disorder and others do not – for example, would some groups view hitting someone back after being hit as reflective of such a disorder, and other groups consider such a response as a reasonable defense of self? Considerations of fairness might also concern what has been termed *social desirability bias* – the tendency for a person to adjust their responses to be in line with what the person thinks is the expected answer (Duckworth & Yeager, 2015). In other words, for some individuals and some contexts, affirmation of the statement "If somebody hits me, I hit back" might be viewed as a sign of strength, and thus potentially overreported (i.e., endorsed by some people who do not actually hit back when hit), and for others such affirmation might be seen as a weakness and thus potentially underreported (i.e., not endorsed by some people who do actually hit back when hit).

Such complexities in how concepts are defined and interpreted and controversies about how tests are used have led psychometricians to expand strategies for assessing fairness, including the central concept of *measurement invariance* (also known as a *lack of measurement bias*; AERA/APA/NCME, 2014; Camilli, 2006; Xi, 2010). Formally defined in subsequent paragraphs, measurement invariance broadly entails the degree to which a measure's questions operate similarly across groups. As early twenty-first century societies again contend with systemic inequities, and movements call for equity and antiracism, the need is urgent for psychologists to comprehensively consider measurement invariance (Han et al., 2019).

Despite the need for examining measurement invariance as an aspect of fairness, the capacity of the field to do so is limited. Partly the limited field capacity reflects minimal training of students and scholars in psychometrics, and particularly item response theory (IRT) approaches. Over a decade ago, a national survey of graduate programs in psychology in the United States, for

instance, found that two-fifths offered no training in IRT and less than one in ten offered full coverage (Aiken et al., 2008). A Canadian survey likewise found few offerings in advanced statistics, including structural equation modeling (Golinski & Cribbie, 2009). Limited coverage was again identified in a recent US national survey that found nearly one quarter of graduate programs completely lacked coverage of psychometrics in introductory statistics courses and another fifth restricted coverage to a single class period or less (Sestir et al., 2021). The consequences of limited field capacity are amplified by the complexity of early strategies for empirically identifying measurement invariance. Iterative approaches for measurement invariance testing are particularly labor intensive, especially when groups are numerous (Cheung & Lau, 2012). Implementing these approaches therefore required particularly advanced levels of programming skill. And, substantive scholars required basic understanding of the techniques in order to best understand the rationale for such investment of time and effort and in order to draw inferences for theory and practice from the volumes of results.

These challenges may contribute to the relative lack of publications documenting invariance for measures commonly used in psychology. For instance, in a large-scale analysis of fifteen widely used measures in social and personality psychology (such as the Rosenberg Self-Esteem Scale), whereas nearly all measures demonstrated good evidence of internal consistency, only one demonstrated good evidence of measurement invariance (Hussey & Hughes, 2020). A review of a representative sample of articles from the *Journal of Personality and Social Psychology* also revealed that the majority of articles reported only reliability coefficients as structural validity evidence; the review authors noted that although they "observed numerous studies which tested hypotheses about numerous populations (e.g., age-groups, cultures) . . . only one tested measurement invariance" (Flake et al., 2017).

The purpose of this tutorial is to support developmental scientists in using and interpreting one recently developed technique for empirically identifying measurement invariance and adjusting for the invariance that is revealed, the *alignment method* (Asparouhov & Muthén, 2014, 2023; Muthén & Asparouhov, 2014). The alignment method was developed for cross-national research (Asparouhov & Muthén, 2014; Marsh et al., 2018; Muthén & Asparouhov, 2014), and has been applied more extensively in that field than in other areas of psychology (e.g., Bansal et al., 2022; Bordovsky et al., 2019; Bratt et al., 2018; Gordon et al., 2022; Lansford et al., 2021; Rescorla et al., 2020). The alignment method differs from other measurement invariance techniques in making it straightforward to allow for *partial invariance* in which some questions similarly reflect a construct across groups and other questions differ in their

relationship to the construct across groups. Other measurement invariance techniques have been designed to detect whether invariance holds or not, and offer less guidance or require more complicated strategies when invariance is rejected.

