Mixed Methods for Psychological Measurement

Using Critical Realism to Reframe Incommensurability

David F. Feldon

First published 2024

ISBN: 978-0-367-70129-1 (hbk)
ISBN: 978-0-367-70132-1 (pbk)
ISBN: 978-1-003-14471-7 (ebk)

Chapter 7

A Critical Turn

DOI: 10.4324/9781003144717-7

The funder of the Open Access version of this chapter is Utah State University.

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

7 A Critical Turn

David F. Feldon

In a society driven by data, there is enormous power placed in the hands of people who create measures. Their assumptions and beliefs about the world shape the ways in which they attempt to "carve nature at its joints" by conceptualizing the tools that can detect and parameterize the phenomena delineated by those joints. However, psychological phenomena rely on individuals' constructed sensemaking to complete their instantiation. As discussed previously, a complete critical realist ontology requires *both* an independent external reality (nomothetic) and perspectival internal reality (idiographic). Similarly, a phenomenographic methodological framework holds that the perspectival observations and experiences of reality are no less real than the physical events with which they interacted. In both frameworks, the basis of measurement is a causal relationship between a latent construct and its manifest indicators (Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2003). Thus, to fully capture the nature of psychological phenomena requires that the inherent variation in idiographic phenomena across individual perspectives be incorporated into measurement in an ontologically robust way.

Accordingly, attempts to measure psychological phenomena must engage mechanisms to capture idiographic components (in addition to nomothetic elements) as integral to, rather than detracting from, valid measurement. In the status quo, the interpretation of measurement models typically relies upon interpretation of mean trends across persons and leaves residual variance uninterpreted. Doing so leaves undifferentiated the idiographic information and random error present in residuals, ossifying the incommensurability between the full phenomenon and the interpreted model. By default, this centers nomothetic aspects of the targets of measurement and parses them fully from the idiographic aspects, treating the former as worthy of measuring and the latter as noise that hinders that endeavor. Fundamentally, this reduces measurement validity, because it presumes that an intrinsic part of the real signal is noise. Indeed, the exclusion of idiographic facets of a phenomenon from measurement prevents the attainment of

DOI: 10.4324/9781003144717-7

Guttman's (1944) deterministic conjoint additivity, because the phenomenon that generated the measured value is not fully reflected in the attained score.

Along these same lines, theories that are built on measures omitting consideration of idiographic factors will skew the construction of theory and the interpretation of results. Reliance on measurement validated and interpreted only in relation to the nomothetic aspects of a phenomenon inherently foregrounds aspects of latent phenomena that are easily quantifiable while forcing to the background aspects that are less so (Di Fiore, Kuc-Czarnecka, Piano, Puy, & Saltelli, 2023). Further, the readily quantified facets of phenomena yield data that can shift understandings of what is "normal" or "typical" or "average" in public discourse to reflect such uneven understandings (Amoore, 2020). Absent idiographic elements of phenomena, these understandings of normal can readily be assumed to reflect "natural" conditions rather than conditions that manifest specific perspectives that align with those of societally empowered groups (Padilla, 2004). In other words, the exclusion of idiographic elements heightens incommensurability, because it truncates relevant information about aspects of a target phenomenon that could otherwise be integrated with in the measurement model. Further, the incommensurability introduced will be skewed such that perspectives of those least likely to be called upon in the development and interpretation of a measure based on their positionalities are most likely to be eliminated.

Within this framework, it becomes clear that measures, in fact, cannot be constructed independent of idiographic influences. Any "self-evident" characteristics of a measure or the phenomenon it assesses are only self-evident to those whose idiographic lenses map sufficiently to those of the scholars who originated the framework within which the work is enacted (Bonilla-Silva & Zuberi, 2008). Thus, the positioning of measurement as objective through academic and political discourse falsely denies the existence and valid experiences of those people whose perspectives cannot be reconciled to the perspectives reflected in any given measure. Further, it instantiates incommensurability between the measure and portions of the natural population from which it was intended to collect data. Where measures are built on unmitigated incommensurability, they cannot be valid.

The Positivist Tradition of Precluding Perspective in Measurement

The tendency to approach measurement as perspective-free is part of a legacy of positivism stretching back to the initial development of statistics. Pearson explicitly took the position that measures and the statistical relationships amongst them could be used to eradicate the metaphysical and theoretical from the human sciences, yielding a completely objective basis for knowledge (Norton, 1979). However, the basis for evaluating such supposed objectivity was the presumption of self-evident truth embedded in eugenicist beliefs that all differences between people were inherited and that certain social and racial groups possessed superior traits to others (Gould, 1996). As noted by MacKenzie (1978, p. 54), "to define heredity as the correlation of parents and offspring indicates the a priori nature of Pearson's hereditarianism; that the correlation could be due to the similarity of parental and offspring environments was not even considered."

Similarly, Spearman's concept of a general intelligence factor (*g*) presupposed that "all individuals possess a general mental capacity called 'general intelligence' which enters with some (and varying) degree into all the diverse types of cognitive activity" (Urbach, 1974, p. 102). From this assumption follows that general intelligence would vary across individuals as a matter of heredity and asserts a latent univariate structure (with deviation in individual abilities explained by secondary, ability-specific factors) by which to measure the phenomenon using the "hotchpot" of mental ability tests developed by Binet.

It should be noted that Spearman's (1927, p. 66) perspective completely precluded the possibility that these tests could be correlated without being indicators of a single underlying construct:

> In any case, the fact that the hotchpot test-series have high correlations with one another, or in any other way actually "work," is no proof whatever that they do this by virtue of any impossible "levels" or averages. *As is much more natural*, every virtue possessed by the hotchpot procedure will find its genuine explanation in the doctrine from which this procedure really emanated.
>
> (emphasis added)

Thus, "carving nature at its joints" entailed an intuition consistent with the foundational eugenicist principle of complete heredity that recommended diverse tests as a collection of imperfect indicators for a single underlying phenomenon to be measured.

Indeed, Spearman himself (1930, p. 301) pointed to the role that his worldview played in driving his work, far outstripping any perspective-free interpretation of data:

> My conviction was accompanied by an emotional heat which cannot, I now think, be explained on purely intellectual grounds. The main source of this heat I take to have been—little as I admitted this at the time—of an *ethical* nature. Sensualism and associationism tend strongly to go with hedonism, and this latter was (and is) to me an abomination.
>
> (emphasis in original)

The association articulated here is striking, as hedonism and intelligence neither had nor have any theoretical relationship or empirical foundation. The relationship seems to lie in eugenicist assumptions about the superior intelligence of certain races and racist assumptions of alleged immorality of non-white and non-affluent communities. Thus, Spearman's intellectual work was grounded in his intuition, perspective, and subjective beliefs (i.e., idiographic elements) and played an inherent role in the construction, validation, and interpretation of measures—even when developed under a positivist approach wherein all measures were asserted to be direct apprehensions of a singular and supposedly objective reality that could be fully characterized without theory. With a different set of perspectival beliefs, Spearman's conceptualization of intelligence and its measurement would likely have differed in a variety of ways. Foundationally, his approach to testing and the interpretation of data highlights the role of idiographic elements at every phase of measurement conceptualization and implementation. This is in line with perspectives on validity that understand it not as an inherent property of a measure, but rather as an attribute intertwined with its development, use, and interpretation (Messick, 1992).

Current Considerations

Psychological scientists acknowledge the limitation of disregarding the idiographic at the level of data collection and generalizability when residents of only WEIRD (Western, educated, industrialized, rich, democratic; Henrich, Heine, & Norenzayan, 2010) countries account for the overwhelming majority of scientific data. However, similar concern over the development of the tools used to collect those data is less frequently expressed. Indeed, even (and perhaps especially) within WEIRD countries, the prevailing experiences and worldviews of scientists reflect power structures built on discrepancy and exclusion—often along the societally manufactured fault lines of race and socioeconomic status. For example, in the United States, academic employment is neither equally nor equitably distributed amongst demographic groups, with white, non-Hispanic individuals comprising a disproportionate number of awarded doctorates (National Center for Science and Engineering Statistics, 2023) and subsequent faculty positions (National Center for Educational Statistics, 2016). Likewise, first-generation college students represent 27% of Ph.D. students, though they represent 37% of students attaining bachelor's degrees (Mitic, 2022; NCES, 2016). Accordingly, the population of individuals most likely to develop, validate, and interpret measures in academic research settings are substantially less likely to hold worldviews and positionalities that reflect the diversity of experiences present in the natural population.