We not only provide an accessible introduction to the key concepts undergirding the alignment method but we also: (a) show how to implement it in the software package *Mplus* using algorithms written by the alignment approach's authors, (b) provide an R package for reading the volumes of results (openly accessible through *GitHub*), and (c) detail how to interpret the results. Importantly, our focus is on the kinds of multi-category (e.g., Likert, 1932) questions common in psychology (such as the five- and three-category response options in the examples of measuring tendencies toward aggression provided earlier). In contrast, existing tutorials and applications of the alignment method have primarily focused on continuous and dichotomous items (e.g., Sirganci et al., 2020). We also differ from prior coverage of the alignment approach with categorical items (e.g., Svetina et al., 2020) in demonstrating how to convert the results to probability units. Probability units help make the results meaningful to substantive scholars and broader stakeholders. In other words, percentages are familiar to many given widespread use, whereas a model coefficient (such as a logit) may be less familiar. In the context of measurement invariance, a difference of 50 to 30 percent may be seen as large, whereas a difference of 41 to 39 percent may be seen as small, when comparing the chances that members of one group versus another would be rated to have "hitting back" behavior be extremely characteristic of them, despite being estimated to have equal latent tendency toward aggression. To illustrate how to implement the alignment method and interpret its results, we offer an empirical example, with code and data available in supplementary materials. Before considering this empirical example, we begin with a conceptual introduction to measurement invariance followed by a formal presentation of central mathematical models.

1.1 Introduction to Measurement Fairness

Much has been written about fairness in measurement, including from those contesting historical uses of standardized testing, from those suggesting ways to conceptualize cross-cultural variations in concepts and their measurement, and from those proposing specific strategies to psychometrically test for invariance and to address its absence (Dorans & Cook, 2016; Hui & Triandis, 1985; Johnson & Geisinger, 2022; Moss, 2016). In this section, we introduce a portion of these writings relevant to understanding the alignment method. Given the limited training in psychometrics across the field of psychology

reviewed earlier, we start with a general introduction to concepts of measurement and then discuss the importance of considering intersectionality and categorical items when testing for measurement invariance.

1.1.1 General Measurement Concepts

Similar to regression models allowing psychologists to see if empirical evidence is consistent with theoretical expectations about how one construct relates to another, psychometric models allow psychologists to see if empirical evidence is consistent with theoretical expectations about which knowledge, behaviors, expressions, and attitudes reflect a latent construct. Different from the core fundamentals of regression modeling, however, terminology and epistemology vary considerably across the psychometric literature. In the limited space of a tutorial, we are selective in what we cover.

One way we are selective is related to terminology, where we prioritize the term *measure* whenever possible as we are discussing concepts and offering interpretations. Some psychometric writing instead uses the terms *tests* or *assessments*. Likewise, we aim to use the term *questions* whenever possible to encompass what are sometimes referred to as *items* or *prompts*. One reason for our prioritization of the terms measure and question is to avoid implying that the alignment method can only be used with standardized or academic tests and assessments. Another reason is that we find the terms measure and question can be received by some audiences as more neutral. In contrast, the terms test and assessment can call to mind uses that are high stakes or that imply universally defined and expressed constructs. When introducing terminology and discussing mathematical models, however, we use the words test and item when doing so reflects conventions (e.g., item response theory; differential item functioning, item-level invariance). Here, our goal is to make this tutorial accessible to those already familiar with these conventional terms and to make the cited references accessible to those who want to learn more after reading the tutorial. Even as we do so, we encourage continued reflection and renaming in the field to prioritize inclusivity of terminology.

Another way we are selective in the context of the tutorial is in our focus on measurement invariance in general and the alignment method in particular. This focus allows us to limit our presentation to a set of concepts and techniques that can be covered within space constraints. At the same time, this focus can place out of sight the ways in which testing for measurement invariance is one aspect of a broader project of continuous measure improvement. We thus emphasize that we do in fact see measurement invariance testing as one component of an iterative process of accumulating and considering multiple pieces of evidence

before any particular use of a measure. This process may include evidence provided by a measure's developers, yet would also include evidence in the local use context and sharing of ownership, data, and interpretations with an array of stakeholders, including those responding to the measure and those impacted by its scores. Such consideration of local evidence and inclusion of an array of stakeholders are central components of fairness in general, and are relevant to considerations of measurement invariance in particular. The need to revisit the evidence for each potential use reflects the reality that the groups relevant to consider in relation to measurement invariance will differ across applications, and those being measured and impacted by scores will have insights into the meaning of constructs and their expressions. This inclusivity is especially important in fields where measures were historically developed by and with persons of limited diversity, and a critical gaze can illuminate areas of historical bias in the field itself and opportunities for future equity.

With such a critical gaze, we can draw from the body of psychometric models and writings to be part of a more inclusive approach, while recognizing historical biases. The field of psychometrics has itself evolved over time in understanding, reflecting, and changing its approach to fairness. As an example, the Standards for Educational and Psychological Testing, published collaboratively by the American Psychological Association along with the major educational and measurement societies (AERA/APA/NCME, 2014), reflects the latest in a series of publications dating back to the 1950s. The most recent standards elevated fairness as a "fundamental validity issue" that is an "overriding foundational concern" with a central issue being "equivalence of the construct being assessed" across groups (p. 49). This fundamental nature of fairness is in contrast to the prior standards, published in 1999, which limited fairness to specific populations (e.g., persons with disabilities, English language learners; Johnson & Geisinger, 2022).