This is not to say that the ability of measures to accommodate idiographic aspects of phenomena is grounded in the identity of the measurement developers. As Covarrubias and Vélez (2013, pp. 272–273) note, measures "speak about the underlying views and biases of those who generated them," which does not inherently exempt "scholars of color [from] reproducing the same problematic ends as their white counterparts." Rather, I argue that measures themselves must be crafted and validated to include idiographic elements of target phenomena in addition to the nomothetic precisely because any sets of "views and biases" that supported the development of *any* measure cannot stand in place of empirically identified idiographic elements that will vary both within and across populations.

To some extent, current statistical models, such as MIMIC models have begun to engage this logic when incorporating items that measure idiographic sources of variance that impact existing nomothetic measures. However, part of the foundational difference between these prior applications and the ones illustrated in this book lies in the rationale for application of these models. When treated as covariates to account for "nuisance" variance in item responses, they remain solely focused on nomothetic constructs. When treated as being not only a necessary part of analysis to create comparable results, but as a part of the underlying measure of the construct, then the variance they explain is not nuisance, but instead an important facet for understanding the latent dimensionality of the construct and can be viewed as intrinsic to the measure.

Social Power and Measurement

As noted at the outset of this chapter, the act of measurement entails the exercise of social hierarchy and power within society. At a basic level, constructing a measure dictates how others understand the structure of a phenomenon. However, as those measurements are drawn into analyses, they also anchor understanding of the individual datum in the larger context of the data in the sample and the identified trends extrapolated from it. Thus, measurement not only informs but also *de facto* reinforces beliefs about what is worth measuring (i.e., what is valued), how it should be measured (i.e., what is epistemically legitimated), and what values are "normal" in relation to a baseline (i.e., who or what should be the standard against which to draw comparisons) (Amoore, 2020; Di Fiore et al., 2023).

For these reasons, most common forms of measurement—especially those targeting psychological phenomena—cannot be conceptualized as existing outside of a social space or escaping its influences. Indeed, methods of inquiry can be infused with a wide range of values (Bonilla-Silva & Zuberi, 2008), despite their genesis within dominant societal frameworks of white supremacy, eugenics, and imperialism (Gould, 1996; Norton, 1979; Zuberi, 2001). Accordingly, the axiology of measurement is a vital part of evaluating consequential validity (Messick, 1995), where the application of measurement can varyingly align with the values that shaped the development and deployment of those tools.

One approach that bridges axiological and theoretical frameworks to engage measurement and statistical analysis is critical quantitative analysis, or CritQuant, which adopts a lens of critical race theory to evaluate data (Gillborn, Warmington, & Demack, 2018). While the approach continues to develop across multiple variants, Gillborn and colleagues posit five core principles (p. 169):

1 The centrality of racism.
2 Numbers are not neutral.
3 Categories are neither "natural" nor given: For "race" read "racism."
4 Voice and insight: Data cannot "speak for itself."
5 Using numbers for social justice.

Collectively, these principles emphasize the ways in which personal and societal perspectives infuse and shape measurement and quantitative analysis. Especially in the United States and Western Europe, racism is a pervasive and persistent influence that is historically intertwined with the development of both measures and statistical methods (Gould, 1996; Zuberi, 2001). It persists as a de facto framework that centers and prioritizes white norms and expectations unless deliberately engaged and disrupted (Mohajeri, 2021). In Hawkman's (2020, p. 404) words, whiteness is:

> An ever-shifting, hierarchical, hegemonic power structure and identity construct that informs the ways individuals view themselves and society and is predicated on dehumanizing the racial other.... Whiteness has also been described as the water in which white people swim.... It is all around them, keeping them afloat.