The latest standards also embrace a unified validity framework (Messick, 1989). What had been seen as distinct types of validity (e.g., content, criterion, consequential) are now recognized as multiple pieces of validity evidence that are brought together when making a decision about whether a measure is suitable for a particular use. Fairness in general, and measurement invariance in particular, can be seen as one aspect of this body of validity evidence. The body of validity evidence is also now seen as continually accumulating, rather than static at the time a test was published. The latest standards advise decision-makers to use a range of strategies, including various psychometric models, as they make a determination regarding the extent to which the full body of evidence supports a proposed use for a measure. In other words, whereas historically a decisionmaker might have cited internal consistency reliability

or factor analyses reported in a publisher's manual, contemporary decision-makers would be encouraged to either locate evidence of validity for their specific use or, if none was available, to build such evidence. This evidence would include demonstrating that the measure's questions demonstrated measurement invariance across the relevant groups and in the local context of a specific application.

Tests of measurement invariance in general, and the alignment method in particular, thus offer empirical evidence related to a measure's validity. Results from testing measurement invariance, including with the alignment method, can be combined with other aspects of validity evidence to inform conclusions about multiple aspects of measurement fairness (Byrne et al., 2009; Davidov et al., 2014). At a conceptual level, if the alignment method indicated very little evidence of measurement invariance, decisionmakers might want to reconsider whether and how the construct is defined across groups. If partial invariance was identified, the instances of non-invariance might be probed to consider whether groups differed in terms of what knowledge, behaviors, expressions, or attitudes were reflective of varying levels of the construct. This probing might include how group members interpret the measure's questions that demonstrate non-invariance. This probing might also include considering the extent to which their responses are affected by social stereotypes and norms related to the measured construct. And, this probing might include considering whether the context in which the measure is administered heightens aspects of social desirability.

Strategies to use might include those from the field of measurement theory and practice for using psychometric model results to iteratively improve measures, including by engaging substantive experts to precisely define constructs and to write questions to reflect those constructs (Boulkedid et al., 2011; Evers et al., 2013; Lane et al., 2016; Wolfe & Smith, 2007a, 2007b). Examining patterns of results, in dialogue with diverse stakeholders and informed by scientific and indigenous concepts, literatures, and practices can lead to tentative interpretations (Chilisa, 2020; Sablan, 2019; Sprague, 2016; Walter & Andersen, 2016). Such tentative interpretations might be examined through future revisions of the measure. Complementary methods such as cognitive interviewing, item reviews, and focus groups can also inform interpretations. Throughout this process, collaborators may see that many aspects of measurement fairness are interrelated, as scrutinizing a measure's questions may lead to revised understandings of concepts, and altered definitions of concepts may result in updating a measure's questions. Engaging a range of stakeholders and variety of methods supports iterative and continuous improvement, including representatives from those being measured as well as content and methods experts, all inclusive of the groups being assessed.

As an example, we collaborated with a school district to iteratively improve a measure of students' social-emotional competencies using psychometric approaches including the alignment method. The school district engaged students, teachers, principals, and other stakeholders in interpreting the results. In a Student Voice Data Summit, for instance, students thought different patterns of socialization influenced why high school boys were more likely to endorse a question about "staying calm when stressed" as easy to do compared to high school girls, believing that boys were less likely to admit feeling stress compared to girls, who were more often socially encouraged to discuss their emotions freely. Following findings indicating measurement non-invariance between students who identified as Latino and Latina, the district's research and practice teams partnered on a project to adapt lessons in their social-emotional learning curriculum based on the findings (Gordon & Davidson, 2022).

1.1.2 Importance of Considering Group Intersections

A limitation in historical considerations of measurement fairness, including in psychology, has been a focus on a small number of groups. Indeed, early methods and tutorials often assumed two groups, one focal and one reference (Finch, 2016). With the critical gaze discussed previously, this approach can be seen as problematic, by assuming a binary and privileging one group as focal and othering the second reference group. Cross-national research in contrast more often considered measurement invariance across many groups. The alignment method arose in the latter context, designed to facilitate empirical identification and adjustment of measurement invariance with many groups. Although this cross-national application still tended to consider groups of a single type (multiple nations), the alignment method can be further extended to consider groups defined by layering together multiple aspects of identities, what we refer to as *multilayered groups*.