This sentiment highlights the presumed normalcy of whiteness, such that it is readily supposed to be the default social state rather than an active imposition of status quo power structures and cultural norms that convey societal and financial resources. Through the presumption of universality for those norms, it is anticipated that any standards or understandings which adhere to them can be validly applied to those without equitable access to the societal machinery that reinforces it. In this sense, whiteness is not simply a construct of skin color or race; it becomes synonymous with social norms that are enforced by those whose privilege is reinforced by the reification of those norms. Accordingly, those who do not subscribe to those normative behaviors and values as a function of their experiences linked to gender, sexuality, social class, and/or disability are subject to a power structure that positions them as outside a meritorious norm (Stewart & Nicolazzo, 2018).

From this perspective, the interpretation of numbers cannot be objective, because they are understood in the context of conventions that originate from social perspectives. Consequently, data and the interpretation of those data cannot be independent of the identities and perspectives of researchers making claims based on those data. Researchers claiming to attend neutrally to data are ignoring default assumptions that socially constructed and defined race somehow connotes deficits as inherent natural properties, which extend to the definition and selection of categories that are actively constructed based on social relationships and meanings. Thus, to avoid perpetuating bias and inequities that may readily be propagated through quantitative research, CritQuant work must actively undertake efforts to use and rebuild these tools to offer fairer and more just understandings and policies in society.

One approach to enhancing fairness in measurement is to detect instances of differential item function by demographic or cultural group and investigate them through targeted qualitative inquiry to understand why the psychometric properties of an item do not perform as expected (David, Hitchcock, Ragan, Brooks, & Starkey, 2018; Hitchcock & Johanson, 2015). However, such efforts are often used to remove rather than understand differential item functioning, essentially reverting to a purely nomothetic framework when the use of such strategies are to eliminate DIF without foundationally building more inclusive or differentiated measures. In other words, this approach perpetuates the tradition of treating idiographic variance as noise rather than signal and misses opportunities to make measures more complete by grappling with both the nomothetic and the idiographic in measurement development and interpretation. Accordingly, other approaches to instrument development and validation specifically engage marginalized communities to capture meanings authentic to those groups and anchor instrument development on those meanings (e.g., Sablan, 2019). Another approach is to ground specific instrument development in literature developed by and about members of a particular social or cultural group (Pérez Huber, Vélez, & Solórzano, 2018; Toldson, 2019).

CritQuant also encourages critical analyses of the ways in which constructs developed by dominant groups may shape the nature of measures in ways that can obscure the meanings or lived experiences of marginalized groups and consequently sustain false causal or presumed narratives (Stage, 2007; Zuberi, 2001). Covarrubias and Vélez (2013, p. 273) explain:

> We believe that the potential of this work rests not in its ability to be "objective" and "un-biased" but in how we foreground our positionality in connection to the research and contextualize our findings and analysis in relationship to our causal theories of how the world

> operates. Masking our intentions any other way gives undue power to statistical methods, when, in actuality, power rests in the theories used to interpret social data, whether implicitly or explicitly.

Broadly, efforts to prevent the perpetuation of racialized or other biased/biasing conclusions from measurement through CritQuant emphasize the purpose or intent to bring to bear a critical lens, as specific analytic techniques themselves (e.g., regression) can be applied across frameworks. As Stage (2007, p. 9) notes, "If we focus solely on research methods—arguably the less interesting of a researcher's concerns—we see little difference between the positivistic approach and the critical quantitative approach." While this observation is likely true in the application of statistical analyses, it is not clear that the approaches to measurement should be quite so ubiquitous, as the formation of a measure dictates the ways in which phenomena are encoded into numeric representations. These numeric representations are easily reified to take on a meaning of their own beyond a probabilistic reflection of an underlying latent variable and motivate actions in their own right (Gould, 1996; Strunk, 2023). Sablan (2019, p. 198) further notes that "too few methodological guidelines can leave an empirical gap, where [critical]-intending scholars have few quantitative cases, with little methods diversity, to turn to for exemplars." Accordingly, this book has offered several new approaches to measurement that deliberately draw on both idiographic and nomothetic sources of information to construct and interpret measures in Chapters 4–6.