One way to define such multilayered groups would be to cross-classify multiple variables. If an existing data source had classifications of sex (e.g., male, female) and race-ethnicity (e.g., Black, White, Asian, Latino/a), then eight multilayered groups might be defined (i.e., Black male, Black female, White male, White female, Asian male, Asian female, Latino, Latina), for instance. Groups could also be defined in flexible ways, such as if some participants preferred to label their own identities or to not use labels, including those identifying as queer, nonbinary, or fluid. And, groups could be defined using theoretical paradigms that consider how systems of power intersect in ways that may amplify, mute, or transform one another dynamically, as in the

concept of intersectionality (Crenshaw, 1989). By facilitating this flexibility, the alignment methods might be used by scholars to interrogate measurement fairness from a range of theoretical perspectives and incorporate social-justice oriented modern data science, (Covarrubias & Vélez, 2013; Garcia et al., 2018; Sablan, 2019). To achieve larger sample sizes in various groups of interest, integrative analyses of multiple datasets might be used (Fujimoto et al., 2018).

1.1.3 Importance of Accounting for Multi-Category Items

Many psychological measures include questions that have multiple categories, such as Likert-type response structures and the five- and three-category response options offered in earlier examples. Yet, similar to the origins of regression modeling, numerous psychometric methods were first developed assuming continuous variables, and psychologists often continue to rely on these methods. We demonstrate how to use the alignment method with multi-category items. Doing so better conforms the model assumptions with the data. Doing so also allows for presentation of results in ways that are meaningful to substantive scholars and to a range of stakeholders: the probabilities of choosing various categories. Doing so additionally reduces the chance of overlooking important aspects of measurement invariance that are revealed in the category probabilities.

Although the statement that psychometric models used should be designed, implemented, and interpreted recognizing the questions' multi-category structure seems obvious, it has been common for scholars and analysts to adopt models designed for continuous items when questions are multi-category (Rhemtulla et al., 2012). Even when models designed for multi-category questions are used, interpretations of the substantive meaning of results can be incomplete (Gordon, 2015; Meitinger et al., 2020; Seddig & Lomazzi, 2019). Analogous to applying regression models designed for continuous versus multi-category outcomes, results can sometimes be robust across specifications (Long, 1997; Long & Freese, 2014). Yet, robustness across specifications should be evaluated in any particular application and not assumed. And, when models appropriate to multi-category outcomes are used, interpretation requires additional steps to convert to substantively meaningful metrics (e.g., probabilities vs. logits; Long, 1997; Long & Freese, 2014).

In other words, when a measure asks individuals to choose among a set of categorical responses to a question, results are meaningful when reported in terms of response probabilities. We might find, for instance, that 60 percent of one group versus 40 percent of another group are predicted to "strongly agree" with a statement, despite both groups being estimated to have the same level of

the underlying construct being measured. We could contrast this result with another where, say, the predicted percentages were 51 percent for the first group and 49 percent for the second group. Here, we discuss how the alignment method makes such calculations. Again, our goal is to make it easier to take this step from estimation to interpretation, given the volumes of results produced by the alignment method and given the need to convert results to meaningful metrics. This goal is consistent with implementation science and related strategies for encouraging the adoption of advanced methods (King et al., 2019; Sharpe, 2013). Reporting in meaningful units makes results more accessible to a range of stakeholders, including those being measured and the substantive scholars, practitioners, policymakers, family, peers, and community members who draw inferences from the scores (Gordon & Davidson, 2022; Moss, 2016). In other words, many will be familiar with percentages and probabilities from day-to-day usage, whereas fewer may be familiar with the logits. In line with modern statistical reporting standards, probabilities also allow stakeholders to consider the real-world importance of a difference, beyond its statistical significance.

1.2 Introduction to Psychometric Methods for Testing Measurement Invariance

There are two general types of psychometric models that can accommodate multi-category questions: *item factor analysis* (*IFA*) and *item response theory* (*IRT*; Embretson & Reise, 2000; Liu et al., 2017; Millsap, 2011). Each has been used to test for measurement invariance. The alignment method uses IFA during estimation. The alignment method also allows results to be translated to IRT format. By presenting both approaches, we support readers connecting to their own prior study of one or both of these methods as well as to the related literatures on each method. We also demonstrate the ways in which each tradition offers insights into measurement invariance.

1.2.1 Review of General Concepts of Item Factor Analysis and Item Response Theory

Psychometric models generally aim to produce empirical evidence regarding the extent to which responses to a measure's set of questions are consistent with the presence of the proposed latent construct. Many psychologists will be familiar with factor analysis, although likely with its most typical presentation assuming continuous items. Here, *factor loadings* are often of focus, capturing the strength of the association between the item and the latent construct. Underlying the factor analytic model are a series of regressions of the