Application of Methods

Chapter 4—Quantitative Member Checking

Chapter 4 introduces the concept of quantitative member checking, which elicits information about idiographic sensemaking from a respondent associated with a selected response option from a closed-ended measurement item. Specifically, the respondent can indicate how well an item aligns with their imagined optimal answer. Represented on a Likert scale, this information is incorporated into the measurement model through structured residuals and quantifies the extent of incommensurability between the item writer and the item respondent. When incorporated into a multiple-indicator multiple cause (MIMIC) model (Jöreskog & Goldberger, 1975), it acts as a way to predict residual variance, therefore differentiating non-random residual variance due to incommensurability between item writer and respondent from random "error." Mathematically, this approach can be implemented to explain an increased proportion of non-random variance in the manifest response variables.

Doing so identifies the impact of alignment between the sensemaking of the respondent and the framing of the item on the residual variance rather than the overall manifest item variance. This model apportions the variability of any observed indicator into that attributable to the latent factor and that attributable to the structured residual, resulting in what could be framed as an assessment of validity at the item level in the sense that the selected response option truly reflects the target latent construct. Accordingly, the approach can be used to remove the invisibility of assumptions held by the developer of the measure and gauge their impact as a function of the incommensurability between their perspective and the perspective of respondents. Where items reflect greater proportions of variance in the structured residuals, they can be identified as limiting the ability of the measure to directly reflect idiographic differences and avoid treating meaningful differences in perspective and sensemaking as "error." This captures the impact that experiential differences can have on the validity of measures at the item-level and can support the development and evaluation of instruments that reflect, rather than mask, the impacts of personal experience or context on derived scores which have historically been understood as "standardized."

Chapter 5—Idiographic Measures

Chapter 5 uses idiographic data to construct a modified latent growth model through carefully defined categorical variables. Fitting models that first restricted all variance not due to the group trend to a time-specific residual, these models 1) operationalized residuals as the difference between the observed values of individual participants at a given point in time and the group trend line and 2) identified positive autoregression between those residuals over time. While this modeling approach was an unconventional application of a latent growth model, it is important to note that the aggregate between-person trend is what is most often interpreted in applied work. While variability can be estimated around this trend, it often is measured in a way that does not reflect the within-person stability of results in reference to the group trend. Ultimately, these results demonstrated that idiographic constructs at the individual level sustained within-person trajectories through semantic space that could be identified in relation to the interpreted between-person trend. Further, residual variance that reflected individual departures from the normative group trajectory were interpretable as meaningful signal rather than conceptualized as random "error" variance.

These individual trajectories were also matched against qualitative case studies to confirm that the idiographic meaning incorporated into the criteria for value assignment of binary categorical variables was not irretrievably reduced. By attaining convergence in the interpretation of residual magnitude and qualitative interview data, we demonstrate that the treatment of residual variance at the level of the individual preserves idiographic information extensible in longitudinal modeling. Further, it illustrates how validity may be understood within persons as a form of measurement stability, and not solely as a between-person concept.

When individual-level residuals demonstrate positive relationships across timepoints, it highlights the non-arbitrary nature of idiographic information and positions individual idiographic trajectories to be interpreted in relation to group trends. Evaluating intra- and inter-individual variance within a model based on idiographic data emphasizes that data from individuals who diverge from the group trend are *neither* interchangeable *nor* represent error, in contrast to Novick's (1966) assumptions of interchangeability and random response in factor models. Instead, the meaning of residuals is anchored in the continuity of sensemaking that occurs within individuals over time. Individuals do not need to experience a phenomenon identically or within a minimal margin of uncertainty for this modeling approach to be interpretable while grounded in idiographic rather than nomothetic measurements of the target constructs. Although some quantitative methods have been developed that can theoretically encompass this viewpoint by disaggregating between- and within-person variance (e.g., Curran, Howard, Bainter, Lane, & McGinley, 2014; Hamaker et al., 2015), the demonstration provided in Chapter 5 lays groundwork for the further development of statistical methods that accommodate idiographic data as intrinsic and interpretable in relation to phenomena rather than idiosyncratic or arbitrary noise around a normative standard. In doing so, the approach facilitates the development of measurement models that do not erase the positionality or personally constructed meanings of individuals, even if those individuals represent a minority of the sample or are members of historically marginalized communities whose perspectives are not typically reflected in developed instruments.

Chapter 6—Idiographic Moderation in Measurement Models

Chapter 6 introduces the use of an idiographic construct as a moderator of factor loadings and intercepts to resolve DIF and item parameter drift (IPD). This use of idiographic data directly incorporates idiographic meaning into measurement models and enhances the ability of a measure with DIF/IPD to be used meaningfully across individuals whose different experiences impact their response patterns. Using moderated nonlinear factor analysis (MNLFA; Bauer, 2017), it positions idiographic factors as intrinsic to measurement models in a way that both enhances measurement invariance and more fully links the measure to both the nomothetic and idiographic facets of the target phenomenon.

From a critical lens, one of the historic concerns about quantitative measurement is that it treats data drawn from individuals as interchangeable (Novick, 1966), which inherently divorces individuals' identities, positionalities, and idiographic constructed meanings from the inferences that can be drawn. Likewise, differences in socially constructed categories such as race are conceptualized as discrete values to be integrated as independent variables or covariates with direct effects on only a single aspect of a model that belies the multifaceted and ongoing interactions that shape human experiences (Stewart, 2008). Idiographic MNFLA reinfuses measurement models with idiographic meaning by permitting categorical variables reflecting individuals' constructed and articulated sensemaking, identity, or identity facets independently or in combination to influence the estimated relationships between individual items and the latent variable associated with the target phenomenon simultaneously in multiple ways. Moderating *both* metric (factor loading) *and* scalar (intercept) values provides an avenue for idiographic factors to account for both how individuals make sense of questions and their likelihood of selecting a response value. To the extent that accommodating these influences in the model decreases DIF and IPD, it enhances the comparability of scores between groups and across time through the inclusion and valuing—rather than the erasure and presumption of irrelevance—of personal meaning.

Conclusion

Audre Lorde, a renowned feminist, pointed out at a 1984 conference on the scholarly study of women's lives that important perspectives and voices were completely absent from both the scholarship presented and the scholars themselves: Black women, lesbians, and women from developing countries. In her comments, she argued that the feminist work of the conference would not ultimately yield the societal impacts it aspired to because the perspectives brought to bear were those of societally normed academe. Specifically, she argued that "the master's tools will never dismantle the master's house. They may allow us to beat him temporarily at his own game, but they will never enable us to bring about genuine change" (Lorde, 2007, p. 112).

In this sense, a mixed methodological approach to psychometrics offers opportunities to deliberately engage idiographic perspectives in parity with nomothetic aspects of target phenomena. Doing so offers new tools to reduce the incommensurability between measures and the underlying human phenomena that they try to capture. The incorporation of idiographic facets preserves the voices and perspectives of the people from whom data are collected. These tools retain rather than discard the sensemaking that reflects their life histories, world views, and positionality within society. Embracing these as essential to valid measurement rather than preventing it yields tools that are not bound to the presumed universality of specific perspectives which tend to dominate social and scientific discourse at the cost of invisibility for those with perspectives that differ. Use of new tools for measurement holds the potential to elevate our understanding of phenomena in ways that fundamentally alter the power structures represented and fueled by classical nomothetic approaches to measurement.

References

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others.* Durham, NC: Duke University Press.

Bauer, D. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526.

Bonilla-Silva, E., & Zuberi, T. (2008). Toward a definition of white logic and white methods. In T. Zuberi, & E. Bonilla-Silva (Eds.), *White logic, white methods: Racism and methodology* (pp. 3–30). New York, NY: Rowman & Littlefield Publishers, Inc.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, MA: Cambridge University Press.

Covarrubias, A., & Vélez, V. (2013). Critical race quantitative intersectionality: An anti-racist research paradigm that refuses to Let the numbers speak for themselves. In M. Lynn, & A. Dixson (Eds.), *Handbook of critical race theory in education* (pp. 270–286). New York, NY: Routledge.

Curran, P., Howard, A., Bainter, Lane, S., & McGinley, J. (2014). The separation of between-person and within-person components of individual change over time: A latent curve model with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*, 879–894.

David, S., Hitchcock, J. H., Ragan, B., Brooks, G., & Starkey, C. (2018). Mixing interviews and rasch modeling: Demonstrating a procedure to develop and instrument that measures trust. *Journal of Mixed Methods Research*, *12*, 75–94.

Di Fiore, M., Kuc-Czarnecka, M., Piano, S., Puy, A., & Saltelli, A. (2023). The challenge of quantification: An interdisciplinary reading. *Minerva*, *61*, 53–70.

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179.

Gould, J. (1996). *The mismeasure of man.* New York, NY: W. W. Norton & Company.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116.

Hawkman, A. M. (2020). Swimming in and through whiteness: Antiracism in social studies teacher education. *Theory & Research in Social Education*, *48*(3), 403–430.

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

Hitchcock, J. H., & Johanson, G. A. (2015). Applying a mixed methods framework to differential item function analyses. *Research in the Schools*, *22*(1), 1–14.

Jöreskog, K. G., & Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Lorde, A. (2007). The master's tools will never dismantle the master's house. In N. Bereano (Ed.), *Sister outsider: Essays and speeches* (pp. 110–114). Berkeley, CA: Crossing Press.

MacKenzie, D. (1978). Statistical theory and social interests: A case-study. *Social Studies of Science*, *8*, 35–83.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. Research report.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Mitic, R. (2022). *Insights into first-generation doctoral students. CGS Research in Brief.* Washington, DC: Council of Graduate Schools.

Mohajeri, O. (2021). "Fly on the wall" moments reveal whiteness-at-work for contested white graduate students. *International Journal of Qualitative Studies in Education*, *35*(4), 393–409.

National Center for Education Statistics, U.S. Department of Education. (2016). Postsecondary education: Characteristics of postsecondary faculty. In *The condition of education 2016* (pp. 222–225; NCES 2016-144). Washington, DC: National Center for Education Statistics.

National Center for Science and Engineering Statistics (NCSES) (2023). *Doctorate recipients from U.S. universities: 2022. NSF 24-300.* Alexandria, VA: National Science Foundation.

Norton, B. (1979). Charles spearman and the general factor in intelligence: Genesis and interpretation in the light of sociopersonal considerations. *Journal of the History of the Behavioral Sciences*, *15*, 142–154.

Novick, M. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18.

Padilla, A. (2004). Quantitative methods in multicultural education research. In J. Banks, & C. McGee Banks (Eds.), *Handbook of research on multicultural education* (2nd ed., pp. 127–145). Mahwah, NJ: Jossey-Bass.

Pérez Huber, L., Vélez, V., & Solórzano, D. (2018). More than 'papelitos:' a QuantCrit counterstory to critique Latina/o degree value and occupational prestige. *Race Ethnicity and Education*, *21*(2), 208–230.

Sablan, J. R. (2019). Can you really measure that? Combining critical race theory and quantitative methods. *American Educational Research Journal*, *56*(1), 178–203.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan and Co., Ltd.

Spearman, C. (1930). C. Spearman. In C. Murchison (Ed.), *A history of psychology in autobiography* (Vol. *1*, pp. 299–334). Worcester, MA: Clark University Press.

Stage, F. K. (2007). Answering critical questions using quantitative data. *New Directions for Institutional Research*, *133*, 5–16.

Strunk, K. (2023). QuantQueer: Renovating quantitative methods through queer and critical theoretical traditions. *Educational Studies*, *60*, 5–18.

Stewart, Q. (2008). Swimming upstream: Theory and methodology in race research. In T. Zuberi, & E. Bonilla-Silva (Eds.), *White logic, white methods: Racism and methodology* (pp. 111–126). New York, NY: Rowman and Littlefield Publishers, Inc.

Stewart, D. L., & Nicolazzo, Z. (2018). High impact of [whiteness] on trans* students in postsecondary education. *Equity & Excellence in Education*, *51*(2), 132–145.

Toldson, I. (2019). *No BS (bad stats)*. Boston, MA: Brill Sense.

Urbach, P. (1974). Progress and degeneration in the i. Q. Debate. *British Journal for the Philosophy of Science*, *25*, 99–135.

Zuberi, T. (2001). *Thicker than blood: How racial statistics lie*. Minneapolis, MN: University of Minnesota Press.